



**HAL**  
open science

# **SIMDIAL, un paradigme d'évaluation automatique de systèmes de dialogue homme-machine par simulation déterministe d'utilisateurs**

Joseph Allemandou

► **To cite this version:**

Joseph Allemandou. SIMDIAL, un paradigme d'évaluation automatique de systèmes de dialogue homme-machine par simulation déterministe d'utilisateurs. Informatique [cs]. Université Paris Sud - Paris XI, 2007. Français. NNT : . tel-00620803

**HAL Id: tel-00620803**

**<https://theses.hal.science/tel-00620803>**

Submitted on 8 Sep 2011

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

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

# **SIMDIAL**

Un paradigme d'évaluation automatique  
de systèmes de dialogue homme-machine  
par simulation déterministe d'utilisateurs

Thèse en Informatique

LIMSI/CNRS

FRANCE TELECOM

par

JOSEPH ALLEMANDOU

Membres du jury :

MM JEAN CAELEN                      RAPPORTEURS  
DANIEL LUZZATI

MM JOSEPH MARIANI                  EXAMINATEURS  
STÉPHANE CHAUDIRON  
JEAN-YVES ANTOINE  
LAURENCE DEVILLERS  
LAURENT CHARNAY  
MURIEL LAUVERGNE



*Pour Mélissa  
et toutes celles et tous ceux  
qui m'ont aidé et m'aident encore  
à faire que demain nous sera meilleur qu'aujourd'hui.*



# Remerciements

En premier lieu je tiens à remercier grandement JOSEPH MARIANI, qui a dirigé mes recherches durant ces trois dernières années. Ses remarques et conseils, toujours justes et constructifs, m'ont permis de mener à bien ce travail dans des conditions remarquables.

Par ailleurs cette thèse est le fruit de nombreuses heures de travail de la part de mes encadrants, LAURENT CHARNAY, LAURENCE DEVILLERS et MURIEL LAUVERGNE. Outre leurs nombreuses qualités scientifiques, ils m'ont guidé vers la meilleure des voix possibles pour le déroulement de mon travail, même lors d'avis divergents. Ils méritent mes remerciements les plus chaleureux et sincères.

Je souhaite aussi adresser une pensée à toutes celles et tous ceux dont j'ai croisé le chemin grâce à ces trois années de recherche. Des collègues de laboratoire que j'ai beaucoup fréquentés, comme DOM, LISA, ERWAN, et FRED, à ceux que j'ai connu de façon plus lointaine, comme MORGANE, GÉRARD, MAGALIE, MERIAM, CEDRIC, LUDO, ANTOINE (les deux), CÉLINE, et j'en oublie ... Quelques rencontres Armoricaïnes, avec TITI, HÉLÈNE, MAUD, TOF, VICTOIRE, ALLEN et MARINE. Merci aussi à ma PETITE BRETONNE, pour la gaieté et la richesse de mon quotidien.

Enfin je veux souligner la place particulière qu'a eu ma famille ces dernières années, et notamment celle de ma Fratrie, LÉA et MATTHIEU, qui ont toujours su voir en moi l'homme que je m'efforce d'être.

Merci du fond du coeur.



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivations . . . . .	1
1.2	Contexte pratique . . . . .	2
1.3	Contexte théorique . . . . .	2
1.4	Contribution . . . . .	3
1.5	Plan du document . . . . .	5
<b>I</b>	<b>Dialogue homme-machine et évaluation</b>	<b>7</b>
<b>2</b>	<b>Le dialogue homme-machine</b>	<b>9</b>
2.1	Bref historique . . . . .	9
2.2	Différents niveaux d'interaction . . . . .	10
2.2.1	Interaction par écrit . . . . .	10
2.2.2	DTMF . . . . .	11
2.2.3	Interaction vocale contrainte . . . . .	11
2.2.4	Interaction vocale naturelle . . . . .	11
2.2.5	Interaction multimodale . . . . .	12
2.3	Caractéristiques du dialogue spontané . . . . .	12
2.3.1	Au niveau de l'énoncé . . . . .	13
2.3.2	Au niveau du dialogue . . . . .	14
2.4	Les systèmes de dialogue homme-machine . . . . .	16
2.4.1	Architecture standard . . . . .	16
2.4.2	Modélisation de l'interaction . . . . .	18
2.4.3	Défis scientifiques . . . . .	20
2.5	Conclusion . . . . .	21
<b>3</b>	<b>L'évaluation de systèmes</b>	<b>23</b>
3.1	Généralités . . . . .	23
3.1.1	Définition . . . . .	23
3.1.2	Buts, valeurs et attributions . . . . .	24



3.1.3	Observation et jugement . . . . .	24
3.1.4	Objectif vs. subjectif . . . . .	25
3.2	Différents points de vue . . . . .	26
3.2.1	Approches ergonomiques et technologiques . . . . .	26
3.2.2	Utilisateur et usager, utilisabilité et usabilité . . . . .	27
3.2.3	Évaluation qualitative et quantitative . . . . .	27
3.2.4	Système perçu en boîte noire ou transparente . . . . .	28
3.2.5	Évaluation avec mesure ou référence . . . . .	29
3.2.6	Modélisation et évaluation . . . . .	29
3.3	L'évaluation de systèmes de dialogue . . . . .	30
3.3.1	Les buts de l'évaluation . . . . .	30
3.3.2	L'intimité avec le système à évaluer . . . . .	31
3.3.3	Les aspects dynamiques et co-construits du dialogue . . . . .	33
3.3.4	La généricité . . . . .	36
3.4	Conclusion . . . . .	37
<b>4</b>	<b>État de l'art sur l'évaluation de SDHM</b>	<b>39</b>
4.1	Cadres d'évaluation . . . . .	39
4.1.1	EAGLES I et II . . . . .	39
4.1.2	DISC . . . . .	40
4.2	Évaluation de l'interprétation sémantique . . . . .	41
4.2.1	Évaluation hors contexte de dialogue . . . . .	42
4.2.2	Évaluation en contexte de dialogue . . . . .	47
4.3	Évaluation de la gestion de l'interaction . . . . .	49
4.3.1	Utilisateurs réels . . . . .	50
4.3.2	Simulation stochastique d'utilisateurs . . . . .	52
4.3.3	Simulation déterministe d'utilisateurs . . . . .	55
4.4	Conclusion . . . . .	56
<b>II</b>	<b>Le paradigme SIMDIAL</b>	<b>59</b>
<b>5</b>	<b>Choix méthodologiques et technologiques</b>	<b>61</b>
5.1	Choix méthodologiques . . . . .	61
5.1.1	Approche générale . . . . .	61
5.1.2	Généricité . . . . .	62
5.1.3	Évaluation avec mesure ou référence . . . . .	63
5.1.4	Utilisateurs réels ou simulés . . . . .	63
5.1.5	Simulation stochastique ou déterministe . . . . .	64
5.1.6	Critères d'évaluation . . . . .	65
5.1.7	Conclusion . . . . .	66

5.2	Choix technologiques . . . . .	66
5.2.1	Architecture et interface . . . . .	66
5.2.2	La gestion de l'interaction . . . . .	68
5.2.3	Les autres modules . . . . .	70
5.3	Évaluer l'évaluation . . . . .	71
5.4	Conclusion . . . . .	72
<b>6</b>	<b>Le simulateur d'utilisateurs</b>	<b>73</b>
6.1	Modèle de l'interaction . . . . .	73
6.1.1	Modèle du domaine . . . . .	73
6.1.2	Modèle de la tâche . . . . .	75
6.1.3	Modèle du dialogue . . . . .	77
6.1.4	Modèle de l'utilisateur simulé . . . . .	78
6.1.5	Algorithme de génération du contenu de la réponse . . . . .	79
6.2	Phénomènes perturbateurs . . . . .	79
6.2.1	Hésitation et phénomènes linguistiques . . . . .	81
6.2.2	Ambiguïté et phénomènes dialogiques . . . . .	82
6.3	Autres modules . . . . .	83
6.3.1	Annotation et validation des corpus . . . . .	83
6.3.2	Interprétation sémantique . . . . .	84
6.3.3	Génération en langue naturelle . . . . .	85
6.3.4	Diagnostic automatique . . . . .	88
6.4	Coût de transformation du simulateur . . . . .	90
6.5	Conclusion . . . . .	91
<b>7</b>	<b>Applications du paradigme SIMDIAL</b>	<b>93</b>
7.1	Évaluation de l'application PlanResto . . . . .	93
7.1.1	Observations . . . . .	93
7.1.2	Analyses . . . . .	96
7.1.3	Exemples de dialogues . . . . .	98
7.2	Évaluation de l'application DialogueBourse . . . . .	104
7.2.1	Observations . . . . .	105
7.2.2	Analyses . . . . .	106
7.2.3	Exemples de dialogues . . . . .	108
7.3	Conclusion . . . . .	112
<b>8</b>	<b>Conclusion et Perspectives</b>	<b>113</b>
8.1	Conclusion . . . . .	113
8.1.1	Méthodes et modèles employés . . . . .	113
8.1.2	Propriétés du paradigme SIMDIAL . . . . .	114
8.2	Perspectives . . . . .	116

<b>III</b>	<b>Annexes</b>	<b>119</b>
<b>A</b>	<b>Tâche PlanResto</b>	<b>121</b>
<b>B</b>	<b>Tâche DialogueBourse</b>	<b>123</b>
<b>C</b>	<b>Test perceptif</b>	<b>125</b>
	<b>Bibliographie</b>	<b>143</b>

# Liste des tableaux

4.1	Exemple d'énoncé annoté pour le paradigme MADCOW . . . . .	43
4.2	Exemple de jeux de test pour le paradigme DCR . . . . .	45
4.3	Exemple d'énoncé dérivé pour le paradigme DEFI . . . . .	47
4.4	Exemple d'énoncé annoté pour le paradigme PEACE . . . . .	48
6.1	Concepts et Valeurs génériques . . . . .	74
6.2	Concepts et Valeurs spécifiques à la tâche PLANRESTO . . . . .	74
6.3	Concepts et Valeurs spécifiques à la tâche DIALOGUEBOURSE . . . . .	75
6.4	Liste des actes du modèle de dialogue . . . . .	77
7.1	Dialogues générés et tours de paroles associés pour la tâche PLANRESTO . . . . .	95
7.2	Tours de parole par dialogue pour la tâche PLANRESTO . . . . .	95
7.3	Dialogue entre PLANRESTO et le simulateur ressemblant à un dialogue humain-machine . . . . .	99
7.4	Dialogue avec l'application PLANRESTO où les énoncés du si- mulateur ne semblent pas humains . . . . .	101
7.5	Dialogue où l'application PLANRESTO et le simulateur bou- clent de façon infinie . . . . .	104
7.6	Dialogues générés et tours de paroles associés pour la tâche DIALOGUEBOURSE . . . . .	106
7.7	Ratios pour les dialogues aboutis de la tâche DIALOGUEBOUR- SE . . . . .	106
7.8	Dialogue entre DIALOGUEBOURSE et le simulateur ressem- blant à un dialogue humain-machine . . . . .	109
7.9	Dialogue entre DIALOGUEBOURSE et le simulateur ressem- blant à un dialogue humain-machine . . . . .	110
7.10	Dialogue non abouti pour cause de reprise à zéro de DIALO- GUEBOURSE . . . . .	112
A.1	Répartition des valeurs du thésaurus relatif à la tâche PLAN- RESTO . . . . .	121

A.2	Répartition des annotations dialogiques de verbalisation pour le domaine de la tâche PLANRESTO . . . . .	121
A.3	Répartition des annotations sémantiques de verbalisation pour le domaine de la tâche PLANRESTO . . . . .	122
B.1	Répartition des valeurs du thésaurus relatif à la tâche DIALOGUEBOURSE . . . . .	123
B.2	Répartition des annotations dialogiques de verbalisation pour le domaine de la tâche DIALOGUEBOURSE . . . . .	123
B.3	Répartition des annotations sémantiques de verbalisation pour le domaine de la tâche DIALOGUEBOURSE . . . . .	124
C.1	Résultats du test perceptif . . . . .	125
C.2	Énoncés produits par la stratégie de verbalisation . . . . .	126
C.3	Transcriptions automatiques d'énoncés produits par des humains . . . . .	127

# Table des figures

2.1	Représentation fonctionnelle standard d'un SDHM . . . . .	17
5.1	Interface du simulateur avec un système . . . . .	67
6.1	Modèle de la tâche pour PLANRESTO . . . . .	76
6.2	Modèle de la tâche pour DIALOGUEBOURSE . . . . .	76
6.3	Algorithme de génération du contenu de la réponse . . . . .	80



# Chapitre 1

## Introduction

### 1.1 Motivations

À l'heure actuelle le domaine des interfaces homme-machine est largement étudié. Il devient même un sujet majeur de recherche dans la communauté informatique. Par ailleurs de plus en plus d'outils électroniques sont utilisés par les gens dans leur vie de tous les jours.

Cette incursion de l'électronique dans les maisons et les bureaux est dûe non seulement à la facilité avec laquelle il aide à résoudre de nombreuses tâches ainsi qu'à son prix de plus en plus attractif, mais aussi aux interfaces qui facilitent de plus en plus son utilisation.

Les téléphones portables et les ordinateurs personnels sont d'excellents exemples d'outils ayant un succès croissant grâce à leur facilité d'utilisation.

Les progrès réalisés ces dernières décennies dans les domaines de la reconnaissance de la parole, de la synthèse vocale, de la compréhension et de la génération du langage naturel ont rendu possible le développement d'interfaces vocales et de systèmes dialoguants qui deviennent aujourd'hui de plus en plus courants[Luzzati, 1995].

Néanmoins les interfaces vocales réellement proposées aux utilisateurs n'ont pas encore dépassé les capacités des interfaces standard, ni en terme de résolution des tâches, ni en terme d'ergonomie.

Les évolutions à apporter aux interfaces vocales afin qu'elles deviennent réellement plus intéressantes pour les utilisateurs que les interfaces classiques sont multiples; des problèmes de reconnaissance vocale au manque de flexibilité générale du système. Leurs améliorations ont pour point commun la nécessité de les évaluer.



## 1.2 Contexte pratique

Ce travail de trois ans a été co-encadré par le laboratoire universitaire LIMSI et par l'entreprise FRANCE TELECOM, qui a aussi financé la thèse.

Le laboratoire de recherche et développement ayant accueilli ces recherches travaillait sur une technologie générique de système de dialogue pour la recherche d'information nommée ARTIMIS [Sadek et al., 1997]. Cette technologie permet de produire des SDHM dont le niveau d'interaction, via texte ou signal sonore, est l'interaction naturelle avec l'utilisateur (voir 2.2.4).

L'encadrement universitaire a été effectué par des membres de l'équipe Traitement du Langage Parlé. Les sujets de recherche traités par cette équipe sont la reconnaissance automatique de la parole, la compréhension de la langue naturelle ainsi que les systèmes de dialogue.

En outre, de par la proximité des sujets et la participation des deux laboratoires d'encadrement à l'action Technolanguage organisée par le LIMSI, il m'a été possible de contribuer à la campagne d'évaluation MEDIA. Cette campagne portait sur l'évaluation des composants d'interprétation sémantique hors et en contexte de dialogue (voir 4.2.2).

Le projet de thèse tel qu'il avait été défini peut alors être résumé comme suit : création d'un paradigme outillé d'évaluation de systèmes de dialogue, permettant d'évaluer automatiquement de façon indépendante de l'application et de la technologie de dialogue sous-jacente au système dialoguant. Ce paradigme permettra d'alimenter et de corroborer des études psychologiques effectuées dans les laboratoires de FRANCE TELECOM.

## 1.3 Contexte théorique

Aujourd'hui il n'existe pas, dans la communauté de recherche, de méthodologie standard ni même de pratique communément admise permettant d'évaluer et de comparer des systèmes de dialogue. La nature dynamique et co-construite du dialogue rend en effet difficile la constitution tant de jeux de tests que de mesures précises offrant un référentiel d'évaluation commun à plusieurs tâches [Devillers et al., 2002].

Néanmoins des méthodologies d'évaluation pour les systèmes de dialogue oraux ont été développées par des projets de taille conséquente, sans pour autant fournir de paradigme standard, en commençant par le projet francophone B2 AUPELF-UREF [Mariani, 1998], l'évaluation par DEFI [Antoine et al., 2002b], les projets européens EAGLES [King et al., 1996], DISC [DISC

Consortium, 1999], SUNDIAL [Gibbon et al., 1997] ainsi que les projets ATIS [Hirschman et al., 1992].

Le paradigme PARADISE est à l'heure actuelle le plus abouti puisqu'il a permis de confronter, lors d'évaluations réalisées dans le cadre du projet COMMUNICATOR [Walker et al., 2001] aux États-Unis, des mesures quantitatives et qualitatives obtenues auprès de vrais sujets sur différents composants d'un système de dialogue. La confrontation avec de vrais sujets a deux défauts majeurs. D'une part elle n'est pas reproductible, ou à tout le moins pas aisément, et d'autre part elle est coûteuse en termes de temps et de ressources.

Les évaluations mettant en jeu des utilisateurs réels sont évidemment indispensables à un moment du développement d'un système. Cependant il est possible, comme le propose tout un courant actuel de recherches, de contourner les défauts de cette approche en utilisant des utilisateurs simulés permettant de générer artificiellement des dialogues et d'évaluer ainsi les systèmes. La majeure partie des simulateurs d'utilisateurs réalisés aujourd'hui est fondée sur des apprentissages statistiques.

## 1.4 Contribution

L'évaluation d'un système de dialogue passe nécessairement par l'observation de ce système, et donc par l'observation de dialogues auxquels a pris part le système. Le moyen le plus aisé pour observer de nombreux dialogues de systèmes est alors de *le faire dialoguer*. Le paradigme SIMDIAL a donc été réalisé en utilisant la technique de simulation d'utilisateurs afin de dialoguer avec les systèmes et ainsi d'observer leurs comportements.

Le paradigme SIMDIAL propose d'évaluer des systèmes de dialogue en interagissant avec eux en fonction de différents paramètres relatifs à l'évaluation souhaitée. Cette approche permet ainsi d'évaluer les systèmes de façon automatique et en contexte de dialogue, par la co-construction dynamique des dialogues avec les systèmes évalués.

Pour éviter de simuler des utilisateurs à partir de modèles statistiques, ce qui demande un grand nombre de données, la démarche est ici de générer des dialogues « réalistes » à partir d'utilisateurs simulés de façon déterministe. Ainsi les comportements des utilisateurs simulés sont précisément définis tout en étant variables en fonction de différents paramètres.

Le choix de simuler des utilisateurs de façon déterministe dans le paradigme SIMDIAL a été fait pour deux raisons majeures. D'une part pour réduire les coûts associés à l'annotation des corpus d'apprentissage, et d'autre

part pour pouvoir réaliser des évaluations comparatives. En effet puisque les comportements dialogiques du simulateur sont précisément déterminés, il devient possible de comparer différents systèmes face à ces comportements.

Par ailleurs afin d'assurer une généricité maximale du simulateur face aux systèmes avec lesquels il doit dialoguer, leur niveau d'interaction dans le paradigme *SIMDIAL* est celui du langage naturel écrit. Ce choix permet théoriquement d'interfacer le simulateur avec tout système de dialogue, puisque ce dernier interagit normalement en langage naturel. De plus le fait d'utiliser le niveau de la transcription textuelle plutôt que celui du signal sonore pour l'interaction permet de générer automatiquement un grand nombre d'énoncés similaires à ceux produits par des humains. Une stratégie de génération automatique d'énoncés à partir de petits corpus et de thésaurus sur la tâche a ainsi fait l'objet d'un brevet déposé par *FRANCE TELECOM*, auquel a participé le *LIMSI*.

Un ensemble de critères d'évaluation classiquement utilisés est également mis en œuvre dans le paradigme *SIMDIAL*, comme l'aboutissement de la tâche à résoudre ou le nombre de tours de parole par dialogue pour la résoudre. Par ailleurs il est intéressant de noter que les mesures de ces critères n'utilisent pas de traces des systèmes évalués, ce qui maximise la généricité du paradigme à ces derniers.

D'autres phénomènes d'ordre linguistiques et dialogiques sont aussi définis et utilisés à des fins de diagnostic. Ces critères sont exploités à partir des traces des dialogues générés mais certains pourraient également être mis en œuvre de façon interactive.

Le paradigme *SIMDIAL* est avant tout une plateforme outillée, mais il pourrait être enrichi de nombreuses mesures tant quantitatives que qualitatives (proches des mesures en psycho-ergonomie).

Enfin le paradigme *SIMDIAL* a été mis en application par un simulateur d'utilisateurs ayant permis d'évaluer deux applications sur des tâches différentes. Ces expériences ont permis de valider d'une part que le moteur de simulation des utilisateurs était bien indépendant des systèmes de dialogue testés, et d'autre part que la réalisation de différentes tâches par la même architecture de dialogue était réalisable à faible coût.

Ont ainsi été montrées la portabilité de ce paradigme d'une tâche à une autre ainsi que la richesse des mesures exploitables dans la plateforme (mesures hors contexte et en contexte de dialogue, comme par exemple la robustesse aux hésitations et le relâchement de contraintes). Les perspectives sont maintenant de tester ce paradigme sur des systèmes inconnus afin de vérifier

la robustesse et la généralité de la simulation mise en œuvre.

## 1.5 Plan du document

La première partie de cette thèse présente de façon générale le dialogue homme-machine ainsi que l'évaluation de systèmes de dialogue. Le domaine du dialogue homme-machine (chapitre 2) est très large et s'appuie sur de nombreux autres, tels que le traitement du signal pour la reconnaissance de la parole, la représentation des connaissances pour l'interprétation ou la gestion du dialogue, ou encore la psychologie cognitive pour juger des comportements ou de la satisfaction des utilisateurs. L'évaluation de systèmes de dialogue (chapitre 3) est un sujet en émergence ces dernières années, notamment suite aux avancées dans le domaine du dialogue homme-machine. Enfin un état de l'art sur les cadres et méthodologies d'évaluation des systèmes de dialogue est proposé en chapitre 4.

Dans la seconde partie de ce document, le paradigme *SIMDIAL* est détaillé. Le chapitre 5 expose les choix ayant mené à la conception du simulateur, par rapport aux notions présentées dans les chapitres précédents. Le chapitre 6 précise alors les modèles et algorithmes employés pour réaliser le simulateur, ainsi qu'une analyse du coût de transformation de ce dernier pour évaluer une tâche différente de celle initialement prévue. Enfin la présentation des résultats de deux évaluations réalisées grâce au simulateur sont présentées en chapitre 7.

Les conclusions et perspectives sont enfin exposées en chapitre 8.



Première partie

Dialogue homme-machine et  
évaluation



# Chapitre 2

## Le dialogue homme-machine

Voir les machines converser naturellement et intelligemment avec les humains est, outre un rêve illustré par les romans et films de sciences-fiction, un domaine de recherche scientifique fécond ces cinquante dernières années. Ce domaine tient une place particulière dans le domaine de l'Intelligence Artificielle en ce que si l'ordinateur s'arroge la prérogative humaine du langage et se met à dialoguer, c'est qu'il est en passe de devenir intelligent [Moor, 1976].

Cette dernière idée est la clé de voûte du fameux test de Turing [Turing, 1950] qui consiste à faire dialoguer un juge avec deux sujets via un terminal d'ordinateur. Sachant qu'un des sujets est humain et que l'autre est un ordinateur, le juge cherche à démasquer l'ordinateur qui, s'il est « intelligent », fera illusion.

### 2.1 Bref historique

Dans sa thèse de doctorat, Laurent Charnay [Charnay, 1999] propose, grâce à l'analyse de trois systèmes de dialogue Homme-Machine (SDHM) historiquement archétypiques, à savoir Eliza [Weizenbaum, 1966], Shrdlu [Winograd, 1972] et GUS [Bobrow et al., 1977], une explication de l'évolution du dialogue Homme-Machine des années quatre-vingt à nos jours. Selon lui les enjeux auxquels est confronté le dialogue Homme-Machine en tant que domaine de recherche sont de trois catégories :

- **Une problématique de la simulation** qui vise à faire croire à un utilisateur qu'un système le comprend.
- **une problématique de la représentation** en ce qu'un système manipule des représentations pour simuler une compréhension qui se



concrétise dans sa réponse.

- **une problématique de l’interaction** puisqu’un système est censé interagir avec un humain de telle sorte que ce dernier soit satisfait.

Du point de vue de la simulation, la problématique est toujours de rigueur. En effet un ordinateur ne manipule que des représentations symboliques qui n’ont pas de sens intrinsèque. La soi-disant compréhension des demandes des utilisateurs n’est en fait qu’une simulation de cette compréhension.

Par ailleurs les progrès accomplis dans les représentations utilisées par les SDHM ainsi que dans la modélisation de l’interaction permettent aujourd’hui la réalisation de systèmes avec lesquels il est possible de résoudre des tâches par le dialogue. Ces systèmes sont généralement cités comme « systèmes de dialogue finalisé » en ce qu’ils ont pour finalité de résoudre la tâche objet du dialogue.

## 2.2 Différents niveaux d’interaction

Les avancées technologiques de ces dernières années ont notamment permis une progression quant aux capacités d’interaction des systèmes de dialogue. En effet les systèmes cités précédemment en 2.1 ne dialoguaient que par écrit alors qu’aujourd’hui les SDHM peuvent utiliser des composants de reconnaissance vocale capables de transcrire de la parole continue<sup>1</sup> pour des locuteurs différents.

Les grandes étapes par lesquelles sont passés les SDHM du point de vue de leurs capacités d’interaction sont décrites ci-dessous.

### 2.2.1 Interaction par écrit

Les premiers SDHM utilisaient uniquement la modalité écrite pour interagir avec l’utilisateur. C’était à l’époque la seule modalité possible à mettre en œuvre, les systèmes de transcription automatique de la parole (reconnaissance vocale) n’existait pas encore (ou n’étant pas assez performants). L’évolution des SDHM vers d’autres modalités d’interaction est d’une part due au fait que la modalité usuelle de dialogue pour un humain est la parole, et d’autre part à l’impossibilité d’exploitation commerciale de tels systèmes du temps où l’ordinateur personnel et l’Internet n’étaient pas de mise. En effet une des applications commerciales majeures des SDHM passe par le

---

<sup>1</sup>Le système transcrit pendant que le locuteur parle

média téléphonique. Aujourd'hui l'explosion du dialogue par écrit via Internet donne un second souffle à cette modalité.

### 2.2.2 DTMF

Systèmes et utilisateurs interagissent via un téléphone avec des modalités différentes. Le système produit des énoncés au format sonore grâce à des mécanismes de synthèse vocale ou des énoncés pré-enregistrés alors que l'utilisateur se sert des touches du téléphone pour choisir parmi une liste de propositions ou fournir un numéro de compte par exemple. Cette modalité a longtemps été et est toujours utilisée dans les services vocaux interactifs (SVI). En effet le DTMF permet une interaction homme-machine par téléphone sans nécessiter de système de reconnaissance vocale pas forcément fiable. De plus la restriction des choix possibles pour l'utilisateur à ceux qui sont gérés par le système facilite la conception des SVI et les rend plus fiables.

L'interaction était alors assez éloignée de ce qu'il est courant de nommer « dialogue ».

### 2.2.3 Interaction vocale contrainte

Pour les systèmes de reconnaissance vocale, le nombre de mots reconnus est un critère de performance important. Plus ce nombre est petit, mieux les mots appartenant au vocabulaire sont reconnus. Par ailleurs les taux d'erreurs de reconnaissance vocale ont longtemps affaibli la confiance donnée aux systèmes de transcription automatique de la parole. En effet pour un énoncé reconnu, il était difficile d'affirmer que c'était effectivement celui qu'avait prononcé l'utilisateur.

Pour ces raisons les premiers SDHM exploitant la reconnaissance vocale étaient équivalents aux systèmes DTMF à ceci près que les choix de l'utilisateur étaient communiqués vocalement plutôt qu'avec des touches. L'interaction était vocale et contrainte à ce que le système proposait. L'utilisateur n'ayant jamais l'initiative de ce qu'il allait pouvoir dire, ces SDHM sont dits à initiative système (voir 2.3.2).

### 2.2.4 Interaction vocale naturelle

Les progrès en reconnaissance vocale permettent aujourd'hui de reconnaître de la parole continue sur des vocabulaires de grande taille (d'environ 3000 mots à plus de 60000 mots dont des formes fléchies [L. Lamel, 1995]) avec des taux d'erreurs suffisamment faibles pour que les énoncés reconnus

puissent être exploités de façon fiable. Les SDHM ont tiré parti de cette progression en permettant aux utilisateurs de faire leurs requêtes « de façon naturelle », autrement dit comme s'ils s'adressaient à un humain. Évidemment le système peut guider l'utilisateur ou lui faire des propositions auxquelles il pourra répondre, mais globalement l'utilisateur n'est pas contraint dans sa réalisation de la tâche.

### 2.2.5 Interaction multimodale

Partant de l'idée qu'une interaction entre un humain et un système peut ne pas être réduite à une seule modalité, qu'elle soit vocale, écrite, visuelle ou autre, une communauté de recherche explore depuis quelques années l'interaction multimodale. La multimodalité peut être définie comme l'utilisation conjointe de plusieurs modalités, de façon complémentaire, redondante ou encore en opposition.

Par exemple pour informer quelqu'un de la localisation de la ville de Paris en France, l'utilisation conjointe de modalités visuelles comme la présentation d'une carte de la France et le pointage de la ville de Paris est plus efficace que d'utiliser la modalité vocale afin de décrire où la ville se situe.

Cette façon d'interagir avec le système est très prometteuse quant aux possibilités d'interaction imaginables pour le futur. Il n'en reste pas moins vrai que la façon de présenter ou de recevoir des informations sur plusieurs modalités ainsi que la représentation de l'information face aux modalités qui la véhiculent sont encore des sujets de recherche en devenir [Rouillard and Caelen, 2001, Horchani et al., 2007].

## 2.3 Caractéristiques principales du dialogue spontané

En premier lieu une précision sémantique s'impose : le dialogue spontané dont il est question ici caractérise la *production langagière spontanée en contexte de dialogue*, et pas le fait de dialoguer spontanément.

La production langagière en contexte de dialogue est généralement spontanée, à l'oral comme à l'écrit. En effet il est rare de préparer précisément ses phrases avant un dialogue. Cette spontanéité de la production langagière explique la présence de phénomènes que la préparation des phrases cherche justement à éviter. Ceci dit il paraît clair que les SDHM ont pour objectif de résoudre une tâche par le dialogue spontané oral comme écrit. En effet la communication spontanée synchrone par écrit grâce aux outils de messagerie instantanée est de plus en plus utilisée aujourd'hui, particulièrement par les

jeunes. Il n'en reste pas moins vrai que la nécessité de préparer son discours, oral ou écrit, avant de communiquer avec un SDHM a de quoi rebuter plus d'un utilisateur !

### 2.3.1 Au niveau de l'énoncé

Le Groupe de Travail « Compréhension Robuste » du Groupe De Recherche « Information - Interaction - Intelligence » (GT 5.5 GDR I3) qui a fait suite à l'action de recherche concertée B2 AUPELF-UREF [Caelen et al., 1997, Mariani, 1998] propose dans [Antoine et al., 2002a] la liste suivante de phénomènes, observables jusqu'au niveau de l'énoncé.

#### **Erreurs de reconnaissance :**

Dans le cas de systèmes à modalité vocale, ce phénomène caractérise les erreurs de reconnaissance vocale. Pour la modalité écrite il représente alors les mots mal orthographiés et autres fautes de frappe.

#### **Énoncés complexes :**

Utilisables pour les modalités écrites comme orales, ils caractérisent à la fois les énoncés contenant des requêtes multiples<sup>2</sup> et les objets structurellement complexes de la langue comme par exemple les négations, les coordinations.

#### **Procédés du discours spontané :**

Cette catégorie comprend, pour la modalité vocale, les phénomènes propres au langage oral, parmi lesquels les hésitations, les répétitions, les auto-corrrections, les incises<sup>3</sup>. Le discours spontané écrit comprend notamment les abréviations, les omissions d'articles, ainsi que le « langage SMS ».

#### **Dislocations :**

À l'oral ce phénomène caractérise les énoncés dont l'ordre des mots ne respecte pas une syntaxe classique de l'écrit. Pour la modalité écrite les dis-

---

<sup>2</sup>Plusieurs demandes en un seul énoncé

<sup>3</sup>Lorsqu'un mot ou groupe de mot est intercalé dans un énoncé sans nécessairement avoir de rapport avec le sens initial de cet énoncé.

locations ne sont pas à observer au niveau des mots mais des lettres dans les mots.

### Problèmes de couverture :

Ce phénomène, défini quelque soit la modalité, est associé à la notion de *couverture du domaine*. Pour dialoguer sur un domaine, il est nécessaire d'avoir connaissance des concepts associés à ce domaine. L'ensemble des concepts manipulables par un système pour un domaine est nommé sa couverture. Les problèmes de couverture caractérisent ici la précision ainsi que la complétude de la couverture d'un domaine par les systèmes.

### 2.3.2 Au niveau du dialogue

Contrairement aux caractéristiques du niveau de l'énoncé qui définissent des phénomènes statiques vis-à-vis de l'interaction, les caractéristiques du niveau dialogique sont liées à la dynamique et à la co-construction du dialogue. Le terme couramment utilisé pour décrire les phénomènes dynamiques dans la littérature est celui de *stratégie*. La liste présentée ci-dessous reprend un ensemble de stratégies existantes pour le dialogue homme-machine.

#### L'initiative

Elle est primordiale pour expliquer le déroulement d'un dialogue. Elle représente, à chaque tour de parole, le degré de liberté laissé à celui qui va s'exprimer par celui qui vient de s'exprimer. L'initiative est dite centrée sur un des participants si celui-ci guide l'interaction. Si les deux participants se laissent la possibilité de guider l'autre, l'initiative est dite mixte.

Il est important de ne pas confondre l'initiative et l'alternance des tours de parole. Par exemple si  $\alpha$  vient de demander à  $\beta$  l'heure qu'il est, c'est à  $\beta$  de prendre la parole mais c'est  $\alpha$  qui a l'initiative puisque c'est lui qui a initié l'échange.

Il est généralement considéré trois degrés possibles d'initiative pour un SDHM. Il est question d'*initiative système* quand le système pose une séquence précise de questions et n'attend de l'utilisateur que des réponses à ses questions. Dans le cas d'*initiative utilisateur*, l'utilisateur ne peut que poser des questions au système et ce dernier ne fournit de réponse qu'à la question posée sans demander plus de détail. Enfin l'*initiative mixte* est un partage généralement coopératif de l'initiative entre utilisateur et système, dans le but d'atteindre l'objectif de l'utilisateur. Localement aux tours de

parole l'initiative est toujours soit utilisateur, soit système, mais elle change au cours des situations de dialogue.

### La confirmation

Les erreurs étant fréquentes lors de l'activité de dialogue, la *confirmation* de ce qui a été compris à un instant donné s'avère d'importance. Ce point est particulièrement vrai pour les SDHM vocaux, leur reconnaissance vocale et leur interprétation sémantique étant moins « performantes » que celles d'un humain. De plus l'humain a des stratégies de confirmation qui dépassent de loin les possibilités des SDHM, notamment grâce à la perception d'indices non-verbaux. L'étude de tels indices, comme par exemple la prosodie des locuteurs [Litman and Forbes, 2003] ou encore les émotions décelables dans la voix [Devillers et al., 2005], est en émergence ces dernières années.

Dans le cadre du dialogue homme-machine, la confirmation passe nécessairement par le dialogue. Elle est dite *explicite* quand le système demande à l'utilisateur si l'information transmise a bien été comprise. Quand le SDHM inclut l'information à confirmer dans un énoncé contenant d'autres informations, la confirmation est dite *implicite*. Cette dernière stratégie suppose d'une part que l'utilisateur corrige le système s'il perçoit que celui-ci a mal compris, et d'autre part que le système soit à même de comprendre et de tenir compte de corrections proposées par l'utilisateur.

### Le relâchement de contraintes

Il s'applique lorsqu'une requête n'a pas de solution. La réaction basique à une requête sans solution est de proposer une nouvelle recherche. Le relâchement de contraintes ou relaxation est une autre solution qui est généralement plus coopérative et moins coûteuse. Elle consiste à affaiblir ou relâcher les contraintes sur la requête déjà réalisée.

Différentes façons de relâcher les contraintes sont alors envisageables, comme par exemple de demander à l'utilisateur s'il a des préférences quant au critère à changer, ou de proposer des solutions estimées « proches » de celle recherchée. Le choix entre ces méthodes dépend alors de l'espace des solutions possibles face à la requête de l'utilisateur ainsi que de la pertinence des solutions approchées du point de vue des utilisateurs.

### La désambiguïsation

Aussi nommée clarification, elle intervient dans les cas où plusieurs possibilités sont envisageables face à une même information. Par exemple lors-

qu'un énoncé contient deux informations différentes alors qu'une seule est attendue, quand deux interprétations sont possibles pour un même mot.

Comme pour la relaxation, plusieurs possibilités sont exploitables, d'une part en fonction des cas à clarifier et d'autre part en fonction de la façon dont la clarification veut être effectuée. Il est par exemple possible de choisir une des informations ambiguës et d'en faire une confirmation implicite ou de demander explicitement un choix entre ces informations.

## 2.4 Les systèmes de dialogue homme-machine

Comme présenté en section 2.2, différents modes d'interaction sont possibles en fonction des SDHM. Le travail réalisé pour cette thèse s'applique à des SDHM utilisant la modalité orale ou écrite. La majorité des SDHM existants utilise l'une ou l'autre de ces modalités et l'étude des comportements de SDHM face à la production discursive spontanée est un challenge actuel du domaine.

### 2.4.1 Architecture standard

A l'heure actuelle il est courant de représenter un SDHM vocal de façon modulaire et fonctionnelle en distinguant les composants de reconnaissance et de synthèse vocale<sup>4</sup>, d'interprétation sémantique et de génération en langue naturelle, ainsi que de gestion de l'interaction (voir figure 2.1).

1. Le rôle du module de reconnaissance vocale est de transcrire le signal de parole de l'utilisateur en une chaîne de caractères correspondant à ce que l'utilisateur a dit. D'autres données peuvent aussi être produites par ce composant comme des informations sur la prosodie ou sur le taux de confiance accordé à la transcription réalisée. Dans ce cas le résultat produit par le composant n'est pas au format « texte simple » mais dans un formalisme spécifique, comme par exemple des treillis de mots.
2. Le module de synthèse vocale a le rôle symétrique de celui de la reconnaissance vocale. Il transforme une chaîne de caractères en signal sonore compréhensible par un utilisateur. De façon symétrique au composant de reconnaissance vocale, des données peuvent être fournies au composant de synthèse vocale pour spécifier des comportements vocaux, comme des changements de rythme, de prosodie etc.

---

<sup>4</sup>Ces composants ne sont évidemment pas présents dans les SDHM à modalité écrite.

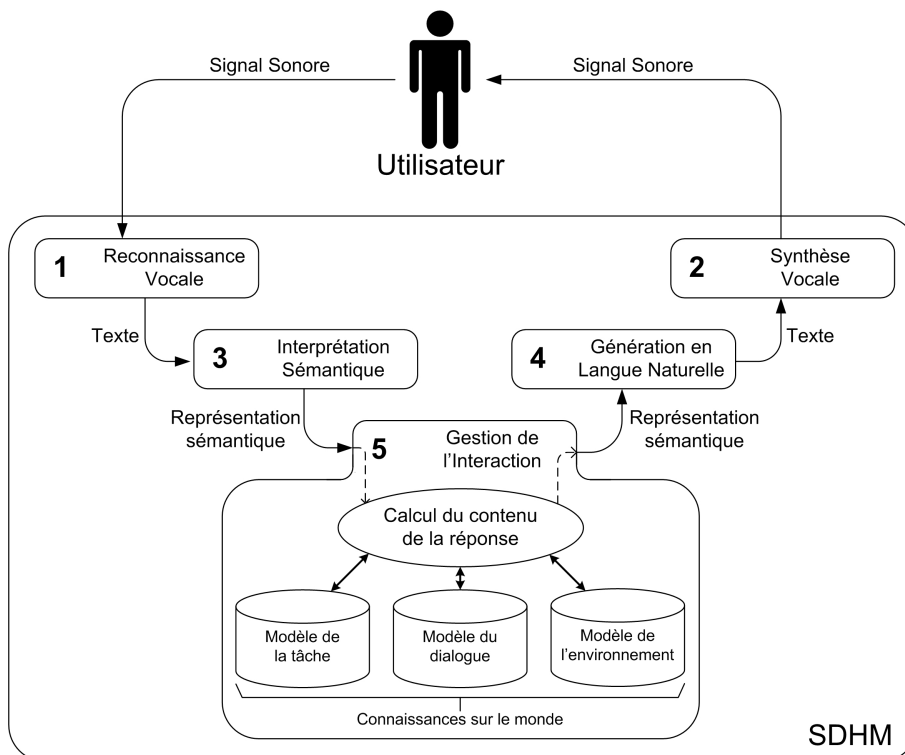


FIG. 2.1 – Représentation fonctionnelle standard d'un SDHM

- Le module d'interprétation sémantique peut-être considéré comme un traducteur automatique. Il reçoit en entrée un énoncé produit en langue naturelle au format texte<sup>5</sup> et doit fournir en sortie le même énoncé traduit dans un langage formel de représentation de contenu d'énoncés. Il est à noter que le formalisme de représentation de contenu d'un énoncé conditionne de façon importante les interprétations réalisées (voir 5.2.2).
- Comme la synthèse vocale face à la reconnaissance vocale, le module de génération en langue naturelle a la fonction symétrique du module d'interprétation sémantique. En effet son rôle est de transformer un énoncé produit dans un formalisme de représentation de contenu en un énoncé en langue naturelle. Là encore le formalisme de représentation de contenu d'énoncés influe fortement sur les possibilités en terme de verbalisation en langue naturelle.

<sup>5</sup>Ou dans le formalisme de sortie de la reconnaissance vocale.



5. Le module de gestion de l'interaction est la pièce centrale du SDHM (voir section suivante). Il doit fournir le contenu de la réponse du système dans le formalisme de représentation de contenu d'énoncés. Pour ce faire il reçoit en entrée le contenu de l'énoncé de l'utilisateur déterminé par l'interprétation sémantique. Le calcul de la réponse est alors effectué en fonction de ce contenu, du contexte d'interaction ou historique du dialogue, qui contient les énoncés déjà produits, et des « connaissances sur le monde »<sup>6</sup>.

Ces dernières sont des données utilisables par le composant de gestion de l'interaction pour choisir le contenu de sa réponse. Elles sont généralement regroupées en trois modèles, celui de la tâche, du dialogue, et de l'environnement. Ces modèles sont distingués du composant de calcul de la réponse d'une part afin de rendre ce dernier plus générique face à eux et d'autre part pour mieux différencier les connaissances utilisées et les façons de les traiter.

## 2.4.2 Modélisation de l'interaction

Depuis trente ans différentes approches ont été utilisées tant pour modéliser l'interaction que pour représenter de façon adéquate les connaissances nécessaires à ces modélisations, comme par exemple dans [Rosset, 2001, Sadek et al., 1997, Rosset et al., 2005, Nguyen and Caelen, 2003, on the France Telecom 3000 Voive Agency Corpus, 2007]. Une description plus précise des modèles présentés ci-après est disponible dans [Pietquin, 2004].

### Modèle de la tâche :

C'est une représentation formelle de la tâche que cherche à résoudre le SDHM. La façon dont elle est représentée est souvent liée à l'approche choisie pour calculer la réponse du système.

Une façon de représenter les tâches a été proposée dans Minsky [Minsky, 1975] et est fondée sur le concept de « *frame* », qui est une structure de données représentative d'une situation stéréotypique. Une autre approche présentée dans [Smith and Hipp, 1994], utilise des axiomes comme formalisme de représentation, ce qui est utile quand le calcul de la réponse est réalisé par déduction logique. La tâche peut aussi être représentée comme dans [Walker et al., 1997c], par une matrice d'attributs - valeurs (voir 4.3.1) où l'information qui doit être échangée entre l'utilisateur et le système est représentée par des paires ordonnées d'attributs et de valeurs correspondantes.

---

<sup>6</sup>Sont parfois distingués le gestionnaire d'interaction et le générateur de contenu sémantique.

Il est à noter que sont parfois séparées les représentations de la tâche à résoudre de celles du domaine d'application de cette tâche. Par exemple la tâche peut être une recherche d'information et le domaine les restaurants parisiens.

### **Modèle du dialogue :**

C'est un formalisme qui vise à représenter les mécanismes en œuvre lors du déroulement d'un dialogue. Il dépend généralement d'un modèle de la tâche et souvent les représentations qu'il manipule sont des actes de dialogues ou actes de langages.

La modélisation du dialogue a emprunté chronologiquement différentes techniques. La plus ancienne [Reichman, 1981] est l'utilisation de grammaires sur les actes dialogues, considérant qu'il existe des régularités dans les séquences d'actes échangés. Une méthode plus récente [Carberry, 1990] est l'utilisation de modèles basés sur les plans, qui considèrent que les actes de dialogues font partie du plan des participants. Une autre méthode est fondée sur la théorie conversationnelle des jeux qui essaie de regrouper les idées des deux approches précédemment citées [Power, 1979]. Les échanges sont alors guidés à la fois par des règles et par le but général du participant. Une dernière approche est de considérer le dialogue comme une activité conjointe des participants plus que comme les activités dissociées de chacun d'eux [Edmonds, 1993]. Elle exploite la notion d'accord mutuel pour expliquer les clarifications et confirmations dans le dialogue.

### **Modèle de l'environnement :**

Il contient des paramètres sur l'environnement du gestionnaire d'interaction. Il devrait par exemple contenir des informations sur les erreurs potentielles des autres composants du SDHM, pour pouvoir les gérer. De même des informations sur les utilisateurs, parfois nommées « modèle utilisateur » [Fischer, 2001], sont contenues dans cet ensemble de paramètres. Par exemple le degré d'expertise par rapport à l'utilisation du système est une information pertinente pour un meilleur comportement du système face aux utilisateurs. Enfin des événements extérieurs au système peuvent modifier son état interne, comme par exemple des mises à jour de bases de données, ou des fluctuations de marché. Bien qu'indépendant de la tâche, le modèle de l'environnement peut modifier l'état interne et donc le comportement du système.

### Calcul du contenu de la réponse :

L'utilisation des modèles décrits ci-dessus pour générer le contenu de la réponse par le gestionnaire d'interaction peut prendre différentes formes.

Une des approches est d'utiliser des algorithmes de preuve de théorème [Smith and Hipp, 1994], exploitant généralement la programmation logique. L'inconvénient majeur de ces approches reste leur rigidité face aux stratégies mises en œuvre. Une autre méthode plus utilisée pour calculer le contenu des réponses exploite des automates à états finis. Le comportement complet du gestionnaire d'interaction est alors défini en termes d'états et de transitions entre ces états [McTear, 1998]. L'inconvénient est ici la nécessité de description complète de tous les états du gestionnaire d'interaction. Une troisième possibilité, dite de *remplissage de formulaire*, exploite une représentation de la tâche dans un format équivalent à une matrice d'attributs-valeurs [Devillers et al., 2003]. Les réponses sont alors générées en fonction des cases vides ou pleines de la matrice. La limite de cette approche est de guider le dialogue uniquement par la tâche, ce qui restreint la gestion de phénomènes inattendus comme les répétitions par exemple. Une dernière méthode nommée auto-organisée permet au gestionnaire d'interaction d'adapter ses comportements en fonction de représentations internes qu'il construit [aust and Schroer, 1998]. La difficulté réside alors dans l'impossibilité d'expliquer certains comportements.

### 2.4.3 Défis scientifiques

Parmi le grand nombre de questions relatives aux SDHM oraux, les deux présentées ci-après sont considérées comme des problèmes majeurs et des questions ouvertes.

#### La généralité

Bien que des SDHM aient été implémentés pour de nombreuses applications différentes, la tâche d'adaptation d'un système vers un nouveau domaine d'application reste difficile, surtout du point de vue du gestionnaire d'interaction. En effet, chaque application d'un système à une nouvelle tâche nécessite la définition des réactions à générer en fonction des situations observées. Évidemment, ce travail est plus ou moins fastidieux en fonction d'une part des différences entre les tâches, et d'autre part de la difficulté d'adaptation du système considéré. D'où des efforts pour chercher à rendre plus génériques les systèmes face aux tâches qu'ils permettent de résoudre.

## L'évaluation

Il n'existe pas à l'heure actuelle de méthodologie d'évaluation standard pour les SDHM, pas plus qu'il n'existe de critère d'évaluation standard. Différentes campagnes, comme le projet francophone AUF- Arc B2 [Mariani, 1998], l'évaluation par DEFI [Antoine et al., 2002b], les projets européens EAGLES [King et al., 1996], DISC [DISC Consortium, 1999], SUNDIAL [Gibbon et al., 1997] ainsi que les projets ATIS [Hirschman et al., 1992], ont été menées, qui eurent plus ou moins de succès en fonction des résultats qu'elles permettaient d'obtenir. De même des critères ont été proposés lors de ces évaluations, dont il est aujourd'hui reconnu que leurs résultats ne suffisent pas à caractériser les SDHM de façon souhaitée. La communauté de recherche relative à l'évaluation des SDHM continue de proposer des paradigmes et des critères en fonction des besoins de chacun mais la définition d'un standard n'est pas encore effective.

## 2.5 Conclusion

La réalisation d'un système informatique à même de dialoguer avec un humain pour résoudre une tâche met en jeu un très grand nombre de notions. Elles peuvent très « bas-niveau », comme par exemple les problèmes acoustiques liés au canal téléphonique ou encore la reconnaissance prosodique pour l'interprétation d'émotions, ou très « haut-niveau », comme l'interprétation du locuteur ayant l'initiative sur le tour de parole ou encore la nécessité de désambiguïser certaines interprétations.

Un système de dialogue est donc dépendant, sinon de toutes, au moins d'un nombre certain de ces notions, ce qui le rend complexe tant à réaliser qu'à maintenir et à faire évoluer. Dès lors la notion d'évaluation paraît primordiale pour l'avancée de ces systèmes. Le chapitre suivant expose donc la problématique de **l'évaluation de SDHM**. En premier lieu seront présentés des concepts généraux sur l'évaluation de systèmes, suivies en section 3.2 des différentes approches généralement employées pour évaluer. Enfin la dernière partie présentera les caractéristiques spécifiques de l'évaluation des systèmes de dialogue.



# Chapitre 3

## L'évaluation de systèmes

Les récentes campagnes et projets tels que DARPA MADCOW [Hirschman et al., 1992], B2 AUPELF-UREF [Mariani, 1998], DARPA Communicator [Walker et al., 2000], Technolanguage MEDIA [Devillers et al., 2004], témoignent de l'intérêt croissant pour l'évaluation des systèmes de dialogue et de leurs composants. Ce sujet est d'importance tant pour le monde industriel que pour la communauté de recherche sur le traitement automatique des langues. Il est aussi d'importance pour ses apports tant des points de vue méthodologique que de la recherche appliquée. À l'heure actuelle il n'existe toujours pas de paradigme ni de mesures communément admis pour évaluer les systèmes de dialogue Homme-Machine.

### 3.1 Généralités

#### 3.1.1 Définition

L'évaluation est une activité qui consiste à **donner une valeur à l'objet sur lequel elle porte.**

Elle est réalisée de façon quasi permanente et souvent inconsciente par tout un chacun. En effet elle est présente à chaque fois qu'il est nécessaire de juger de la valeur de quelque chose, des situations les plus anodines aux plus importantes. Le fait même de juger une situation comme importante ou anodine relève de l'évaluation. De plus cette activité s'applique aussi bien à des objets concrets qu'à des concepts abstraits.

Par exemple un piéton qui veut traverser une route n'a généralement pas conscience qu'il évalue la vitesse et la distance qui le sépare de la voiture qui arrive. De même un musicien accordant son instrument évalue la différence

de fréquence entre deux notes, le professeur évalue le savoir ou savoir-faire de l'élève.

La nécessité d'évaluation dans des domaines tels que l'éducation ou la recherche a ouvert la voix de la clarification de cette activité trop souvent considérée comme allant de soi.

### 3.1.2 Buts, valeurs et attributions

L'activité d'évaluation consiste en l'attribution d'une valeur à une entité. Elle nécessite donc un ensemble de valeurs possibles et une méthode pour attribuer une de ces valeurs à l'entité évaluée. De plus les valeurs utilisées pour caractériser l'entité évaluée peuvent être représentatives de données concrètes ou pas.

Par exemple pour évaluer le poids de quelqu'un, un ensemble de valeurs généralement utilisé est celui des entiers naturels associé à une unité de poids. Une méthode d'attribution de la valeur de poids peut alors être de mesurer celui-ci à l'aide d'un pèse-personne.

Le choix d'un ensemble de valeurs et d'une méthode d'attribution pour réaliser une évaluation est très dépendant du but guidant celle-ci. En effet dans le cas d'une personne souhaitant sauter en parachute, l'évaluation de son poids n'est nécessaire que pour valider que ce dernier ne dépasse pas celui supporté par le matériel de saut. L'évaluation porte alors sur le dépassement ou non d'une valeur seuil pour le poids de la personne, et pas sur le poids en lui même. Dans ce cas l'ensemble des valeurs d'évaluation peut-être restreint aux deux valeurs « oui » ou « non ».

Cet exemple montre la nécessité d'explicitation précise des objectifs, ensembles de valeurs et méthodes d'attribution choisis lors de la définition ou de l'exposition d'une évaluation.

### 3.1.3 Observation et jugement

Concernant la méthode d'attribution d'une valeur à un objet, il est important de préciser qu'à moins d'avoir des connaissances préalables ou de porter un jugement *a priori*, cette activité n'est pas possible sans observation. En effet c'est l'observation de l'objet qui permet d'obtenir les informations nécessaires au jugement et donc à l'évaluation.

La méthodologie d'évaluation se compose donc de deux sous-activités

distinctes à savoir l'observation de l'objet et le jugement (ou qualification) des résultats obtenus par l'observation. Par exemple l'observation du savoir d'un élève par un devoir sur table permet son jugement en une note.

La distinction entre observation et jugement, tant au niveau des ensembles de résultats d'observation et de jugement que des méthodes d'observation et de jugement, est fondamentale pour la bonne définition de ce qui est évalué et de la façon de l'évaluer.

### 3.1.4 Objectif vs. subjectif

Bien que les notions de résultats *objectifs* et *subjectifs* soient souvent employées dans le cadre de l'évaluation, leur définition est rarement explicite.

La façon implicite et usuelle de définir ces notions passe par la séparation des résultats obtenus par des sondages sur les utilisateurs, dits subjectifs, de ceux obtenus autrement, dits objectifs.

Cette définition est précise pour ce qui est des résultats subjectifs, mais concernant les résultats objectifs, la définition par opposition à l'ensemble des résultats subjectif semble trop évasive.

La notion de résultat **objectif** est liée à l'accord inter-personnel sur la définition de ce qui est observé. Par exemple il est possible de dire que le poids du sauteur en parachute est un résultat objectif car la définition de ce qu'est le poids est la même pour tout un chacun, autrement dit qu'il existe une unité de mesure partagée pour le poids.

De plus cette idée est consistante avec la notion d'accord inter-annotateur utilisée pour démontrer la validité « objective » d'annotations. En d'autres termes le fait qu'un ensemble d'annotations puisse être utilisé de façon objective face à des données est démontré par un calcul d'accord entre plusieurs personnes ayant utilisé cet ensemble d'annotations sur un même ensemble de données.

Il est enfin à noter que cette définition précise celle communément admise sans s'y opposer.



## 3.2 Différents points de vue

### 3.2.1 Approches ergonomiques et technologiques

Les approches dont il question ici exploitent les notions de résultats objectifs ou subjectifs explicitées précédemment en 3.1.4.

Les *approches technologiques* sont généralement fondées sur l'observation de facteurs objectifs car non dépendants de la perception des utilisateurs. Certains de ces facteurs sont spécifiques aux systèmes considérés et d'autres définis de façon plus générale par les communautés de recherche. Par exemple le nombre de tours de parole d'un dialogue est un facteur dont la définition est communément admise.

A l'inverse les approches ergonomiques exploitent souvent des résultats fournis par des utilisateurs. Ces résultats portent sur des facteurs dont la définition peut parfois permettre des observations objectives. Par exemple la comparaison entre deux systèmes sur la longueur de leurs énoncés est observable objectivement alors que la satisfaction des utilisateurs à utiliser le système n'est observable que de façon subjective. En effet bien que certains phénomènes observables objectivement soient reconnus ayant un impact sur la satisfaction des utilisateurs, le manque d'accord sur une définition de la satisfaction utilisateur observable objectivement la rend nécessairement subjective.

Une solution possible pour exploiter conjointement les travaux réalisés par les deux communautés est de chercher les éléments communs à l'une et l'autre. De façon assez naturelle les facteurs utilisés pour des évaluations ergonomiques et utilisables objectivement pour des évaluations technologiques semblent être de bons candidats.

Un point important à noter concerne l'idée que l'observation objective de résultat rende caduque leur observation subjective. Bien qu'intuitivement raisonnable, cette idée est à rejeter. En effet il est tout à fait possible, face à un même observable, qu'un système soit objectivement meilleur qu'un autre mais pourtant moins bien perçu par les utilisateurs.

Une dernière précision à apporter concerne les facteurs utilisés pour des évaluations ergonomiques et dont la définition ne permet pas l'observation objective. L'incapacité à fournir objectivement des résultats sur ces facteurs est liée au fait qu'ils ne soient pas encore suffisamment bien modélisés de façon objective. De nombreux travaux comme par exemple la modélisation

des émotions [Devillers, 2006] ouvrent la voix de l'objectivation de facteurs aujourd'hui purement subjectifs.

### 3.2.2 Utilisateur et usager, utilisabilité et usabilité

Les définitions discutées ici s'appuient sur celles proposées par Stéphane Chaudiron dans [Chaudiron, 2001].

La distinction entre *utilisateur* et *usager* rend compte d'une différence de point de vue sur l'individu se servant d'un système. Un utilisateur est alors un individu qui se sert d'un système avec un but précis tandis qu'un usager est individu perçu comme un acteur social dans un contexte socio-technique.

Le concept d'utilisateur implique ici l'individu dans une relation « personnelle » avec la machine et comprend donc les notions ergonomiques individuelles telles que le niveau de compétence ou la charge cognitive. A l'inverse l'usager est perçu comme un représentant d'un contexte d'usage du système et pas en tant qu'individu pris isolément.

Cette distinction permet ainsi de clarifier les notions d'*utilisabilité* et d'*usabilité*, fréquemment confondues dans la littérature scientifique francophone.

L'utilisabilité représente alors la capacité de « maniement » de l'outil technique du point de vue de l'utilisateur tandis que l'usabilité traduit l'activité sociale que représente l'usage de cet outil. En reprenant les termes de Stéphane Chaudiron, l'usabilité désigne « la capacité d'un système à être effectivement utilisé dans des conditions satisfaisantes et pour l'accomplissement de tâches définies par l'utilisateur ».

### 3.2.3 Évaluation qualitative et quantitative

L'*évaluation qualitative* d'un système fournit un jugement de valeur sur le fonctionnement du système en fonction de concepts utilisés dans le domaine d'application du système.

L'*évaluation quantitative* repose sur l'analyse de critères mesurables objectivement, ou *métriques*, pour caractériser le fonctionnement d'un système.

Les différences majeures entre ces approches résident dans :

- la subjectivité des concepts utilisés pour une évaluation qualitative face à l'objectivité des résultats d'une évaluation quantitative.

- le caractère explicatif d'une évaluation qualitative par rapport au caractère descriptif d'une évaluation quantitative.
- l'objectif de découverte d'hypothèses pour une évaluation qualitative face à la validation ou preuve d'hypothèses pour l'évaluation quantitative.

L'utilisation de l'une ou l'autre de ces approches, ou encore d'une approche mixte ou hybride, est fortement liée au degré d'avancement dans la modélisation du domaine d'application du système à évaluer. L'évolution de la modélisation d'un domaine commence généralement par la modélisation qualitative par des concepts abstraits générant des hypothèses, suivie par des (in)validations quantitatives de ces hypothèses par des métriques. Peut alors avoir lieu une hybridation des approches, exploitant des métriques quantitatives pour permettre des analyses d'ordre qualitatif.

### 3.2.4 Système perçu en boîte noire ou transparente

Un système perçu en *boîte noire* peut être représenté comme une entité qui, en fonction des données qui lui sont fournies en entrée « *input* », génère un comportement « *output* ».

En *boîte transparente*, le fonctionnement d'un système est perçu à travers le comportement des différents modules qui le composent.

Une évaluation en boîte transparente permet notamment des évaluations des différents modules du système évalué ainsi que des représentations utilisées par ces modules. Outre ces évaluations, elle permet aussi potentiellement de cibler les modules en cause dans un dysfonctionnement du système. Cette dernière propriété, dite de diagnostic, est très intéressante du point de vue du concepteur en ce qu'elle lui permet de cibler de quel(s) module(s) vien(nen)t les problèmes.

A l'inverse une évaluation en boîte noire porte sur les performances globales du système face à la tâche pour laquelle il a été conçu. Elle ne permet pas de diagnostic technique mais se prête particulièrement bien à la comparaison de systèmes d'architectures différentes. Enfin cette évaluation s'applique aussi très bien à des approches ergonomiques de l'évaluation en ce qu'elle permet de faire abstraction de toute modélisation technologique pour se centrer sur l'utilisateur.

### 3.2.5 Évaluation avec mesure ou référence

Dans le cadre de l'évaluation de systèmes en boîte noire, un système est une entité qui, à une situation, fait correspondre un comportement. De ce fait l'évaluation d'un système perçu en boîte noire correspond à l'évaluation des comportements du système face à des situations.

L'explicitation du fait que l'évaluation porte sur des comportements rend possible une distinction entre deux façons d'évaluer, selon la nécessité ou non de **représentation des comportements**. Lorsque l'évaluation a lieu en comparant les comportements d'un système évalué avec des comportements de référence (qu'ils soient « idéaux » ou pas), leur description est nécessaire. À l'inverse, l'évaluation par application de mesures aux comportements n'implique pas nécessairement cette description.

Il sera dans la suite du document question d'**évaluation avec référence** lorsque l'évaluation nécessite la description de comportements et d'**évaluation avec mesures** lorsque cette description n'est pas nécessaire. Par ailleurs l'évaluation par application de mesures à des comportements évalués et de référence est assimilable à une évaluation avec référence puisqu'elle nécessite la représentation des comportements.

### 3.2.6 Modélisation et évaluation

L'idée de prendre en considération le modèle du système évalué défini lors d'une évaluation provient de la communauté de recherche sur le diagnostic automatique. Cette communauté développe des méthodes permettant de diagnostiquer automatiquement si un système fonctionne bien ou pas. Pour ce faire, la méthode est d'utiliser un modèle souvent simplifié du système considéré, définissant soit le bon fonctionnement du système, soit ses cas de panne, soit les deux.

Dans le cas d'un système perçu en boîte noire, la distance à considérer met en jeu le modèle du système défini par l'évaluateur face au système à évaluer. Pour un système perçu en boîte transparente, ce sont les représentations associées aux modules du système évalué qui doivent être considérées.

Il est aussi à noter que la définition du modèle utilisé pour représenter le système évalué est plus facile dans le cas d'une évaluation par référence que dans le cas d'une évaluation par mesure. En effet, l'évaluation par référence nécessitant l'explicitation des comportements du système évalué et du système de référence, le modèle faisant référence est explicite. Dans le cas d'une

évaluation avec mesure, le modèle permettant l'obtention d'un résultat est implicite à la mesure, ce qui rend difficile sa caractérisation.

De façon générale, la conscience d'une différence potentielle entre le modèle de ce qui est évalué et sa réalité est primordiale pour évaluer l'évaluation, ou en d'autres termes pour caractériser la valeur de l'évaluation réalisée. En effet, les erreurs décelées par l'évaluation ne seront pas les mêmes en fonction des modèles utilisés pour représenter le système évalué.

Par ailleurs, pour l'évaluation ergonomique de systèmes, l'utilisation d'une même représentation pour évaluer et mettre en œuvre un système est une source d'erreurs de jugement. Par exemple dans le cas d'évaluations ergonomiques comparatives, il semble légitime qu'un système utilisant la même représentation que l'évaluateur ait de meilleurs résultats que les autres (voir 3.3.2).

### 3.3 L'évaluation de systèmes de dialogue

L'évaluation des systèmes de dialogues est un sujet difficile par essence vu l'aspect dynamique et co-construit du dialogue [Devillers, 2004]. Il est difficile aussi pour des questions de généricité et de pouvoir diagnostique des paradigmes d'évaluation. Enfin un système de dialogue finalisé ayant pour objectif premier la réalisation d'une tâche via une interaction avec un humain, l'étude des usages en situation de dialogue prend une place importante dans la problématique de l'évaluation de SDHM [Chaudiron, 2001, Chaudiron and Ihadjadene, 2002].

Dans cette section sont définis les différents points permettant de caractériser une évaluation de SDHM.

#### 3.3.1 Les buts de l'évaluation

Comme présenté en 3.1.2, la définition des buts d'une évaluation est fortement corrélée à la définition d'une part de l'ensemble des valeurs d'évaluation possible et d'autre part de la ou des fonctions d'attribution de ces valeurs. Le groupe de travail EAGLES (voir section 4.1) a proposé une catégorisation générale des buts d'évaluation de façon indépendantes des ensembles de valeurs et des fonctions d'attribution :

- **Évaluation de progrès** : Quelle est la « valeur » du système relativement à son concurrent ou à une version précédente ?

- **Évaluation diagnostique** : Ce but est de préciser d'où viennent (ou ne viennent pas) les problèmes<sup>1</sup>.
- **Évaluation d'adéquation** : Le système a-t-il atteint un objectif fixé ?

Un premier point à noter est que ces catégories ne sont pas disjointes, il est par exemple possible de réaliser des évaluations de progrès diagnostiques, ou encore des évaluations d'adéquation comparatives.

Par ailleurs avec seulement trois catégories la quasi totalité des buts d'évaluation généralement considérés sont couverts. En effet ces derniers contiennent souvent dans leurs définitions soit des précisions sur l'ensemble de valeurs d'évaluation utilisé, soit des précisions sur l'objet évalué, et sont contenus dans une des trois catégories du point de vue de leur objectif. Par exemple une évaluation de rentabilité est une évaluation d'adéquation (possiblement diagnostique) où l'objectif est un ratio.

Quelques précisions peuvent quand même être apportées à cette catégorisation. La liste qui suit présente des sous-buts d'évaluation ainsi que leur positionnement dans la catégorisation générale.

- *La non-régression* a pour objet de contrôler qu'une nouvelle version d'un système ne fonctionne pas moins bien que la précédente. Du point de vue de la catégorisation, cet objectif est à placer dans les évaluations de progrès, mais il est intéressant de noter qu'il définit en réalité l'**objectif dual** du progrès.
- *Les comparaisons inter-systèmes ou inter-versions* qui sont classées dans les évaluations de progrès mais qui diffèrent généralement beaucoup du point de vue des observations souhaitées lors de l'une ou l'autre de ces évaluations.
- *Les diagnostics modulaires ou fonctionnels* qui se différencient en ce que le premier concerne des composants réels du système évalué alors que le second exploite des modèles pour définir le bon ou mauvais fonctionnement du système relativement à ces modèles.

### 3.3.2 L'intimité avec le système à évaluer

L'intimité avec le système évalué est représentative de la connaissance de ce système en tant qu'évaluateur. Cette connaissance peut être d'ordre

---

<sup>1</sup>Les prédictions peuvent porter sur des sous-composants du système ou sur des comportements prédéfinis du système.

technique et/ou conceptuelle. La connaissance technique d'un système est représentée par les notions de boîte noire et transparente alors que sa connaissance conceptuelle est représentée par la proximité de modèle. Les sections suivantes précisent ces deux points.

### **Boîte noire et transparente**

Comme présenté en 3.2.4, il est possible pour évaluer de considérer le système à différents niveaux d'abstraction fonctionnelle, soit comme une boîte noire dont seules les entrées/sorties sont considérées, soit comme une boîte transparente où sont perçus les modules le composant. Dans ce dernier cas, différents niveaux de finesse d'observation sont possibles, comme par exemple les entrées/sorties des modules ou des traces internes de comportements ou de traitement.

Ramenée à l'évaluation de SDHM, cette différence de perception revient à évaluer soit le système de dialogue complet grâce aux interactions menées avec lui, soit chacun des sous-composants du système<sup>2</sup>, avec pour chacun une méthode et un contexte d'évaluation propre.

Dans le cadre de l'évaluation de SDHM, la perception du système en boîte noire est utilisée généralement par les approches ergonomiques ainsi que par quelques approches technologiques. La perception du système en boîte transparente est quant à elle généralement utilisée par des approches technologiques pour évaluer un ou des composants du SDHM (voir 4).

### **Proximité de représentations**

Cette idée a été explicitée en 3.2.6 dans son cadre général. Ramenée à l'évaluation de SDHM, elle est notamment à prendre en compte lors d'évaluations ergonomiques ainsi que lors d'évaluation hybrides ergo-technologiques portant sur des critères non technologiques.

Par exemple il paraît naturel d'évaluer un système de dialogue par une approche technologique face à l'ensemble précis des fonctionnalités pour lesquelles il a été développé. La distance entre le modèle de l'évaluateur et le modèle du concepteur est alors nulle ou quasi nulle.

En revanche une évaluation ergonomique générale de ce système où les seuls critères pris en compte seraient précisément les capacités pour lesquelles le système a été créé n'est pas pertinente. La distance entre les modèles utilisés pour l'évaluation et le développement du système doit ici être maximisée.

---

<sup>2</sup>Voir 2.4.1 pour les sous-composants d'un SDHM généralement admis.

Cet exemple non réaliste explicite bien la nécessité de séparer les évaluations technologiques et ergonomiques du point de vue de la modélisation du système évalué. Les évaluations technologiques doivent être réalisées de telle sorte que la distance entre les représentations de l'évaluateur et celles du concepteur soit minimisée. En revanche les évaluations ergonomiques se doivent d'utiliser des représentations indépendantes des systèmes qu'elles évaluent, dans l'objectif de minimiser le biais associé à la proximité de représentation.

### 3.3.3 Les aspects dynamiques et co-construits du dialogue

Le besoin de comprendre et d'être compris est une nécessité pour le bon déroulement d'un dialogue. Ce besoin implique de la part des acteurs du dialogue le fait de construire **dynamiquement** et **ensemble** leurs interactions. Dans le cadre de l'évaluation de SDHM, ces aspects peuvent prendre différentes formes.

#### Observations objectives et/ou subjectives

Suite à la discussion proposée en 3.1.4, il sera considéré que les observations réalisées pour évaluer un SDHM sont dites subjectives ou objectives en fonction de leurs dépendances ou non aux jugements des utilisateurs.

Les **observations subjectives** portent généralement sur le système entier, mais il est possible de demander à l'utilisateur de spécifier son avis sur certains composants du système ou sur certains aspects du dialogue. Par exemple il est possible d'effectuer des tests perceptifs sur le composant de synthèse, ou sur les capacités de coopération du système. Ce type d'observations est généralement obtenu grâce à des questionnaires remplis par les utilisateurs.

Les **observations objectives** suivantes sont celles couramment utilisées pour l'évaluation des SDHM et de leurs composants.

- **Pour la reconnaissance vocale** : L'observation habituelle est le taux global d'erreurs de reconnaissance. Cette observation est généralement affinée en distinguant les erreurs dues à des insertions, des suppressions, ou des substitutions. Il est aussi possible de pondérer les erreurs en fonction de leur type ou de l'importance du mot en fonction de la tâche. Ce type d'observation nécessite pour chaque énoncé test d'avoir accès à une transcription correcte du signal sonore.



- **Pour l'interprétation sémantique de la langue naturelle :** Sont généralement distinguées l'interprétation sémantique hors et en contexte (voir 3.3.3 et 4.2). De même que pour la reconnaissance de la parole, le taux d'erreurs global affiné en taux d'insertions, taux de suppressions et taux de substitutions est l'observation usuelle. Ce type d'observation dépend de la représentation sémantique de l'énoncé dans le format utilisé par le système évalué. Elle nécessite généralement une annotation manuelle des énoncés dans ce format.
- **Pour le dialogue :** Les observations communément réalisées sur le dialogue sont le nombre et la durée des tours de parole, la durée totale du dialogue, et la complétude de ce dernier par rapport à la tâche objet du discours. Cette dernière observation précise si la tâche dont la réalisation est souhaitée par l'utilisateur a été résolue ou pas, et avec quelle finesse et quelle efficacité. Des observations plus fines sont parfois utilisées, pouvant prendre en compte les stratégies de dialogue, les intentions de l'utilisateur, des phénomènes particuliers (abandon, négociation, répétition).

### Taille du contexte dialogique

En fonction des évaluations, différentes tailles de contexte dialogique peuvent être prises en compte.

- **Un énoncé pris hors de son contexte dialogique :** Seul un énoncé est utilisé, sans information sur le contexte dans lequel il a été produit. Il est alors possible d'observer des scores de reconnaissance vocale et d'interprétation sémantique hors contexte. Pour la reconnaissance vocale, les observations généralement réalisées sont obtenues en comparant la transcription automatique de l'énoncé à une transcription automatique de ce même énoncé. De même pour l'interprétation sémantique, le contenu interprété de l'énoncé est comparé à une annotation manuelle de ce contenu. Pour ce dernier cas, le formalisme dans lequel est représenté le contenu sémantique de l'énoncé joue un rôle important pour l'évaluation, notamment pour des raisons de généralité (voir 3.3.4) et de représentativité de l'annotation en terme d'accord inter-annotateur.
- **Un énoncé pris dans son contexte dialogique :** L'énoncé considéré est relié à son contexte d'apparition et en particulier aux énoncés le précédant dans le dialogue. En plus de scores sur la reconnaissance vocale et l'interprétation sémantique hors contexte, des résultats sur

l'interprétation en contexte de dialogue et sur les capacités du gestionnaire d'interaction sont possibles à obtenir. Un problème récurrent à l'évaluation contextuelle d'énoncés est observable dans le cas d'évaluations ne mettant pas en jeu des utilisateurs réels. En effet se pose alors la question de la *représentation du contexte*. Comment mettre un système dans l'état contextuel dans lequel il devrait être pour observer son comportement face à un énoncé? Il est imaginable de rejouer la portion de dialogue précédent l'énoncé considéré, mais d'un système à l'autre le dialogue à rejouer ne serait pas exactement le même, et ce notamment en fonction des stratégies de chacun des systèmes. Ce problème sera traité plus en détail dans la section 4.2.2.

- **Un dialogue entier** : Sont alors disponibles tous les énoncés du dialogue. Il est donc possible de prendre chacun des énoncés hors contexte de dialogue, dans son contexte dialogique, et en plus d'appliquer des observations objectives sur le dialogue entier (voir section précédente). Prendre en considération des dialogues complets pour l'évaluation d'un SDHM nécessite d'avoir des dialogues mettant en jeu ce SDHM. Cette contrainte, bien que naturelle à première vue, conditionne en réalité la façon d'évaluer le système. En effet l'obtention de dialogues complets avec un système n'est possible, de part la nature co-construite du dialogue, qu'avec des utilisateurs réels ou simulés.

### Corpus, utilisateurs réels ou simulés

Dans cette partie sont distinguées les évaluations réalisées à partir d'interactions entre un utilisateur réel et un SDHM, entre un utilisateur simulé et un SDHM, ou grâce à des batteries de tests, autrement nommées corpus de tests.

- **Corpus de tests** : L'utilisation de batteries de tests est intéressante pour des évaluations où les énoncés sont pris hors contexte de dialogue. En effet il est aisé d'aller chercher un énoncé dans un corpus, de le fournir à un système puis d'évaluer la réponse de ce dernier. Cette méthode a de plus l'avantage d'être automatisable, mais nécessite généralement une annotation manuelle des énoncés du corpus avant d'être appliquée, ce qui est long et coûteux. Dans le cadre d'évaluations contextuelles, cette méthode se heurte aux problèmes de représentation du contexte présentés plus haut.
- **Utilisateurs Réels** : La confrontation d'un SDHM avec des utilisateurs réels est nécessaire pour aller vers des conditions normales d'uti-

lisation du SDHM évalué. Elle est cependant coûteuse, tant en moyens qu'en temps. En effet la mise en œuvre du protocole d'expérience, les passations ainsi que le dépouillement des données représentent un travail considérable. Par ailleurs ces tests doivent porter sur un échantillon suffisamment représentatif de la population ciblée pour obtenir une signification statistique. Enfin la reproductibilité des tests pour des évaluations comparatives entre plusieurs systèmes peut aisément être biaisée par des différences non contrôlées entre les utilisateurs.

- **Utilisateurs Simulés** : Cette méthode met en jeu un système prenant le rôle d'un utilisateur pour interagir avec le SDHM évalué. Cette méthode a notamment pour objectif de permettre d'évaluer un SDHM en contexte de dialogue sans passer par des utilisateurs réels. Ainsi l'évaluation est réalisée de façon automatique (coût réduit), les différences inter-personnelles sont maîtrisées (reproductibilité) et la représentation du contexte dialogique ne pose plus problème puisque le contexte est construit via une interaction. L'inconvénient de cette méthode est de nécessiter un système simulant un utilisateur, avec toute la complexité de conception et de mise en œuvre impliquée.

### 3.3.4 La généralité

Aujourd'hui le nombre de formalismes de représentation, tant au niveau sémantique que dialogique, est approximativement aussi élevé que le nombre de SDHM. Des efforts sont faits pour créer des cadres suffisamment généraux pour permettre une homogénéisation de ces formalismes. Par exemple le projet TrindiKit [Larsson and Traum, 2000], fondé sur la théorie de l'« Information State » [Traum and Larsson, 2003], propose un cadre outillé pour la création de questionnaires d'interactions.

Parallèlement le projet Technolanguage MEDIA [Deville et al., 2003] évalue l'interprétation sémantique d'énoncés à partir d'un formalisme de représentation sémantique unique sur lequel sont projetés les différentes représentations des participants. Il n'en reste pas moins vrai que la diversité des formalismes de représentation est à l'origine de difficultés à trouver des paradigmes d'évaluation généraux [Deville et al., 2002].

Trois axes de recherches sont aujourd'hui explorés :

- Des paradigmes indépendants du système qu'ils évaluent, notés **génériques par rapport au système**.
- Des paradigmes indépendants de la tâche que les systèmes évalués cherchent à résoudre, notés **génériques par rapport à la tâche**.

- Des paradigmes indépendants des deux et notés **génériques par rapport au système et à la tâche**.

### **Généricité par rapport au système évalué**

Parmi les deux types de généricité dont il est question, la généricité par rapport au système est plus facile à obtenir que celle par rapport à la tâche.

Les difficultés pour être générique par rapport au système sont de l'ordre des formats d'entrées/sorties et de l'accord sur la définition de ce qui est évalué, ici un SDHM et/ou ses composants. Il est à noter que ce dernier point est en lien avec la généricité par rapport à la tâche en ce que la définition d'un SDHM passe nécessairement par l'explicitation de la tâche qu'il vise à résoudre.

### **Généricité par rapport à la tâche**

La généricité par rapport à la tâche implique soit (1) de pouvoir évaluer sans avoir à représenter la tâche, soit (2) d'avoir un formalisme de représentation de la tâche qui permette d'y représenter plusieurs tâches.

1. Ce premier cas est assez peu exploité d'une part car la tâche à réaliser joue un rôle prépondérant dans les dialogues [Walker, 1994] et d'autre part car même avec des critères qui ne seraient pas dépendants de la tâche, l'évaluation nécessite l'observation et donc la réalisation d'une tâche<sup>3</sup>.
2. La majorité des paradigmes qui se disent génériques par rapport à la tâche sont à classer dans cette catégorie. Il n'existe pas encore de formalisme de représentation d'une tâche qui soit standard dans la communauté. En fonction des paradigmes leurs degrés d'expressivité permettent de représenter plus ou moins de tâches à différents niveaux de précision.

## **3.4 Conclusion**

Ce chapitre a exposé les principaux concepts de l'évaluation de systèmes ainsi que leur application à l'évaluation de SDHM. Ont aussi été abordés les différentes approches généralement employées lors d'évaluations de systèmes.

Le chapitre suivant présente les différents cadres et paradigmes qui permettent d'évaluer des SDHM par une approche technologique ou hybride

---

<sup>3</sup>Les SDHM dont l'objectif n'est pas de résoudre une tâche ne sont évidemment pas concernés.

ergonomique-technologique. Les paradigmes cités seront, au travers de l'explicitation précise de leurs méthodologies respectives, positionnés face aux concepts présentés dans ce chapitre.

## Chapitre 4

# État de l'art sur l'évaluation des systèmes de dialogue homme-machine

L'évaluation empirique de SDHM a naturellement débuté conjointement à la création de SDHM. Depuis une quinzaine d'années, une communauté de recherche s'est créée qui s'est intéressée tant à la définition de standards pour la caractérisation d'évaluations au travers de *cadres d'évaluation* qu'à la création de *paradigmes* plus ou moins génériques pour évaluer des systèmes de dialogue et/ou leurs composants.

### 4.1 Cadres d'évaluation

Les cadres d'évaluation présentés ci après proposent des outils, sous forme de modèles et méthodes, pour définir et caractériser des évaluations de SDHM. Leur objectif est plus la définition de normes associées à l'activité d'évaluation de SDHM que celle d'évaluations à proprement parler.

#### 4.1.1 EAGLES I et II

Les groupes de travail sur l'évaluation EAGLES-EWG I [King et al., 1996] et II [Blasband et al., 1999] (Expert Advisory Group on Language Engineering Standards - Evaluation Working Group) ont vu le jour suite à une initiative de la Commission Européenne. Les travaux ont duré de 1993 à 1996 pour EAGLES-EWG I et de 1996 à 1999 pour EAGLES-EWG II. Les laboratoires participant aux projets étaient majoritairement universitaires.

Le premier groupe de travail a développé un cadre général pour la conception d'évaluations appliquées aux systèmes et produits du TALN (systèmes de dialogue, analyseurs syntaxiques ...). Le second a cherché à consolider ce cadre ainsi qu'à le promouvoir dans les communautés scientifiques concernées.

Il résulte notamment de ces travaux la proposition de l'ensemble des buts possibles pour une évaluation vue plus haut ainsi qu'une méthodologie de conception des évaluations des outils du TALN, formalisée en appliquant les normes ISO 14598 et 9126 (respectivement sur le processus d'évaluation et sur les critères de qualité) aux outils du TALN. Cette méthodologie se décline en quatre points :

1. Définition des critères de qualités pertinents (caractéristiques générales du système qui intéressent l'évaluateur<sup>1</sup>).
2. Définition des attributs associés aux critères (attributs mesurables dont dépendent les caractéristiques générales).
3. Définition d'une mesure pour chaque attribut<sup>2</sup>.
4. Définition des méthodes pour effectuer les mesures.

Le groupe de travail EAGLES a aussi proposé de distinguer quatre types d'applications, à savoir la commande vocale, la génération de document, les services téléphoniques et les autres applications. Il propose pour chacun de ces types des critères d'évaluation génériques selon une méthodologie que L. Devillers et al. [Devillers et al., 2002] qualifie de comparative avec toute la réserve de mise lorsque doit être caractérisée une proposition qui se situe à un niveau très général.

La méthodologie proposée a été mise en œuvre pour évaluer des vérificateurs de grammaire et d'orthographe dans le cadre du projet TEMAA [des Tombe et al., 1996], quatre systèmes de dialogue dans le cadre du projet ARISE (Automatic Railway Information System for Europe) ainsi que des systèmes de traduction à mémoire (Translation Memories).

### 4.1.2 DISC

Le projet DISC [DISC Consortium, 1999, Dybkjaer et al., 1998] est une « Esprit Long - Term Research Concerted Action » qui a duré de juin 1997

---

<sup>1</sup>L'approche empruntée par EAGLES explicite que les critères généraux sur lesquels est évalué un outil sont dépendants des rôles i.e. des points de vue de chacun des acteurs par rapport à l'outil.

<sup>2</sup>Bien que la norme ISO 9126 définisse deux types de métriques (internes ou externes à l'outil), le groupe de travail n'a pas utilisé cette distinction.

à novembre 1998. Il regroupait des acteurs provenant de laboratoires de recherche universitaires ainsi qu'industriels. L'objectif du projet DISC était dans un premier temps de formaliser, par une analyse ascendante de différents systèmes, les pratiques en usage pour la création et l'évaluation de SDHM et/ou de leurs composants. Dans un second temps il a été dérivé de ces pratiques un guide des « meilleures pratiques » pour la création et l'évaluation de SDHM et/ou de leurs composants.

Six différents *aspects* ont été définis pour caractériser les analyses, à savoir :

- La reconnaissance vocale
- La synthèse vocale
- La compréhension et la génération du langage naturel
- La gestion du dialogue
- Les facteurs humains
- L'intégration des systèmes

L'explicitation des pratiques en usage a été réalisée en analysant, au regard des six aspects présentés plus haut, un ensemble de SDHM et leurs composants. Pour mener à bien cette tâche des grilles de questions et un modèle de cycle de vie ont été utilisés. Ces outils permettaient de caractériser respectivement les propriétés des SDHM et/ou de leurs composants, et leur processus de développement et d'évaluation. Les questions des grilles étaient spécifiques en fonction de si elles s'appliquaient au système complet ou à des composants de celui-ci, tandis que le modèle de cycle de vie s'appliquait indépendamment, dans la plupart des cas, au système ou à ses composants.

## 4.2 Évaluation de l'interprétation sémantique

Ces paradigmes procèdent tous à des évaluations automatiques<sup>3</sup> avec référence et sont génériques par rapport aux systèmes. Leur généricité par rapport à la tâche dépend de la façon dont elle y est modélisée, car même les plus génériques nécessitent une représentation de la tâche à résoudre par le système qu'elles évaluent, ou sa contraposée, voir 6.1.2. La distinction majeure entre les paradigmes d'évaluation de l'interprétation sémantique réside dans leur prise en compte ou non du contexte dialogique.

---

<sup>3</sup>Sans interaction entre le système et un utilisateur réel.



### 4.2.1 Évaluation hors contexte de dialogue

MADCOW [Hirschman et al., 1992, Hirschman et al., 1993]

Le Multi-site Data Collection Working group (MADCOW) a été initié par la DARPA (Defense Advanced Research Projects Agency) aux États-Unis en 1991. Ce groupe de travail a créé un corpus sur le domaine « Air Travel Information Service » (ATIS)<sup>4</sup> dans le but d'évaluer l'interprétation littérale (hors contexte) des systèmes des participants grâce à un paradigme d'évaluation spécifique. Cette campagne fut la première évaluation multi-système de l'interprétation sémantique de SDHM.

Les laboratoires participant à ce groupe de travail étaient AT&T, BBN, CMU, MIT, SRI et Paramax. Deux autres participants, NIST et une équipe du SRI non concernée par l'évaluation, ont réalisé la validation, l'annotation, la distribution ainsi que la sélection des jeux de tests et d'entraînement.

Le corpus, créé entre mai 1991 et février 1992, provenait de plusieurs sources d'une part pour permettre une évaluation commune à plusieurs systèmes sur un grand nombre de données diversifiées, et d'autre part pour réduire les coûts qu'auraient entraînés une telle évaluation par un laboratoire individuel. Il était constitué exclusivement d'énoncés d'utilisateurs annotés en fonction de deux informations :

1. la référence de l'énoncé au contexte d'interaction :
  - Classe A : indépendants du contexte.
  - Classe D : dépendants du contexte avec les liens appropriés pour le retrouver.
  - Classe X : non évaluable car le contenu de l'énoncé ne fait pas partie de la tâche.
2. Les réponses valides d'un système à cet énoncé en fonction de la base de données ATIS. Ces réponses allaient de la moins locale, ne comportant que les informations demandées, à la plus bavarde, contenant toute la surinformation intéressante possible.

Le corpus comprenait 10400 énoncés d'entraînement<sup>5</sup>, 300 énoncés annotés pour un test d'entraînement, 1000 énoncés annotés pour le test réel et 1000 énoncés supplémentaires mis de côté pour un test ultérieur.

---

<sup>4</sup>Le domaine ATIS est représenté par une base de données fournie par l'Official Airline Guide (OAG) en 1990. Cette base de données contient environ 25 tables conservant et reliant des informations sur le transport aérien aux États-Unis. Il s'y trouve par exemple les vols, les prix de ces vols, les compagnies aériennes, les villes, les aéroports, les services etc.

<sup>5</sup>Ces énoncés provenaient de 280 utilisateurs sur 5 sites.

Le paradigme, présenté en 1990 lors du *Third DARPA Speech and Language Workshop* par M. Bates, suppose qu'une « bonne » réponse d'un système à une question qui lui est posée nécessite une « bonne » compréhension de la question de la part du système. Les énoncés des utilisateurs avaient pour format la transcription écrite d'énoncés vocaux et les réponses du système étaient les résultats de sa recherche dans sa base de données. Cette approche avait donc pour avantage d'être assez générique par rapport au système évalué, les formats utilisés étant assez standard.

Les références permettant l'évaluation étaient associées à chaque énoncé fourni en entrée sous la forme de deux ensembles d'éléments de la base de données, un minimum EMin et un maximum EMax. Le système répondait correctement du point de vue de l'évaluation si l'ensemble des éléments rendus par le système satisfaisaient deux conditions : il devait être inclus dans l'ensemble EMax et devait inclure l'ensemble EMin (EMin et EMax étant propres à chaque énoncé). Le tableau 4.1 présente un exemple d'énoncé et d'ensembles de réponse minimal et maximal tirés de [Antoine and Caelen, 1999].

Il est aussi à noter que les classes A, D ou X des énoncés devaient donner lieu à deux évaluations, une sur les énoncés de la Classe A et l'autre sur ceux de la Classe D. Pratiquement seule la première a été réalisée par manque de paradigme pour l'évaluation en contexte.

<i>Quels sont les tarifs sur les vols Paris-Toulouse arrivant avant 15h ?</i>				
Ensemble Minimal		Ensemble Maximal		
Code-Tarif	Numero-Vol	Code-Tarif	Tarif	Numero-Vol
Plein.Ciel	AF_2137	Plein.Ciel	615 Fcs	AF_2137
Azur	AF_2137	Azur	875 Fcs	AF_2137

TAB. 4.1 – Exemple d'énoncé annoté pour le paradigme MADCOW

Une des limitations du paradigme MADCOW est lié à la comparaison entre la réponse de référence (sous forme de deux ensembles) et le besoin de l'utilisateur d'obtenir des informations utiles. La « méthode du comparateur » (*comparator method* dans le texte) peut considérer des réponses comme correctes alors que le système n'a pas bien interprété l'énoncé. Par exemple si tous les vols qui intéressent l'utilisateur sont quotidiens, comment savoir si le système a bien compris qu'il était question d'un vol pour un jour précis et pas d'une réponse canonique ? De plus, certaines informations qui paraissent

utiles peuvent être notées comme erronées puisqu'il n'existe pas de catégorie « partiellement correcte » pour les réponses.

Un deuxième problème se pose avec les systèmes à initiative mixte. En effet le paradigme tel qu'il a été conçu classe les questions qui sont des réponses à une question initiale du système comme non évaluables puisqu'elles font référence au contexte (Classe D). Durant l'évaluation, toutes les réponses d'un système sous forme de question seront comptées comme incorrectes puisque seules les réponses de la base de données peuvent être correctes.

### **DCR [Antoine and Caelen, 1999, Antoine et al., 2000]**

Le paradigme DCR a été présenté par Jean-Yves Antoine en 1999 dans [Antoine and Caelen, 1999]. Cette publication a fait suite aux trois articles [Zeiliger et al., 1997] [Antoine et al., 1997] et [Antoine et al., 1998] dans lesquels les auteurs présentaient le paradigme DQR. Ce dernier proposait une méthodologie d'évaluation qui nécessitait un système capable de dialoguer sur ses propres connaissances, ce qui est ambitieux. Le paradigme DCR utilise la même idée de base que DQR, à savoir que l'objet de l'évaluation n'est plus soit la réponse du système soit une représentation sémantique de l'énoncé.

L'objectif du paradigme DCR était d'améliorer certains aspects de MADCOW tout en restant dans une évaluation de même type à savoir l'évaluation de l'interprétation d'énoncés par les systèmes. Les améliorations portaient notamment sur la généralité de l'évaluation réalisée, la prédictivité de cette évaluation par rapport à des phénomènes propres à la communication homme-machine orale, et le fait d'assimiler une réponse correcte du système à une interprétation correcte d'un énoncé par le système.

Après une analyse critique du paradigme MADCOW, Jean-Yves Antoine propose deux conditions nécessaires à la mise en œuvre d'une évaluation générique et prédictive.

Pour la prédictivité, l'évaluation doit être discriminante tout en restant quantitative, donc objective. C'est-à-dire qu'elle doit être composée de sous-sessions d'évaluations portant chacune sur les phénomènes linguistiques à évaluer (ici, inattendus de la parole spontanée et résolution de co-référence). Ces phénomènes étant indépendants du contexte applicatif choisi, un diagnostic détaillé sera disponible sur ces phénomènes et sera, dans une certaine mesure, généralisable à d'autres domaines. Cette généralisation sera conditionnée par une caractérisation préalable des usages des phénomènes évalués à l'aide d'analyses de corpus. L'évaluation de phénomènes purement linguistiques ne permettant pas de juger du caractère finalisé de la communication

homme-machine orale, il faut cependant conserver en plus de l'approche discriminante une approche d'évaluation globale.

Concernant la généralité par rapport au système, l'évaluation doit être envisagée au niveau de la compréhension sans nécessiter pour autant de représentations communes. La solution proposée est de fonder l'évaluation non pas sur une référence en sortie mais sur une comparaison interne au système de référence en entrée. Ce moyen permet de ne faire intervenir que les représentations internes du système.

La méthodologie du paradigme DCR repose sur la définition de jeux de tests spécifiques pour chaque phénomène évalué. Elle se focalise explicitement sur deux types de phénomènes, à savoir les inattendus propres à la parole spontanée et la résolution de référence. Un troisième phénomène est implicitement évalué, qui est le but central de la requête. Ce dernier phénomène permet de montrer que le paradigme DCR peut se substituer à une évaluation de type MADCOW.

Un test DCR se compose de deux énoncés et d'un résultat. Les énoncés sont soit de *demande* D soit de *contrôle* C. Le *résultat* R est le booléen issu de la comparaison de D et de C, qui prend la valeur vrai si et seulement si C contient une donnée présente dans D. Le tableau 4.2 présente deux jeux de tests en exemple.

<b>D</b>	:	Quels sont les tarifs sur les vols Paris-Toulouse ?
<b>C</b>	:	Quelles sont les prestations ?
<b>R</b>	:	FAUX
<b>D</b>	:	Quels sont les tarifs sur les vols Paris-Toulouse ?
<b>C</b>	:	Quels sont les tarifs ?
<b>R</b>	:	VRAI

TAB. 4.2 – Exemple de jeux de test pour le paradigme DCR

Appliquer un test consiste alors en l'interprétation des deux énoncés D et C par le système évalué, puis à la comparaison par unification des deux représentations internes obtenues par le système pour chacun des énoncés. Si elles s'unifient sur un point au moins, le résultat renvoyé est vrai. Pour savoir si le système a « bien compris » les deux énoncés, il faut comparer le résultat renvoyé par l'unification des représentations internes à celui fourni dans le jeu de tests.

Une mise en œuvre de la méthodologie DCR a été présentée dans l'ar-

ticle [Antoine et al., 2000]. L'évaluation hors-contexte du système LAMB-DACOMP du VALORIA y est explicitée ainsi que la classification des jeux de tests et les résultats obtenus.

Lors de cette expérience un biais possible a été détecté, dû à l'utilisation d'un énoncé de contrôle. En effet si l'énoncé de contrôle est mal interprété par le système, l'évaluation qui découle de la comparaison de D et de C peut être incorrecte : si D est bien interprété alors la comparaison de C et de D peut détecter une erreur alors qu'il n'y en a pas et si D est mal interprété alors il est possible d'obtenir un résultat sans erreur qui en masque une !

### DEFI [Antoine et al., 2002b, Antoine et al., 2002a, Antoine, 2001]

le groupe de recherche Groupe de Travail « Compréhension Robuste » du Groupe De Recherche « Information - Interaction - Intelligence » (GT 5.5 GDR-I3), dont le responsable était Jean-Yves Antoine (voir ci dessus), mène des recherches sur l'évaluation diagnostique des systèmes de compréhension. En plus d'une méthodologie d'évaluation nommée DEFI, ce groupe de travail a étudié les problèmes posés à la compréhension de parole, dans l'optique de rendre les évaluations plus prédictives (voir 2.3).

La méthodologie DEFI a pour but d'évaluer de façon automatique, objective et prédictive les systèmes d'interprétation sémantique. Pour ce faire elle s'appuie sur différents jeux de tests en fonction des systèmes et de la liste de phénomènes sur lesquels il est envisagé d'obtenir des résultats. Le fait d'avoir des jeux de tests différents pour chaque système évalué implique que les évaluations ne puissent être comparatives mais les allège d'un autre point de vue puisque les systèmes n'ont besoin d'aucune adaptation (voir le paradigme DCR) ni de représentation commune (voir le paradigme PEACE).

La nouveauté de la méthodologie DEFI réside sur *la création des jeux de tests*. En effet là où les autres approches passent par une modélisation de la tâche, DEFI utilise des énoncés initiaux fournis par le réalisateur du système évalué, qui sont considérés comme étant représentatifs de la tâche. Ensuite des énoncés de tests sont dérivés des énoncés initiaux et triés en fonction des phénomènes (voir section suivante) qu'ils contiennent. La dérivation en énoncés de test doit être considérée comme une réécriture en plus compliqué des énoncés initiaux. Par exemple dans le tableau 4.3, l'énoncé D a été dérivé de l'énoncé initial I en y ajoutant une auto-correction.

I	<i>non le matin à six heures environ</i>
D	<i>non c'est le matin à sept heu non à six heures environ</i>

TAB. 4.3 – Exemple d'énoncé dérivé pour le paradigme DEFI

### 4.2.2 Évaluation en contexte de dialogue

PEACE [Bonneau-Maynard and Devillers, 2000, Devillers et al., 2002, Devillers et al., 2003]

Ce paradigme a été proposé en 2000 par Laurence Devillers *et al* et appliqué dans le cadre du projet européen TechnoLangue pour la campagne d'évaluation MEDIA. Il découle des expériences de l'équipe Traitement du Langage Parlé (TLP) du LIMSI sur l'évaluation de systèmes de dialogue. Il y est montré que le taux de compréhension en contexte est un des paramètres les plus prédictifs de la satisfaction de l'utilisateur. Ce paradigme a pour objectif de permettre une évaluation automatique, comparative et diagnostique de l'interprétation littérale et contextuelle d'un système de dialogue. Il est par ailleurs le seul qui permette une évaluation de l'interprétation sémantique en contexte de dialogue sans nécessiter de réaliser un dialogue avec le SDHM.

Du point de vue de **l'évaluation de l'interprétation sémantique**, PEACE est fondé sur l'idée que dans le cadre de tâches de renseignements liées à une base de données, il est possible de mettre en place une représentation sémantique dite « commune » vers laquelle chaque système pourra convertir sa propre représentation et réciproquement. Pour le paradigme ont donc à être spécifiés une représentation sémantique commune, des unités de références pour l'interprétation littérale et contextuelle ainsi qu'un outil permettant la comparaison de deux ensembles d'unités, l'ensemble de résultats du système et l'ensemble de référence. Il est à noter que même si ces représentations sont dépendantes de la tâche, elle présentent un caractère générique pour toute tâche de renseignements liés à une base de données.

La représentation sémantique commune choisie pour la campagne d'évaluation TechnoLangue/MEDIA devait être capable de représenter complètement selon la tâche le sens des énoncés utilisateurs. Elle devait aussi être comparable au sens d'une relation d'équivalence avec l'ensemble des requêtes possibles. Il a alors été choisie une représentation de type attribut - valeurs . Cette façon de représenter le sens permettant une annotation rapide et fiable des corpus.

À chaque énoncé était associé un mode ainsi qu'une paire attribut - valeur. Le mode représentait le fait que la phrase soit une affirmation, une négation

ou question, et les attributs étaient distingués selon quatre classes.

- les attributs de la base données, qui avaient comme valeur un élément de la base de données.
- les attributs modifieurs, qui modifiaient les attributs de la base de données
- les attributs discursifs, liés à la gestion du dialogue
- l'attribut spécial « argument » avait comme valeurs potentielles les attributs, afin de pouvoir représenter les classes abstraites associées à chacun de ces attributs.

Chaque unité d'évaluation de compréhension littérale se composait alors d'un énoncé utilisateur au format texte et de la transcription de cet énoncé (hors du contexte dialogique) dans la représentation sémantique commune. Les ambiguïtés et références non résolues inhérentes à l'absence de contexte de dialogue étaient préservées dans la représentation. Le tableau 4.4 présente un exemple d'énoncé annoté dans le formalisme précédemment exposé.

Énoncé	<i>Je voudrais un hôtel 4 étoiles dans le neuvième arrondissement</i>
Annotations	+/argument : hôtel +/arrondissement : 9 +/catégorie : 4

TAB. 4.4 – Exemple d'énoncé annoté pour le paradigme PEACE

Pour **l'évaluation en contexte**, la difficulté réside dans le fait de mettre le SDHM évalué dans l'état contextuel souhaité. Pour ce faire le moyen le plus simple est de rejouer avec le SDHM évalué des dialogues ayant déjà eu lieu. Cette méthode n'est pas satisfaisante car les dialogues à rejouer seraient *a priori* différents en fonction des systèmes à évaluer et même entre différentes versions d'un même système.

L'idée du paradigme PEACE est de fournir au système une représentation du contexte dialogique sous forme de paraphrase. Cela est possible grâce à l'utilisation, pour la création des jeux de tests, de dialogues complets. Le contexte est transcrit depuis les traces de dialogues réels dans la représentation sémantique commune. Il faut alors générer la paraphrase contextuelle à partir de la représentation sémantique obtenue. Deux possibilités ont été envisagées dans l'article [Devillers et al., 2002], à savoir l'adaptation du module de génération d'un système de dialogue pour produire automatiquement la paraphrase à partir de la représentation sémantique du contexte, ou la concaténation des phrases précédant l'énoncé cible. Dans un cas comme

dans l'autre, il est nécessaire d'envisager des corrections manuelles de ces paraphrases.

Dans le cadre du projet MEDIA, l'annotation sémantique en contexte de dialogue a été réalisée sur la phase de « ré-interprétation de l'énoncé courant en fonction du contexte », sans inclure celle, postérieure, de « ré-interprétation du contexte en fonction de l'énoncé ». En effet cette seconde phase est dépendante des stratégies de dialogue et des représentations internes des systèmes de dialogue, ce qui débordait de l'objet d'étude du projet.

Quatre propositions de représentation du contexte ont par ailleurs été soumises lors de la campagne, à savoir :

1. **L'approche écologique** où le contexte est représenté par la succession des énoncés échangés ;
2. **Les représentations abstraites** où ne sont disponibles que les annotations des énoncés se succédant ;
3. **L'approche mixte écologique et de représentations abstraites** où les deux informations précédentes sont disponibles ;
4. **L'approche par paraphrase du contexte** initialement proposée dans le paradigme PEACE.

Dans le paradigme PEACE, les paraphrases censées représenter le contexte étaient construites semi-automatiquement à partir des représentations internes à un système de l'historique du dialogue. Pour la campagne MEDIA elles doivent être indépendante des représentations internes des systèmes ainsi que des stratégies de dialogue qui sont différentes suivant les systèmes. Ainsi elle doit donc être construite à partir des interactions existantes. De ce fait, les règles à énoncer pour sa construction étaient complexes et une annotation manuelle coûteuse et subjective. En conclusion les évaluations en contexte de la campagne MEDIA ont été menées dans les configurations **écologique** et **mixte écologique et représentations abstraites** pour le contexte et la requête courante.

### 4.3 Évaluation de la gestion de l'interaction

Le terme de gestion de l'interaction est utilisé ici pour parler des comportements du système évalué du point de vue de l'interaction et pas du module de gestion de l'interaction défini en section 2.4.1. Les paradigmes présentés dans cette section sont, comme ceux présentés dans la section précédente, génériques par rapport aux systèmes. De même leur généricité par rapport à



la tâche dépend de la façon dont la tâche y est modélisée. Enfin trois types de paradigmes ont été distingués, ceux mettant en jeu des utilisateurs réels, ceux simulant des utilisateurs de façon stochastique et ceux simulant des utilisateurs de façon déterministe.

### 4.3.1 Utilisateurs réels

#### PARADISE [Walker et al., 1997c, Walker et al., 1998]

Proposé par Marilyn Walker *et al* en 1997, PARADISE (PARAdigm for Dialogue System Evaluation) propose d'évaluer les SDHM en corrélant des mesures d'évaluation objectives et subjectives. Ce paradigme fait l'hypothèse que l'objectif d'un système de dialogue est la satisfaction de l'utilisateur et que la complétude par rapport à la tâche ainsi que les divers coûts associés au dialogue peuvent être utilisés comme prédicteurs de cette satisfaction. Des données expérimentales permettent ainsi d'obtenir par régression multiple une fonction de prédiction de la satisfaction utilisateur en utilisant des mesures objectives.

Ce paradigme a notamment été appliqué pour la campagne DARPA d'évaluation multi-SDHM Communicator en 2000 et 2001 [Walker et al., 2000, Walker et al., 2001]. Il peut être défini en cinq points.

#### 1. Représentation de la Tâche

L'évaluation de la complétude d'un dialogue par rapport à la tâche nécessite d'avoir une représentation la tâche. Cette représentation doit permettre de mesurer un écart entre une tâche complètement réalisée et une autre partiellement ou incomplètement réalisée. Dans PARADISE il a été choisi d'utiliser une représentation d'une tâche par une matrice d'attribut - valeur ( « Attribute Value Matrices » , AVM). Cette matrice contient les informations qui doivent être échangées entre le SDHM et l'utilisateur durant le dialogue, représentées par un ensemble ordonné de paires [Attribut - Valeurs] possibles<sup>6</sup>. La matrice contient aussi, pour chaque paire, la direction du flot d'information pour représenter qui, de l'agent ou de l'utilisateur, reçoit l'information.

Un scénario exécuté par un utilisateur est alors représenté par l'instanciation d'une AVM indiquant les informations propres à la tâche utilisées pour ce scénario, où chaque attribut est mis de paire avec la valeur utilisée lors du dialogue. Pour chaque scénario réalisé par un utilisateur il y a deux AVMinstanciées : une dite « clé », qui est le scénario

---

<sup>6</sup>Pour les ensembles infinis de valeurs, c'est l'ensemble fini des valeurs trouvées dans les données expérimentales qui est choisi

référence auquel l'utilisateur souhaite arriver, et l'autre dite « donnée » qui est le scénario effectivement produit par l'expérimentation.

### 2. *Mesure du Succès de la Tâche*

Le succès de la tâche pour un dialogue (ou un sous-dialogue) est une mesure de « comment » le SDHM et l'utilisateur sont arrivés à résoudre les besoins d'informations propres à la tâche à la fin du dialogue. Dans PARADISE cette mesure est réalisée par le coefficient Kappa ( $\kappa$ ), défini par Cohen dans [Cohen, 1960], qui corrèle la possibilité de fournir ou d'obtenir une information juste par hasard avec les résultats expérimentaux. Le coefficient  $\kappa$  est calculé avec une matrice de confusion. Elle résume si l'agent a réussi ou non à traiter les besoins d'informations pour une tâche particulière sur un ensemble de dialogues, représentés par un ensemble de scénarios. La matrice de confusion est générée en comparant les AVM « clé » et les AVM « donnée » issues des tests expérimentaux. Elle regroupe les résultats de ces comparaisons pour l'ensemble des dialogues du corpus expérimental.

### 3. *Mesure des Coûts du Dialogue*

Les coûts du dialogue sont des mesures objectives sur les dialogues. Il en existe beaucoup dans la littérature, incluant des mesures d'efficacité (nombre de tours de parole, temps écoulé) ou des mesures sur des phénomènes précis (énoncé inapproprié ou réparé). Dans PARADISE toute mesure de coût est représentée par une fonction  $c_i$  qui peut être appliquée à tout (sous)dialogue.

L'application des fonctions de coût à des sous-dialogues définis par les sous-tâches, nécessite de spécifier, pour chaque énoncé, quels sont ses contributions en terme d'informations propre à la tâche. Pour cela, tous les énoncés sont annotés à l'aide des noms d'attributs de la tâche (définis par les AVM) en fonction des informations sur lesquelles ils portent.

### 4. *Estimation d'une Fonction de Performance*

La fonction de performance est la fonction de prédiction de la satisfaction utilisateur.

Avec les définitions précédentes de coût et de succès par rapport à la tâche, la performance pour tous les (sous)dialogues est définie comme suit :

$$Performance = (\alpha * N(\kappa)) - \sum_{i=1}^n w_i * N(c_i)$$

Ici  $\alpha$  et  $w_i$  sont respectivement les poids de  $\kappa$  et  $c_i$ , et  $N$  est une fonction de normalisation centrée réduite. La normalisation centrée réduite est utilisée afin de résoudre les problèmes d'échelle différentes entre  $\kappa$  et les  $c_i$ . Pour estimer la fonction de performance, les poids  $\alpha$  et  $w_i$  doivent être calculés en considérant leur participation à la satisfaction utilisateur. Cette dernière est typiquement calculée par des sondages où est demandé à l'utilisateur de spécifier un degré d'accord avec des phrases portant sur le comportement ou les performances du système. Pour un ensemble de dialogues pour lesquels sont collectés expérimentalement la satisfaction utilisateur (SU), le succès de la tâche ( $\kappa$ ) et les coûts du dialogue, les poids  $\alpha$  et  $w_i$  peuvent être résolus en utilisant une régression linéaire multiple. Ainsi est obtenue une fonction de performance qui prédit la satisfaction utilisateur ainsi que l'importance de chaque facteur expérimental pour le calcul de cette fonction. Cela permet d'éliminer les facteurs négligeables pour les calculs futurs avec la fonction de performance.

#### 5. Application à des Sous-Dialogues

Une hypothèse du paradigme est que la fonction de performance une fois calculée par régression, est applicable à tous les sous-dialogues du corpus. Autrement dit les facteurs qui sont prédictifs pour la performance globale, basée sur la satisfaction utilisateur, le sont aussi pour la performance locale, soit les sous-dialogues définis par les sous-tâches.

Cette hypothèse permet de prédire la satisfaction utilisateur dans le cas de sous-tâches, ce qui permet d'évaluer les stratégies de dialogue qui s'appliquent sur ces sous-tâches. Par exemple les stratégies de confirmation, de présentation de réponses.

### 4.3.2 Simulation stochastique d'utilisateurs

La simulation stochastique d'utilisateur est fondée sur des techniques d'apprentissage issues du domaine de l'Intelligence Artificielle. Le principe fondamental est d'apprendre avec des corpus annotés les comportements standards des utilisateurs, afin de pouvoir les reproduire.

Un point important à préciser est que la simulation stochastique d'utilisateur est généralement utilisée dans le cadre de l'apprentissage de stratégies de dialogue plus que dans celui de l'évaluation. Cette précision explique que ces approches n'exploitent souvent que le niveau intentionnel du dialogue et pas ceux de la langue et du son.

### Simulation dirigée par l'historique de l'interaction

Cette approche de la simulation d'utilisateur est l'une des plus ancienne [Eckert et al., 1997, Eckert et al., 1998, Levin et al., 2000], tout en étant toujours d'actualité [Georgila et al., 2005a, Georgila et al., 2006]. Elle permet à un simulateur et un SDHM d'interagir au niveau du gestionnaire de l'interaction du SDHM. C'est-à-dire que les informations échangées par les deux protagonistes sont au format des représentations sémantiques du SDHM évalué.

Le modèle utilisé pour gérer l'interaction dynamiquement avec le SDHM évalué est stochastique par n-grammes. C'est-à-dire qu'il est considéré que la probabilité  $P(\textit{phrase})$  qu'un énoncé soit produit par un utilisateur au tour de parole  $n$  est dépendante des  $n - 1$  tours de parole précédents :

$$P(\textit{phrase}) = P(\textit{phrase}_n | \textit{phrase}_{n-1}, \textit{phrase}_{n-2}, \dots, \textit{phrase}_1)$$

Dans [Eckert et al., 1997] cette vision n'est que théorique puisque la réalisation du modèle ne prend en compte que l'énoncé précédent celui à générer :  $P(\textit{phrase}) = P(\textit{phrase}_n | \textit{phrase}_{n-1})$ . Dans [Levin et al., 2000] les ensembles de probabilités sont définis en fonction de trois actes de dialogues plutôt que sur toutes les possibilités de l'interaction. Cette simplification a été réalisée car le premier modèle nécessitait trop de données d'apprentissage pour être mis en œuvre pour des SDHM à initiative mixte.

Dans [Georgila et al., 2005a], l'apprentissage est réalisé par la même méthode des n-grammes. Le dialogue est alors représenté comme une succession de listes contenant des paires  $\langle \textit{Acte de dialogue}, \textit{T\^ache} \rangle$ . Une stratégie d'approximation a par ailleurs été ajoutée afin de produire une réponse même si l'historique précis d'un dialogue n'a jamais été rencontré.

Le modèle présenté dans [Georgila et al., 2006] propose de prendre en compte, en plus des listes de paires  $\langle \textit{Acte de dialogue}, \textit{T\^ache} \rangle$ , le fait que l'utilisateur simulé ait déjà énoncé ou pas des informations, et lesquelles. Cette idée s'apparente intuitivement à la mémoire de l'utilisateur réel. Le modèle associé est donc dépendant de la tâche, du dialogue, et de l'état d'avancement de l'interaction du point de vue de l'utilisateur simulé.

### Simulation dirigée par la tâche de l'utilisateur simulé

Deux approches différentes sont présentées ici, celle de Konrad Scheffler [Scheffler and Young, 2000, Scheffler and Young, 2001] et celle de Grace Chung [Chung, 2004]. Ces deux approches sont sensiblement inspirées de

la même idée à savoir l'utilisation d'un modèle de la tâche, ou en d'autres termes du but de l'utilisateur simulé, pour gérer les interactions d'évaluation. Ces deux approches permettent des interactions soit au niveau intentionnel, soit au niveau de la langue naturelle, soit au niveau vocal à ceci près que les énoncés vocaux ou écrits produits par le simulateur de Grace Chung sont générés par des composants informatiques alors que ceux produits par le simulateur de Konrad Scheffler proviennent de corpus et sont donc de « vrais » énoncés utilisateurs.

Dans l'approche de Konrad Scheffler, la tâche, ou but, de l'utilisateur simulé est représentée par un arbre dont les nœuds sont soit (1) des choix absolus où la réaction ne dépend que des informations perçues du SDHM évalué, soit (2) des choix stochastiques où il y a une probabilité de  $P_1$  qu'on se dirige dans telle branche de l'arbre, de  $P_2$  dans telle autre etc.

L'approche de Grace Chung est de représenter la tâche ou but par un ensemble de frames sémantiques. Le dialogue est alors généré par des choix aléatoires dans ces frames sémantiques.

La différence fondamentale entre ces deux approches est que celle de Grace Chung a un modèle de la tâche figé dans lequel les choix sont fait aléatoirement alors que celle de Konrad Scheffler apprend sa représentation de la tâche ainsi que les probabilités associées aux nœuds stochastiques de l'arbre grâce à un corpus. Il est par conséquent raisonnable de supposer que les interactions générées par le simulateur Konrad Scheffler seront plus proches d'interactions humaines que celles générées par Grace Chung.

Il est par ailleurs à noter que le modèle de Konrad Scheffler inclut un composant de génération stochastique d'erreurs qui permet de simuler des erreurs d'interprétation sémantique dans l'énoncé généré par le simulateur.

### **Simulation dirigée par un modèle complexe de l'utilisateur simulé**

Le travail d'Olivier Pietquin [Pietquin, 2004, Pietquin and Dutoit, 2006] expose un cadre général de simulation stochastique d'utilisateurs à l'aide de réseaux bayésiens. Une simulation des entrées/sorties d'un SDHM est aussi proposée, permettant de brouter les interactions entre utilisateur simulé et SDHM.

Il est à noter que l'approche par simulation stochastique d'utilisateurs a été choisie par Olivier Pietquin pour permettre l'apprentissage de stratégies de dialogue à des SDHM apprenant et pas particulièrement pour l'évaluation. La notion de « satisfaction utilisateur » présentée y est en effet plus représentative d'une fonction de récompense d'un algorithme d'apprentissage que

d'une réelle estimation de la satisfaction d'utilisateurs.

Les tâches réalisables par le simulateur y sont représentées en tant que matrices d'attributs-valeurs (voir 2.4.2), ce qui revient à traiter les tâches de recherche d'informations et de remplissage de formulaire. Les paramètres pris en compte par le modèle sont les suivants :

- Le nombre d'attributs possibles pour la tâche.
- Le nombre d'actes de dialogue réalisables par le système.
- Les actes de dialogue qui sont liés à des attributs de la tâche.
- Les valeurs possibles pour chaque attribut.
- Des seuils d'acceptation de changement d'une valeur pour un attribut<sup>7</sup>.
- Des paramètres pour un calcul automatique représentant la *satisfaction utilisateur*.

Deux systèmes de dialogue ont été affinés du point de vue de leurs stratégies grâce au simulateur. L'un fournissait des informations sur des ordinateurs (marque, type de processeur, vitesse de processeur ...) et l'autre permettait de réserver des billets de train. Les expériences réalisées ont notamment permis d'adapter les stratégies des systèmes à différents niveaux d'expertise des utilisateurs simulés ainsi que d'identifier des scénarios de dialogue problématiques.

### 4.3.3 Simulation déterministe d'utilisateurs

**Simulation par corpus de scenarii [López-Cózar et al., 2002, López-Cózar et al., 2003]**

Cette approche, contrairement aux précédentes, est basée sur une gestion dynamique d'interactions d'évaluations par règles. Elle diffère aussi des autres en ce qu'elle met en jeu un SDHM entier et pas uniquement son gestionnaire d'interaction.

Cette méthode permet de tester le SDHM évalué sur un ensemble de scenarii prédéfinis. Elle nécessite donc la définition des scenarii de test ainsi qu'un corpus d'énoncés utilisateurs, annotés en correspondance avec les scenarii, pour la verbalisation des énoncés du simulateur. Ce dernier est composé d'un gestionnaire de dialogue et d'un générateur de réponse.

Le gestionnaire de dialogue s'appuie sur

- le corpus de scenarii

---

<sup>7</sup>Ces valeurs sont représentatives d'un degré d'expertise de l'utilisateur d'après l'auteur. Elles précisent un seuil, pour les compteurs d'assignation de valeurs à un attribut, au delà duquel le simulateur clot le dialogue.

- une *mémoire*, qui contient le contexte de dialogue
- l'énoncé du SDHM évalué et sa représentation sémantique

Il fournit alors au générateur de réponse une représentation de l'énoncé à générer. Le générateur de réponse va chercher dans un corpus d'énoncés celui convenant le mieux à la représentation donnée par le gestionnaire de dialogue. Ces énoncés sont au format audio et sont donc envoyés au module de reconnaissance vocale du SDHM évalué.

Les règles utilisées pour choisir la représentation de la réponse à fournir au générateur de réponse sont des règles conditionnelles (*if ... then ... else ...*) appliquées successivement ou de façon imbriquées. Les variables utilisées par ces règles sont soit les entrées du gestionnaire de dialogue (énoncé précédent et sa représentation sémantique), soit des données provenant du contexte de dialogue, soit des données provenant du corpus de scénarii (sur le scénario en cours).

Les évaluations réalisées sur un SDHM réel à l'aide de ce simulateur s'appuyaient sur trois critères principaux :

- Insertion/Deletion/substitution de mots par le module de reconnaissance vocale
- Nombre de phrases bien interprétées par le module d'interprétation sémantique
- Nombre de dialogues dont les buts du scénarii ont effectivement été réalisés en moins de 30 échanges

## 4.4 Conclusion

L'évaluation de système de dialogue a longtemps été considérée comme partie intégrante du développement des SDHM et pas comme un domaine de recherche à part entière. Une communauté de recherche s'est créée ces dernières années grâce aux besoins croissants en évaluation des SDHM. Des cadres et paradigmes ont été développés, proposant différentes approches et méthodologies pour évaluer les systèmes de dialogue. Ce chapitre a proposé une description de ces principaux cadres et paradigmes.

La partie suivante présente en trois chapitres le paradigme SIMDIAL, permettant une évaluation automatique de SDHM. Par rapport aux différents paradigmes cités ci-dessus, SIMDIAL permet d'évaluer quantitativement la gestion de l'interaction de SDHM par la méthode de simulation déterministe d'utilisateurs (voir 4.3.3). Le premier chapitre a pour objectif de motiver les choix méthodologiques et technologiques ayant mené à la définition du paradigme. Ces choix seront notamment explicités relativement aux concepts introduits dans la première partie du document. Le deuxième chapitre expose

de façon détaillée la modélisation d'un simulateur d'utilisateurs et le troisième présente des résultats d'évaluation obtenus en appliquant le paradigme à deux SDHM réels.





## Deuxième partie

### SIMDIAL

Un paradigme d'évaluation  
automatique de systèmes de  
dialogue homme-machine par  
simulation déterministe  
d'utilisateurs



# Chapitre 5

## Les choix méthodologiques et technologiques

Ce chapitre motive les choix effectués pour la réalisation du paradigme SIMDIAL. Dans une première partie seront expliqués les choix méthodologiques qui ont mené à réaliser un simulateur déterministe d'utilisateurs pour évaluer des systèmes de dialogue. Dans la seconde le point de vue technologique sera utilisé pour préciser comment a été mis en œuvre le simulateur.

### 5.1 Choix méthodologiques

#### 5.1.1 Approche générale

En premier lieu il faut rappeler que la technologie ARTIMIS est centrée sur la notion de **dialogue naturel** et pas sur l'interprétation sémantique. Le paradigme SIMDIAL doit donc cibler les comportements dialogiques et pas nécessairement l'interprétation sémantique.

Par ailleurs comme explicité en introduction de ce document, cette thèse a pour domaine général l'informatique. Par conséquent une approche plutôt technologique qu'ergonomique pour évaluer les SDHM est envisagée, sans toutefois négliger les critères et analyses psycho-ergonomiques exploitables dans le cadre d'évaluations technologiques.

En outre le paradigme SIMDIAL doit être générique face aux systèmes évalués (voir ci après) ce qui exclut une approche qualitative de l'évaluation. En effet de telles différences résident aujourd'hui tant dans les approches que dans les moyens technologiques mis en œuvre pour réaliser des SDHM qu'il est nécessaire de s'appuyer sur des données quantitatives pour juger de la

qualité réelle des systèmes.

Enfin pour le même besoin de généralité face aux systèmes évalués, une approche où le système est perçu en boîte noire a été choisie. L'évaluation étant focalisée sur la gestion de l'interaction, une approche en boîte transparente aurait nécessité pour observer le composant correspondant

- soit que le formalisme de représentation sémantique de tous les systèmes évaluables soit identique, ce qui aujourd'hui est loin d'être le cas ;
- soit qu'un formalisme de représentation sémantique commun sur lequel auraient été projetés les représentations sémantiques de chacun des systèmes évalués soit défini (voir le paradigme d'évaluation PEACE 4.2.2).

Cette dernière solution est envisageable mais présente le biais d'évaluer à la fois l'interprétation sémantique et la capacité de transformation des énoncés vers le formalisme commun.

### 5.1.2 Généralité

#### Face aux systèmes évalués

Comme précisé dans le paragraphe 1.2, le paradigme SIMDIAL doit être applicable à des SDHM fondés sur la technologie de dialogue ARTIMIS mais aussi permettre de comparer ces SDHM avec d'autres fondés sur d'autres technologies. Autrement dit il doit être général face aux systèmes évalués.

La comparaison implique qu'une *même évaluation* au sens méthodologique soit appliquée aux différents systèmes ou aux différentes versions d'un même système. Il est de plus à noter que lorsqu'il est question d'évaluations avec une approche technologique, il est souvent implicitement admis qu'il soit possible de comparer différentes versions d'un même système, tant pour observer des progressions que pour tester de la non-régression.

#### Face aux tâches à résoudre

La technologie ARTIMIS permettant de développer des SDHM sur des tâches à résoudre différentes, le paradigme SIMDIAL doit lui aussi viser la possibilité d'être appliquée sur différentes tâches, d'où un besoin d'une grande généralité face à la tâche. Plus le paradigme sera général face à la tâche que les SDHM évalués cherchent à résoudre, moins sera coûteux le changement de tâche pour évaluer d'autres systèmes.

De plus la généralité du paradigme SIMDIAL par rapport à la tâche peut permettre d'exploiter indépendamment de la tâche certains critères d'évaluation. En effet il devient possible d'observer les mêmes critères pour

des SDHM réalisant des tâches différentes, ce qui rend possible la comparaison entre ces systèmes, réalisant pourtant des tâches différentes, sur ces critères. De telles comparaisons ne sont aujourd'hui réalisées que lors d'évaluations qualitatives par conséquent d'un manque d'informations quantitatives homogènes.

### 5.1.3 Évaluation avec mesure ou référence

Au regard de l'existant les deux cas ont déjà été expérimentés. En particulier la méthodologie PARADISE évalue avec des mesures et les méthodologies par simulation d'utilisateurs évaluent avec référence. Il est alors intéressant de constater que les mesures mises en œuvre pour évaluer avec la méthodologie PARADISE sont parfois très proches de la définition d'un comportement de référence. Dans le cas d'une mesure sur la résolution de la tâche par exemple, la définition de la mesure revient à expliciter la(les) bonne(s) façon(s) de résoudre la tâche.

Partant de l'idée que la définition d'une mesure suffisamment fine sur les comportements passe par l'explicitation d'abstractions revenant à définir les « bons » comportements, le choix a été fait d'utiliser l'explicitation de ces comportements pour évaluer, i.e. d'évaluer avec référence. L'explicitation des bons comportements est présentée dans la suite du document comme le modèle de l'interaction par rapport auquel sont réalisées les évaluations.

Un autre choix est alors à faire entre évaluer à partir d'interactions avec des utilisateurs réels ou à partir d'interactions dans lesquelles l'utilisateur est simulé.

### 5.1.4 Utilisateurs réels ou simulés

Contrairement à une évaluation avec mesure, l'évaluation avec référence des comportements d'un SDHM permet de se placer dans le cadre de la simulation d'utilisateurs. En effet les comportements du système évalué sont :

- Générés par des situations connues et maîtrisées.
- Prédéfinis pour permettre le « bon » déroulement d'une interaction du point de vue du simulateur.

De plus l'évaluation avec référence dans un contexte réel d'utilisation nécessiterait de définir un modèle de l'interaction qui rendrait compte de toutes les caractéristiques comportementales face auxquelles l'évaluation devrait être réalisée. Or du fait de la variabilité des comportements des utilisateurs un tel modèle serait rapidement soit trop compliqué, soit trop peu représentatif des interactions réelles, voire même les deux. Enfin les évaluations avec des utilisateurs réels sont plus coûteuses que des évaluations où

l'utilisateur est simulé, tant en temps passé qu'en moyens à mettre en œuvre. En effet le coût de la réalisation d'un simulateur est initialement important mais permet de faire diminuer les coûts récurrents d'évaluations avec des utilisateurs.

La méthodologie d'évaluation avec référence s'applique assez simplement à la simulation d'utilisateurs alors que celle avec mesure est mieux adaptée à des évaluations en contexte réel d'utilisation, du fait de la difficulté de représentation de l'ensemble des situations d'évaluation.

Le choix a donc été fait de se placer dans le cadre de la simulation d'utilisateurs, le simulateur implémentant le modèle de l'interaction défini.

Reste alors à choisir entre simulation d'utilisateurs de façon stochastique ou de façon déterministe.

### 5.1.5 Simulation stochastique ou déterministe

Là encore les deux possibilités ont déjà été explorées (voir 4.3.3 et 4.3.2). Une évaluation quantitative des méthodologies stochastiques a été proposée dans [Schatzmann et al., 2005]. L'argumentation qui suit s'inspire de cet article.

L'argument majeur de la méthode stochastique est qu'elle permet de simuler des comportements généralement plus proches de ceux d'un utilisateur réel que dans le cas d'une simulation déterministe. En revanche, la simulation stochastique s'appuyant sur la notion de répétitivité<sup>1</sup>, elle ne permet pas aisément de simuler des comportements peu observés dans les corpus d'apprentissage. De façon plus générale l'apprentissage de comportements étant fondé sur des corpus d'apprentissage, la représentativité de ces derniers face aux comportements de simulation souhaités joue un rôle très important. Enfin, bien que les récents progrès fassent augmenter la qualité des simulations, des métriques statistiques simples permettent toujours de différencier des dialogues réels de dialogues générés par simulation comme par exemple l'uniformité des distributions du nombre de tours de parole pour résoudre une tâche.

À l'opposé, la méthode déterministe génère des comportements maîtrisés et potentiellement aussi précis que souhaités, mais il est nécessaire de les spécifier. La difficulté réside alors dans la modélisation des comportements souhaités. Enfin il est intéressant de noter que cette méthodologie est bien adaptée pour l'observation de corrélations entre les stratégies de dialogue, les comportements de l'utilisateur simulé et les performances du système évalué.

---

<sup>1</sup>L'apprentissage automatique de comportements est fondé sur les comportements récurrents.

Le choix de simuler les utilisateurs de façon déterministe est motivé notamment par le besoin d'expérimentation de certains types de comportements qu'il n'aurait pas été aisé (1) d'observer dans des corpus et (2) d'explicitier via un modèle dans ces mêmes corpus.

Ce choix précise alors que le modèle de l'interaction par rapport auquel sont évalués les systèmes et qui guide les comportements de l'utilisateur simulé est défini par expertise.

### 5.1.6 Critères d'évaluation

Le dernier point nécessaire à la définition du paradigme SIMDIAL est l'ensemble des critères exploités pour juger les systèmes évalués. Ces critères seront détaillés plus précisément dans la section 6.3.4.

Relativement au fait d'évaluer des comportements dialogiques par simulation déterministe d'utilisateurs, le critère principal de jugement est **l'aboutissement des dialogues par rapport à la tâche**. Ce critère précise si la tâche que le simulateur cherche à résoudre a été menée à bien ou pas. Ce critère est essentiel dans le cadre d'une évaluation d'un système de dialogue finalisé car la réalisation de la tâche est censé être l'objectif premier du système évalué<sup>2</sup>[Walker, 1994].

Il est à noter que la réalisation de la tâche ne passe pas nécessairement par la fourniture d'une information. En effet il est possible que pour les critères du simulateur le système évalué n'ait pas de réponse et que le simulateur n'élargisse pas sa recherche.

Un autre critère d'évaluation classiquement utilisé est le dénombrement des tours de parole par dialogue pour les dialogues ayant aboutis. Ce critère dit « d'efficacité » permet de juger de la rapidité de résolution de la tâche en fonction des réglages du simulateur.

Enfin la présence de relaxation de critère et/ou d'ambiguïté est comptabilisée pour chacun des dialogues (voir 2.3.2 pour une définition précise de la relaxation et de l'ambiguïté). Ces phénomènes pouvant jouer sur le nombre de tours de parole pour la résolution de la tâche, il est pertinent de les contrôler.

---

<sup>2</sup>D'autres objectifs premiers pourraient être imaginés comme par exemple la maximisation du temps de communication dans le cadre de systèmes commerciaux.



### 5.1.7 Conclusion

Ainsi, compte tenu des constats réalisés lors de l'analyse de la problématique, des choix méthodologiques sont faits. Le paradigme SIMDIAL doit permettre une **évaluation avec référence, technologique, quantitative et générique des comportements dialogiques** de SDHM perçus en **boîte noire**. Pour ce faire il exploite la **simulation déterministe d'utilisateurs** et les critères d'évaluation d'**aboutissement des dialogues** et d'**efficacité**.

## 5.2 Choix technologiques

### 5.2.1 Architecture et interface

#### Architecture générale du simulateur

La réalisation d'un simulateur d'utilisateurs de systèmes de dialogue revient à réaliser un système de dialogue finalisé puisqu'il s'agit d'un système dialoguant dans un objectif précis. Les différences entre le simulateur d'utilisateurs et le système auquel ce premier sera confronté résident d'une part dans la tâche à résoudre par chacun de ces systèmes (voir 6.1.2) et d'autre part dans la variabilité des verbalisations des énoncés (voir 2.3.1).

Le choix de l'architecture générale du simulateur (ou du système de dialogue simulant un utilisateur) doit par ailleurs tenir compte du problème de la proximité conceptuelle des modélisations (voir 3.2.6). En effet cette architecture impacte la modélisation du système évalué faite au travers de la réalisation du simulateur.

Ces considérations présentées, le choix a été fait d'utiliser, pour le simulateur, l'architecture standard des systèmes de dialogue présentée en 2.4.1.

D'une part cette architecture est le standard de la communauté de recherche sur les SDHM.

D'autre part elle permet d'obtenir un bon niveau de généralité face aux tâches à résoudre (voir section suivante 5.2.2).

Enfin relativement au modèle du système évalué, cette architecture est intéressante puisqu'elle sépare les modules de gestion de l'interaction, d'interprétation sémantique et de verbalisation en langue naturelle. Ainsi la focalisation de l'évaluation sur la gestion de l'interaction est simplifiée puisque cette dernière est séparée de l'interprétation et de la verbalisation.

### Interface du simulateur avec le système évalué

Pour permettre au simulateur de « dialoguer » avec les systèmes évalués, un canal de communication est nécessaire. Partant du postulat généralement vérifié que l'architecture des systèmes évalués est celle présentée en 2.4.1, les canaux de communication peuvent aisément être conçus à trois niveaux différents.

- **Au niveau des représentations sémantiques** entre les composants de gestion de l'interaction. Cette possibilité n'a pas été retenue pour les raisons présentées dans le dernier paragraphe de la section 5.1.1.
- **Au niveau du signal sonore** entre les composants de reconnaissance vocale et de synthèse vocale. Cette possibilité n'a pas non plus été retenue d'une part car elle nécessiterait un très grand corpus sonore d'énoncés utilisateurs, et d'autre part car si un tel corpus existe, il serait plus aisé d'adapter uniquement le modèle de reconnaissance vocale à ce corpus plutôt que de le faire au travers d'évaluations ciblant la gestion de l'interaction.
- **Au niveau du texte transcrit** entre les composants d'interprétation sémantique et de génération en langue naturelle. Ce niveau d'interaction a été choisi (voir le schéma 5.1) car d'une part il est le plus aisé à mettre en œuvre, d'autre part il permet de simuler des erreurs de reconnaissance vocale ainsi que des verbalisations pseudo-humaines (voir 6.3.3), et enfin il est très générique face aux systèmes évalués.

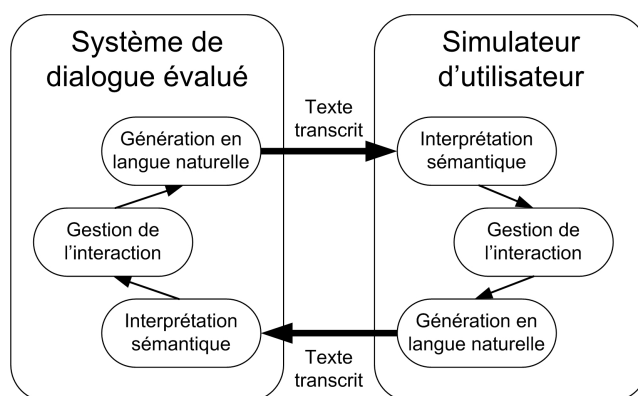


FIG. 5.1 – Interface du simulateur avec un système

## 5.2.2 La gestion de l'interaction

### Les modèles utilisés pour la gestion de l'interaction

Les modèles utilisés pour gérer l'interaction du simulateur doivent d'une part être suffisamment simples pour bien maîtriser les comportements du simulateur et d'autre part être assez génériques pour permettre au simulateur de n'être pas spécifiquement proche de SDHM existants.

Il a été choisi d'utiliser des modèles pour représenter le domaine, la tâche, le dialogue ainsi qu'un autre ensemble de paramètres ici nommé *modèle de l'utilisateur simulé*. Par ailleurs un système de bruitage des énoncés du simulateur est mis en œuvre sous le terme de *phénomènes perturbateurs*.

Les modèles du domaine, de la tâche et du dialogue utilisent des formalismes usuels pour représenter leurs données respectives.

Le modèle du domaine (voir 6.1.1) exploite les concepts généralement admis pour une tâche considérée ainsi qu'un ensemble de valeurs possibles définissant la couverture du modèle. Cette représentation non structurée a été choisie pour sa simplicité ainsi que pour être différente des deux méthodes actuellement utilisées pour représenter les domaines et les tâches, à savoir les réseaux de connaissances et les matrices d'attributs-valeurs.

Le modèle de la tâche est lui aussi différents des deux modèles cités ci dessus et utilise une représentation hiérarchique à base de schémas (voir 6.1.1) s'appuyant sur le modèle du domaine. Là encore cette façon de représenter est simple afin de mieux maîtriser les comportements du simulateur relativement à ce modèle. Un dernier point à noter concerne la séparation des modèles du domaine et de la tâche, permettant ainsi une plus grande généralité. En effet il devient possible d'adapter le modèle du domaine sans toucher à celui de la tâche et vice versa.

Le modèle du dialogue (voir 6.1.3) s'inspire de la philosophie du langage et notamment des actes de langage d'Austin [Austin, 1962] et Searle [Searle, 1969]. Cette approche est aujourd'hui le standard de la communauté de recherche sur le dialogue homme-machine.

Le dernier modèle dont il est question ici est celui précédemment défini comme « modèle de l'utilisateur simulé ». Il détermine des caractéristiques comportementales du simulateur telles que le nombre d'informations fournies par tour de parole ou encore sa stratégie de réponse (voir 6.1.4).

Un point intéressant à noter est que les paramètres de ce modèle définissent bien un utilisateur et pas un usager au sens de Stéphane Chaudiron (voir 3.2.2). En effet l'utilisateur simulé est défini au travers de ce modèle non pas comme un acteur social mais bien comme un individu cherchant à

resoudre une tâche précise.

Par ailleurs ce modèle est traité indépendamment des modèles précédemment cités car il est spécifique à l'approche choisie et n'est donc pas standard dans la communauté. Il définit un ensemble de paramètres pouvant prendre différentes valeurs en fonction des comportements souhaités pour le simulateur.

Enfin l'ajout de phénomènes perturbateurs à une interaction est équivalent à un bruitage du comportement du simulateur dans le but d'obtenir des informations prédictives sur les phénomènes considérés.

Il aurait été possible d'ajouter la notion de phénomène perturbateur au modèle de l'utilisateur simulé. En effet les deux modèles ont un impact sur le comportement du simulateur. L'intérêt de les distinguer réside dans le fait de n'avoir, dans le modèle de l'utilisateur simulé, que des paramètres dont la définition est nécessaire pour générer une interaction. Les phénomènes perturbateurs ne sont pas nécessaires à la génération d'une interaction, ils étendent le pouvoir diagnostic<sup>3</sup> du paradigme SIMDIAL. De plus cette distinction permet de ne pas modifier les différents modèles lors de l'ajout d'un phénomène perturbateur, ce qui rend l'approche plus générique face à ces derniers.

### La représentation du contenu des énoncés

La simulation d'une compréhension au niveau du dialogue implique de disposer des informations sur lesquels la simulation de compréhension est réalisée. Ces informations sont ici d'ordre dialogiques (modèle du dialogue) et sémantiques (modèle du domaine) et sont nommées *contenu des énoncés*<sup>4</sup>. Ce sont des informations de ce type qui sont fournies en entrée, ainsi qu'obtenues en sortie, du gestionnaire d'interaction.

Le formalisme utilisé pour représenter ce contenu est fondé sur celui défini pour la campagne d'évaluation du projet européen TechnoLangue/MEDIA [Maynard et al., 2004] [Devillers et al., 2004]. Il est structuré hiérarchiquement sur deux niveaux.

Le premier niveau représente les informations dialogiques des énoncés. Il utilise pour ce faire le modèle de dialogue du simulateur.

Le second niveau définit le contenu sémantique des énoncés relativement au modèle du domaine. Chaque information y est représentée par un triplet contenant un mode, un attribut et une valeur. Le mode définit si l'information

---

<sup>3</sup>Au sens prédictif.

<sup>4</sup>Par rapport à un énoncé réel dont le contenu est implicite

est affirmée ou niée et l'attribut fournit le type de l'information relativement à la tâche. La valeur est, comme son nom l'indique, le contenu informationnel véhiculé par la portion d'énoncé. Par exemple le nom d'une ville ou un horaire<sup>5</sup>.

Ci dessous un exemple d'un énoncé dont le contenu est annoté dans le formalisme de représentation.

Auriez-vous un restaurant antillais pour moins de vingt euros ?  
Request (< + / spécialité / antillais >, < + / prix inférieur / 20 >)

### 5.2.3 Les autres modules

#### L'interprétation sémantique

L'interface entre le simulateur et le système évalué se situe au niveau des énoncés transcrits au format texte. Par conséquent il est nécessaire d'extraire le contenu des énoncés fournis par le système évalué dans le formalisme de représentation du contenu précédemment exposé. Cette phase d'extraction de contenu est nommée interprétation sémantique par analogie aux SDHM.

Il est à noter que les énoncés dont le contenu est à extraire sont générés par un système informatique. Il semble donc raisonnable que leur variabilité soit suffisamment faible pour pouvoir en extraire aisément des portions d'énoncés types. L'interprétation sémantique du contenu des énoncés est alors réalisé par îlots sémantiques fondés sur l'égalité textuelle des contenus relatifs au domaine, comme par exemple les noms de villes. Cette méthode pourrait être affinée, notamment en ajoutant la possibilité de prise en compte de variantes flexionnelles.

Le choix d'utiliser un module d'interprétation plutôt que d'exploiter directement les représentations sémantiques du système évalué se justifie, comme précisé dans le dernier paragraphe de la section 5.1.1, par des besoins de généralité. La campagne d'évaluation MEDIA a démontré la difficulté d'unification de plusieurs formalismes de représentation sémantique vers un formalisme unique. L'analyse d'énoncés type pour un système est alors peu coûteux comparé au développement d'un traducteur de représentations sémantiques.

La réalisation du module d'interprétation sera présentée dans la partie 6.3.2.

---

<sup>5</sup>C'est le modèle de la tâche qui définit la précision de la représentation et donc des valeurs : un horaire peut être vu comme une heure et des minutes agrégées ou non ...

### La génération en langue naturelle

De façon duale à l'interprétation sémantique, il est nécessaire pour le simulateur de produire des énoncés en langue naturelle. De plus ce dernier prenant le rôle d'un utilisateur, les énoncés qu'il génère doivent d'une part être assez similaires à ceux produits par des utilisateurs réels et d'autre part assez variables pour permettre de tester un large éventail de situations.

La méthode usuelle utilise un corpus d'énoncés réels annotés. Les énoncés y sont cherchés à partir de leur contenu en fonction du contenu que le simulateur doit verbaliser. Cette méthode a pour désavantage de demander de très gros corpus pour permettre la verbalisation des principaux contenus souhaités.

La génération des énoncés du simulateur a donc nécessité le développement d'une stratégie de verbalisation propre au simulateur qui est présentée en section 6.3.3. Cette stratégie présente l'avantage de ne nécessiter qu'un petit corpus et un thésaurus sur la tâche pour verbaliser de façon proche de celle d'utilisateurs réels de nombreux contenus. Un test perceptif a par ailleurs été réalisé pour contrôler la validité des énoncés produits par ce module (voir le paragraphe sur le test perceptif en section 6.3.3).

### Le diagnostic automatique

Des fonctions de diagnostic automatique ont aussi été développées dans le but d'automatiser le plus possible les capacités d'évaluation du simulateur. Ainsi les différents critères présentés en 5.1.6 sont diagnostiqués automatiquement pour les dialogues générés par le simulateur. Par ailleurs pour maintenir une généricité maximale face aux systèmes évalués, les critères diagnostiqués n'utilisent, en plus de l'état interne du simulateur, que les énoncés générés par les systèmes évalués et pas d'informations internes à ces derniers.

Comme précisé précédemment en 5.1.6, les critères diagnostiqués portent sur l'aboutissement des dialogues relativement à la tâche prévue par le simulateur, sur le dénombrement des tours de parole et sur la présence de relaxations ou de désambiguïisations dans les dialogues.

## 5.3 Évaluer l'évaluation

Cette section présente une évaluation qualitative (voir 3.2.3) du paradigme SIMDIAL.

Un premier point à noter est que le paradigme n'a pas pour objectif d'être une référence du « bon » comportement d'un SDHM dans l'absolu, mais plutôt

de rendre possible des tests automatiques pour des situations et problèmes connus. Les différents phénomènes modélisés sont ceux proposés en 2.3, bien que certains comme les énoncés complexes ou les problèmes de couverture n'aient pas été traités.

Il n'en reste pas moins vrai que le paradoxe de l'évaluation reste de rigueur et qu'il est normal de réussir à faire référence pour la communication entre les SDHM et les représentations de l'utilisateur telles qu'elles sont représentées aujourd'hui. En revanche les comportements générés par le simulateur n'ont pas pour objectif d'être représentatifs d'une **réalité** de comportements humains, ni en terme de variabilité des phénomènes observables, ni même en terme de cohérence statistique face à ces phénomènes. Le simulateur a pour but de générer des comportements *proches* de ceux d'utilisateurs réels et dont la variabilité est maîtrisée face à des phénomènes prédéfinis.

Ces comportements doivent donc permettre l'observation et donc le jugement face à des phénomènes précis et dont l'apparition est maîtrisée. Cette approche octroie notamment au paradigme un fort pouvoir diagnostique face aux phénomènes traités.

Enfin il est à noter que le paradigme SIMDIAL n'a pas pour prétention de remplacer les évaluations ergonomiques. En effet il est pour l'heure trop difficile de représenter la variabilité tant des utilisateurs que des phénomènes observables. De plus les nombreuses possibilités d'approches ergonomiques différentes rendent cette activité plus que jamais nécessaire.

## 5.4 Conclusion

Le paradigme SIMDIAL, présenté dans ce chapitre, propose d'évaluer les comportements dialogiques de SDHM perçus en boîte noire. Les évaluations réalisées sont technologiques, quantitatives, avec référence et génériques face aux systèmes évalués et aux tâches à résoudre. Pour ce faire une véritable dynamique de dialogue est instaurée entre le système évalué et un simulateur déterministe d'utilisateurs.

Par ailleurs les différents choix technologiques présentés dans la seconde partie de ce chapitre seront précisés dans le chapitre suivant. La mise en œuvre du simulateur sur deux tâches de recherche d'information différentes y est présentée. Ces tâches portent sur les domaines des restaurants parisiens et des informations boursières.

# Chapitre 6

## Le simulateur d'utilisateurs

Ce chapitre présente la modélisation détaillée du simulateur d'utilisateurs pour évaluer deux systèmes sur des tâches de recherche d'information. L'application nommée `PLANRESTO` permet de chercher des restaurants à Paris et `DIALOGUEBOURSE` permet d'obtenir des informations sur les cours de la bourse ainsi que sur des portefeuilles d'actions.

### 6.1 Modèle de l'interaction

L'algorithme du gestionnaire d'interaction produit le contenu des énoncés à verbaliser. La génération de ce contenu dépend 1) du contenu des énoncés produits par le système évalué et 2) de l'état interne du gestionnaire d'interaction. Comme expliqué dans le chapitre précédent en 5.2.2, les entrées et sorties de l'algorithme utilisent le formalisme de représentation du contenu des énoncés et plusieurs modèles statiques sont exploités pour générer le contenu de la réponse. Ces modèles précisés ci-dessous sont du domaine, de la tâche, du dialogue et de l'utilisateur simulé. Seront aussi présentés les phénomènes perturbateurs utilisés pour bruiteur les interactions.

#### 6.1.1 Modèle du domaine

Le premier point à noter est que les modèles présentés dans cette section sont dissociés pour des raisons de généralité entre les informations propres à la tâche de celles jugées génériques à plusieurs tâches. Le tableau 6.1 présenté plus bas concerne les informations génériques et les deux suivants 6.2 et 6.3 celles des application `PLANRESTO` et `DIALOGUEBOURSE`.

Par ailleurs les modèles des domaines exposés ci après sont des représentations non structurées des informations utilisables pour la résolution par le



dialogue de la tâche considérée. Ces informations sont présentées au format concepts et valeurs associées, ce qui correspond à classer les valeurs possibles de la tâche par leur concept générique. Par exemple à la valeur *Tour Eiffel* est associé le concept *lieu*.

Concept	Explication	Valeurs associées
meta	Gestion de la tâche	quitter, nouvelle recherche, aide, répéter, pas compris
listeNav	Navigation verbale dans une liste	suivant, précédent, premier, dernier, consulter
listeNavNum	Navigation numérique dans une liste	Entier positif
listeProp	Nombre de propositions	Entier positif
bruit	Phénomènes perturbateurs	reconnaissance vocale, hésitation
null	Quand rien à associer	

TAB. 6.1 – Concepts et Valeurs génériques

Concept	Explication	Valeurs associées
restaurant	Nom d'un restaurant	Nom du restaurant <sup>1</sup>
specialite	Nom d'une spécialité	Nom de la spécialité <sup>1</sup>
lieu	Nom d'un lieu	Nom du lieu <sup>1</sup>
prixInf	Prix inférieur à la valeur associée	Entier positif
prixEgal	Prix égal à la valeur associée	Entier positif
prixSup	Prix supérieur à la valeur associée	Entier positif
info	Information relative au sujet précisé	général (présentation système), prix, lieu, spécialité, critère (quand non précisé), ambiance, adresse, plus d'infos, téléphone, restaurant

TAB. 6.2 – Concepts et Valeurs spécifiques à la tâche PLANRESTO

Concept	Explication	Valeurs associées
compte	Type du compte	PEA, Compte titre
indice	Nom d'un indice boursier	Nom de l'indice <sup>1</sup>
action	Nom d'une action	Nom de l'action <sup>1</sup>
infoCompte	Information relative à un compte	composition, valeur, solde espèce, solde SRD, plafond SRD
infoIndice	Information relative à un indice	niveau, niveau clôture, performance
infoAction	Information relative à une action	cote, volume échangé, dividende, PER, BNPA, opinion analystes, rendement, performance

TAB. 6.3 – Concepts et Valeurs spécifiques à la tâche DIALOGUEBOURSE

### 6.1.2 Modèle de la tâche

Dans le cadre du dialogue finalisé, le « bon » déroulement d'une interaction est notamment dépendant de la résolution de la tâche objet du dialogue [Walker, 1994], et par conséquent de la définition de celle-ci .

Le modèle de la tâche défini pour le simulateur représente la tâche contraposée de celle que cherche à résoudre le système évalué. La représentation de la tâche du simulateur joue donc un rôle prépondérant dans le « bon<sup>2</sup> » déroulement des interactions. Les évaluations réalisées sont éminemment relatives à cette représentation.

Comme présenté dans le chapitre précédent, le formalisme choisi pour représenter la tâche est un modèle hiérarchique à base de schémas. Il dépend du modèle du domaine puisqu'il en exploite les concepts. Les concepts définis comme génériques dans le modèle du domaine sont considérés, pour le modèle de la tâche, comme relevant de capacités de niveau méta. Ils ne sont donc pas présentés dans les modèles associés aux applications PLANRESTO en 6.1 et DIALOGUEBOURSE en 6.2.

Le modèle de la tâche développé pour l'application DIALOGUEBOURSE a pour objectif racine un but dit « abstrait » qui représente la conjonction de ses sous-buts. Il est considéré comme résolu quand tous ses sous buts le

<sup>1</sup>Liste des valeurs possible pour chacun des attributs concernés

<sup>2</sup>Au sens de la référence face à laquelle le système est évalué.

sont, qu'ils le soient effectivement par l'obtention d'une réponse ou qu'aucune réponse ne puisse être fournie par le système.

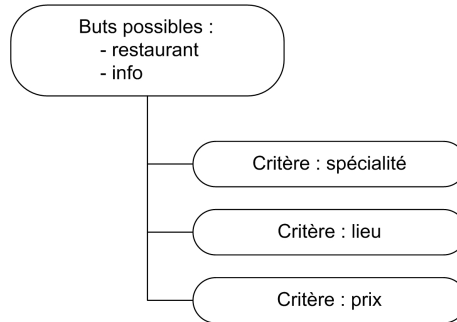


FIG. 6.1 – Modèle de la tâche pour PLANRESTO

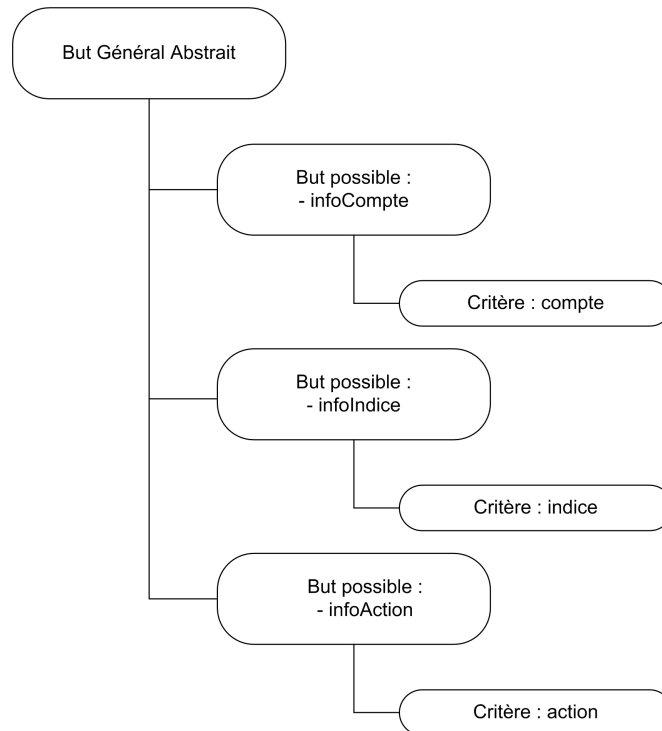


FIG. 6.2 – Modèle de la tâche pour DIALOGUEBOURSE

### 6.1.3 Modèle du dialogue

Le modèle du dialogue est commun aux différentes tâches possibles du simulateur. Comme introduit dans le chapitre précédent, il s'inspire de la philosophie du langage et notamment des actes de langage d'Austin [Austin, 1962] et Searle [Searle, 1969]. Les actes utilisés sont les six actes principaux extraits du formalisme FIPA ACL [FIPA Consortium, 2002] par le consortium MEDIA (voir 4.2.2). Cette sous liste a été extraite pour des raisons d'accord inter-annotateurs, la liste complète ne permettant pas un accord suffisamment élevé entre les annotations. Le tableau qui suit présente ces six actes ainsi que leur signification. Un sujet concernant le niveau de la

Inform(X)	Acte assertant l'information X
Request(X)	Acte de requête de l'information X
Accept	Acceptation de l'assertion de l'énoncé précédent
Reject	Refus de l'assertion de l'énoncé précédent
Open	Acte d'ouverture de dialogue
Close	Acte de fermeture de dialogue

TAB. 6.4 – Liste des actes du modèle de dialogue

représentation dialogique a par ailleurs été débattu lors de la création du formalisme d'annotation pour la campagne MEDIA. Il est en effet intéressant de remarquer que la notion de requête peut être associée soit (1) uniquement à l'acte de surface de l'énoncé, soit (2) à la pragmatique du discours.

1. Seul l'acte de surface de l'énoncé est considéré. Par exemple l'énoncé *Auriez-vous des restaurants chinois ?* est une requête alors que l'énoncé *je cherche un restaurant chinois* est informatif.
2. Dans le cas où la pragmatique du discours est utilisée pour considérer le niveau dialogique des énoncés, les deux exemples cités précédemment seraient des requêtes. En effet bien qu'informatif au niveau de l'acte de surface, le second exemple est bien une demande d'obtention d'un restaurant chinois.

La décision prise par le consortium MEDIA fut d'utiliser l'énoncé de surface pour annoter les énoncés de la campagne. La raison ayant poussé à ce choix est qu'il permet un meilleur accord inter-annotateur pour des énoncés pris hors contexte. En effet la pragmatique du discours dépend considérablement de l'interprétation contextuelle de l'énoncé considéré.

Ce choix a été suivi dans le cadre de cette thèse pour les mêmes raisons, en considérant le besoin d'annotation pour l'interprétation et la génération en langue naturelle.

#### 6.1.4 Modèle de l'utilisateur simulé

Le modèle présenté dans cette partie ne doit pas être confondu avec le « modèle utilisateur » tel qu'il est généralement cité. Ce dernier est, d'après Fisher [Fischer, 2001], le modèle que le système a d'un utilisateur, qu'il soit obtenu *a priori* par spécification du concepteur ou dynamiquement par inférences du système. Dans ce cas il est un moyen pour le système d'être plus sensible aux besoins d'un utilisateur et aux caractéristiques d'une situation donnée.

Le modèle de l'utilisateur simulé décrit ici un ensemble de paramètres qui permettent de faire varier le comportement du simulateur. Ces paramètres modifient le comportement de façon plus ou moins directe en fonction de leurs niveaux d'abstraction face aux modèles utilisés pour gérer l'interaction. La liste qui suit présente ces différents paramètres.

- **Nombre d'informations relatives à la tâche par tour de parole :** C'est le nombre maximum d'informations que le simulateur fournit en un énoncé.
- **Ordre de présentation des critères :** Ordre dans lequel le simulateur verbalisera ses critères s'il n'est pas contraint par sa stratégie et une question du système, voir ci après.
- **Ordre dans lequel les critères doivent être relâchés :** En cas de requêtes sans solution, le simulateur essaie de relâcher un des critères de recherche, i.e. de permettre qu'il soit transformé en un autre. Si la stratégie du simulateur est directive (voir ci-dessous), le critère à relâcher est imposé par l'ordre dont il question ici, sinon le premier critère proposé par le système évalué est accepté.
- **Utilisateur simulé directif ou non directif :** Si sa stratégie est directive, le simulateur ne cherche pas à répondre aux propositions du système évalué, il prend l'initiative dans le discours à tous les tours de parole pour résoudre la tâche selon ses paramètres. Dans le cas de la stratégie non directive, le simulateur cherche au contraire à répondre aux questions du système<sup>3</sup> à tous les tours de parole.

---

<sup>3</sup>Sans pour autant se répéter ...

### 6.1.5 Algorithme de génération du contenu de la réponse

Il exploite les modèles décrits précédemment et le contenu de l'énoncé produit par le système évalué. Il dispose de compteurs pour chacun des critères et des buts du modèle de la tâche. Ces compteurs décrivent le nombre de fois où un critère ou but a été énoncé, ainsi que le nombre de fois où ce critère ou but a été perçu par le simulateur comme une confirmation du système. Ces variables permettent d'une part d'éviter au simulateur de se répéter, et d'autre part d'analyser le déroulement d'un dialogue sans avoir à le reprendre entièrement.

Le calcul du contenu de la réponse se déroule comme présenté par la figure 6.3. Sur ce schéma les entités représentées par des rectangles (resp. rectangles hachurés) sont des procédures (resp. procédures terminales) et celles représentées par des ellipses les résultats possibles des procédures. Il est par ailleurs à noter sur le schéma que l'algorithme de génération du contenu de la réponse permet une gestion des relaxations de critères ainsi que des confirmations du système évalué.

Une précision concernant l'algorithme du simulateur est qu'il dispose d'un mécanisme mettant fin au dialogue après un nombre prédéfini de tours de parole. Ce mécanisme a pour objet de mettre fin aux boucles d'interaction infinies et de restreindre la résolution de la tâche en nombre de tours de parole. Il est aussi impliqué dans le diagnostic automatique d'aboutissement des dialogues, voir 6.3.4.

## 6.2 Phénomènes perturbateurs

Comme précisé en 5.2.2, les phénomènes perturbateurs sont un moyen de bruiteur les interactions du simulateur. Ce bruitage permet d'obtenir des informations prédictives sur les comportements du système évalué face à des phénomènes comme les hésitations, les auto-corrrections ou les ambiguïtés.

Le fait de dissocier l'inclusion de phénomènes perturbateurs des modèles permettant la gestion de l'interaction est avantageux car il n'est pas nécessaire de modifier ces modèles lors de modifications relatives aux phénomènes perturbateurs (ajout, modification ...). L'approche est donc générique face aux phénomènes perturbateurs modélisables.

L'inconvénient réside alors dans le fait de devoir représenter les phénomènes perturbateurs relativement aux modèles de gestion de l'interaction préexistant. Le cas des phénomènes linguistiques est généralement aisé à traiter puisque ne mettant pas en jeu les comportements dialogiques du simulateur (voir ci dessous en 6.2.1). En revanche les phénomènes dialogiques

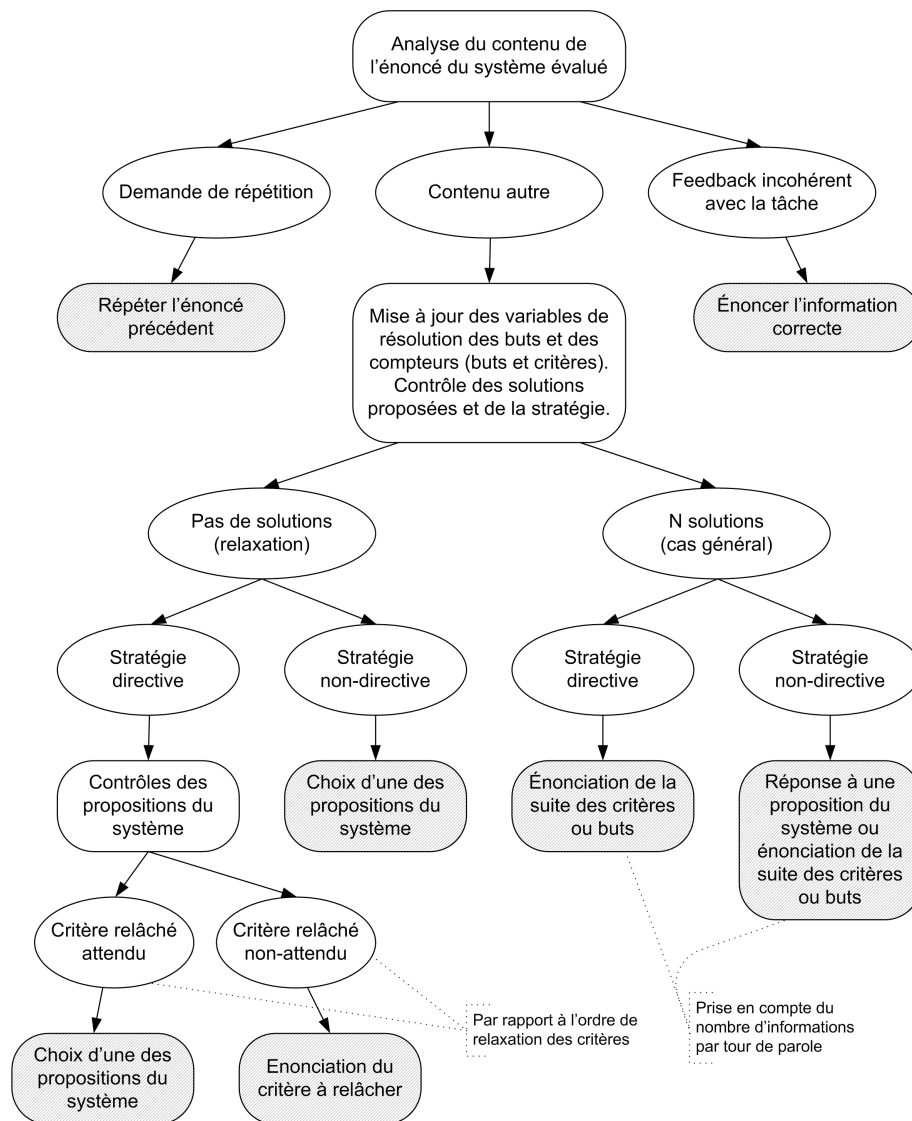


FIG. 6.3 – Algorithme de génération du contenu de la réponse

doivent, pour être générés, être représentés dans les formalismes des modèles de gestion de l'interaction (voir 6.2.2).

### 6.2.1 Hésitation et phénomènes linguistiques

#### L'hésitation

L'ajout d'hésitations ou d'autres phénomènes de l'oral spontané aux énoncés du simulateur ne dépend pas du modèle de l'interaction mais du module de génération en langue naturelle des énoncés. Ce dernier présenté plus loin en 6.3.3 transforme les énoncés du formalisme de représentation de contenu vers la langue naturelle.

La méthodologie employée pour permettre l'insertion d'hésitations dans les énoncés est d'augmenter le formalisme de représentation du contenu des énoncés de la notion d'hésitation. Il existe en effet parmi les concepts et valeurs du modèle du domaine générique présenté en 6.1.1, la paire **bruit/hésitation**. Lorsque cette paire est ajoutée au contenu d'un énoncé, le module de génération en langue naturelle produit un énoncé contenant une hésitation.

Il reste à préciser que la génération d'énoncés contenant des hésitations est réalisée à l'aide de la stratégie de génération des énoncés présentée en 6.3.3 qui nécessite un corpus annoté et un thésaurus sur la tâche considérée. En revanche cette stratégie permet de générer des hésitations pour tous les énoncés souhaités avec seulement une occurrence d'hésitation annotée dans le corpus, d'où un coût très faible en terme d'annotation.

#### D'autres exemples linguistiques

Pour les phénomènes relevant de l'oral spontané comme les auto-corrrections, les répétitions ou les incises, ainsi que pour les erreurs de reconnaissance vocale, la méthodologie employée pour générer les hésitations est exploitable. Il suffit simplement d'ajouter les paires correspondantes aux phénomènes souhaités au formalisme de représentation du contenu, ainsi que d'augmenter le corpus de verbalisation de ces phénomènes.

Du point de vue des annotations du corpus de verbalisation, les phénomènes ne mettant pas en jeu des données relatives à la tâche sont exploitables pour tous les énoncés (les incises et certaines corrections et hésitations). Pour les données relatives à la tâche, soit une ou plusieurs occurrences d'énoncés annotés pour ces phénomènes existent et le problème est résolu, soit il est nécessaire de générer automatiquement ces phénomènes pour les données souhaitées, ce qui n'est pas toujours possible.

Un exemple de phénomène linguistique qui met en jeu le modèle de l'interaction est celui présenté sous le terme d'énoncé complexe. En effet un énoncé est dit complexe quand il contient plusieurs demandes, or la notion de demande est relative la tâche et donc au modèle de l'interaction.



## 6.2.2 Ambiguïté et phénomènes dialogiques

### L'ambiguïté sur les critères de recherche

L'ambiguïté telle qu'elle est modélisée ici ne rend pas compte du phénomène linguistique (resp. pragmatique) qui consiste en la possibilité d'association de plusieurs sens à un mot (resp. une information). L'ambiguïté sur les critères de recherche est définie par le fait de fournir plusieurs informations différentes pour un même critère de recherche au système évalué. Plusieurs stratégies sont alors possibles pour le système évalué, voir le paragraphe sur la désambiguïsation en section 2.3.2.

La génération de ce phénomène est aisée, qu'elle ait lieu sur un seul énoncé ou sur plusieurs. Pour le cas d'une ambiguïté intra-énoncé, il suffit d'ajouter au contenu sémantique une portion informative contenant une valeur différente du ou d'un des critères déjà contenus dans l'énoncé. Pour une ambiguïté inter-énoncé, il suffit d'ajouter au contenu sémantique une portion informative contenant une valeur différente de celle du modèle pour le critère considéré.

Dans le premier cas cité précédemment, l'ambiguïté est immédiate. Dans le second elle est soit immédiate si le critère correspondant à déjà été verbalisé et compris, soit décalée dans le temps si le critère n'a pas été verbalisé. Elle apparaîtra alors à la verbalisation du critère du modèle.

Un point à préciser est que la génération d'ambiguïté, telle qu'elle est proposée ici, n'est pas applicable à des modèles de tâches n'ayant qu'un seul critère de recherche pour leurs buts. En effet dans ce cas tout critère énoncé est directement associé à une ou plusieurs réponses ce qui rend caduque la génération d'une ambiguïté inter-énoncé. Pour cette raison les ambiguïtés n'ont pas été testées avec l'application DIALOGUEBOURSE.

### D'autres exemples dialogiques

L'approche de bruitage des interactions présentée permet de générer d'autres exemples de phénomènes dialogiques, mais la façon de les modéliser ressemble à celle utilisée pour la modélisation des phénomènes linguistiques. En effet la contestation, l'annulation, la demande de remise à zéro ou encore la demande de reformulation sont à modéliser par ajout d'un contenu sémantique relatif au phénomène souhaité.

Il est aussi des phénomènes non modélisables par cette approche. Il faut citer la négation, la disjonction de critère comme dans l'énoncé *je voudrais un chinois ou une pizzeria* ainsi que les contraintes définies négativement comme dans l'énoncé *je voudrais un restaurant asiatique mais pas un chinois*.

En effet même s'il serait possible simplement de faire verbaliser ces exemples par le simulateur, ce dernier ne serait pas à même d'exploiter une réponse du système de dialogue étant donné son modèle de l'interaction.

## 6.3 Autres modules

Un premier point à préciser concerne les modules d'interprétation sémantique et de génération en langue naturelle présentés dans cette section en 6.3.2 et 6.3.3. Ces modules exploitant des corpus annotés, il est à noter qu'une étape de validation de la représentation du contenu des énoncés est nécessaire. Les annotations des corpus utilisés ainsi que leur validation sont ainsi présentés en section 6.3.1.

### 6.3.1 Annotation et validation des corpus

En premier lieu il est à préciser, concernant les annotations, qu'elles ont été effectuées en contexte de dialogue. Ainsi toutes les informations relatives à la tâche ont été annotées sans ambiguïtés grâce au contexte dialogique. Par ailleurs les références n'y ont pas été annotées pour des raisons de complexité de résolution.

Pour la verbalisation d'énoncés par le simulateur, les corpus exploités contiennent 110 dialogues pour la tâche `PLANRESTO` et 40 pour la tâche `DIALOGUEBOURSE`. Ces dialogues ont eu lieu entre des utilisateurs réels et les systèmes et les énoncés verbalisés par les systèmes n'y sont pas annotés.

Pour la tâche `PLANRESTO` le nombre d'énoncés utilisateurs est de 790, soit une moyenne d'environ 7 énoncés par dialogue. À ces 790 énoncés sont associés 1757 annotations dont 800 dialogiques et 957 sémantiques.

Pour la tâche `DIALOGUEBOURSE` 303 énoncés utilisateurs ont été annotés, soit une moyenne d'environ 7 énoncés par dialogue. Ils contiennent 759 annotations dont 312 dialogiques et 447 sémantiques.

Concernant l'interprétation sémantique, les corpus annotés pour les applications `PLANRESTO` et `DIALOGUEBOURSE` contenaient respectivement 52 et 65 énoncés types, et les thésaurus de concepts relatifs à la tâche contenaient 6302 et 1131 valeurs (voir en section `refpara-anx-interp` pour les explications relatives aux énoncés types et aux thésaurus).

Des tableaux fournis en annexe présentent la répartition des annotations dans ces corpus et thésaurus. Pour l'application `PLANRESTO`, les tableaux

A.2 et A.3 sont relatifs aux annotations dialogiques et sémantiques du corpus utilisé pour verbaliser des énoncés, alors que le tableau A.1 expose la répartition des valeurs du thésaurus relatif à la tâche.

De même, pour la tâche *DIALOGUEBOURSE*, les annotations dialogiques et sémantiques du corpus utilisé pour verbaliser des énoncés sont présentées dans les tableaux B.2 et B.3, et la répartition des valeurs du thésaurus relatif à la tâche exposé dans le tableau B.1.

Enfin une validation des annotations effectuées a été réalisée par un protocole de double annotation. Une double annotation a été mise en œuvre pour 10 dialogues extraites aléatoirement du corpus de verbalisation de l'application *PLANRESTO*. Ces 10 dialogues contenaient 66 énoncés, représentant 67 annotations dialogiques et 87 annotations sémantiques. Les coefficients Kappas de Cohen [Cohen, 1960] obtenus sont de 0,88 (resp. 0,84) pour les annotations dialogiques (resp. sémantiques), ce qui valide la cohérence du formalisme de représentation du contenu des énoncés.

### 6.3.2 Interprétation sémantique

Ce module « minimal » transforme les énoncés du système évalué en leur contenu dialogique et sémantique dans le formalisme décrit en 5.2.2. Cette transformation est réalisée en deux temps.

Une première passe associe, par la méthode d'îlots sémantiques, des portions de l'énoncé à transformer à des portions d'énoncés annotées. Les portions d'énoncés de référence annotées proviennent d'une part de l'analyse des énoncés produits par le système à évaluer ainsi que de l'ensemble des valeurs du domaine considéré (voir le tableau A.1 en annexe pour plus de détails). Cette passe permet d'obtenir le contenu sémantique de l'énoncé ainsi que certains contenus dialogiques, dans le cas d'énoncés types où le contenu sémantique et dialogique est connu.

La deuxième passe associe des actes de dialogue aux portions d'énoncés dont le contenu sémantique n'est pas encore lié à un contenu dialogique. Cette association a lieu par analyse de la ponctuation et de marqueurs linguistiques. Par exemple une portion d'énoncé clôturée par un point (resp. un point d'interrogation) est considérée comme une assertion (resp. une requête).

Ce module « minimal » permet une interprétation plus ou moins robuste des énoncés du système évalué (voir le chapitre 7). Il pourrait être amélioré en y ajoutant par exemple la prise en compte de variantes flexionnelles pour les portions d'énoncés de référence.

### 6.3.3 Génération en langue naturelle

Ce module transforme les énoncés produits par le module de gestion de l'interaction du format de représentation du contenu vers des énoncés en langue naturelle. Il permet notamment une génération d'énoncé totalement automatique ou semi-automatique.

Lors d'une évaluation automatique par simulation d'utilisateurs, le système évalué génère un énoncé et le simulateur lui répond. L'évaluation semi-automatique proposée ici correspond à une évaluation où les énoncés fournis en entrée du système évalué sont contrôlés par un humain avant d'être envoyés au système évalué.

Au lieu de ne générer qu'un seul énoncé correspondant au contenu souhaité (voir la stratégie automatique), le simulateur génère tous les possibles en une liste d'énoncés (voir la stratégie semi-automatique). Cette liste est proposée à l'évaluateur humain qui choisit un des énoncés de la liste ou propose un nouvel énoncé correspondant au contenu. Cette approche permet notamment de générer de façon certaine des énoncés ressemblant à ceux d'utilisateurs réels à partir de corpus annotés.

Le dernier paragraphe de cette section est consacré à une présentation des résultats du test perceptif contrôlant la qualité des énoncés produits par la stratégie automatique de génération.

#### Stratégie automatique

Le passage d'un énoncé du formalisme de représentation vers la langue naturelle est réalisé par appariement de contenu sémantique et/ou dialogique avec des énoncés ou des portions d'énoncés d'utilisateurs annotés. Cette méthode requiert donc un corpus annoté à partir duquel les énoncés du simulateur seront générés.

L'algorithme de génération de l'énoncé exploite une succession de recherches sur les annotations des énoncés du corpus. Si une recherche ne donne pas de résultat, la suivante est essayée. Chacune de ces recherches autorise, par rapport à la précédente, soit une déconstruction structurelle du contenu de l'énoncé à générer, soit un surensemble de ce contenu. En d'autres termes cette stratégie réduit à chaque recherche les contraintes sur l'ensemble des portions d'énoncés recherchées, d'où un plus grand choix à chaque étape. Lorsque tout le contenu de l'énoncé a été trouvé, sa verbalisation est construite par concaténation.

Une description séquentielle de l'algorithme est présentée ci dessous.

1. *Recherche exacte* : un énoncé correspondant exactement aux annota-

tions souhaitées<sup>4</sup> est cherché. S'il en existe plusieurs, un choix pseudo-aléatoire permet de n'en obtenir qu'un. Dans le cas où il n'y en a pas dans la base ...

2. *Recherche élargie* : La requête porte sur un énoncé contenant un sur-ensemble des annotations prévues (i.e. un énoncé exprimant une requête plus détaillée ou plus complète). S'il en existe, un choix pseudo-aléatoire permet de n'en obtenir qu'un qui est alors nettoyé des portions d'énoncés ne correspondant pas aux annotations attendues. Dans le cas où il n'en y a pas dans la base ...
3. *Déconstruction dialogique* : Une portion d'énoncé est recherchée pour chacun des actes de dialogue de la liste, indépendamment des autres. Ces portions sont d'abord cherchées en tant qu'énoncés complets puis en tant que sous-ensembles d'énoncés. L'énoncé final est reconstruit à partir des portions obtenues en conservant l'ordre initial des actes de dialogue. Dans le cas où un acte n'est pas trouvé de façon unitaire ...
4. *Déconstruction sémantique* : Le contenu sémantique de cet acte est alors cherché indépendamment de l'acte qui le contient. S'il existe des énoncés contenant les annotations sémantiques souhaitées, un choix pseudo-aléatoire permet de n'en obtenir qu'un, dont ne sera utilisée que la portion dont les annotations sémantiques correspondent à celles attendues. Si le contenu sémantique décorrélé de l'acte n'est pas trouvé ...
5. *Reconstruction* : Chaque triplet sémantique du contenu est recherché de façon indépendante. La portion d'énoncé est alors reconstruite en agglomérant chacun des segments correctement annotés dans un ordre aléatoire. Si pour un ou plusieurs triplets cette recherche s'avère infructueuse ...
6. *Verbalisation minimale* : Les segments retournés sont ceux utilisés pour l'interprétation des énoncés, ce qui permet à la plateforme de verbaliser tout ce qu'elle est en mesure d'interpréter.

### Stratégie semi-automatique

La stratégie semi-automatique est la même que la stratégie automatique si ce n'est qu'au lieu de choisir une verbalisation parmi les possibilités, elle les garde toutes. Les énoncés reconstruits à partir des portions de contenus sont ensuite proposés à un évaluateur humain. Ce dernier a alors le choix

---

<sup>4</sup>L'ordre des actes de dialogue est pris en compte mais pas l'ordre des triplets sémantiques contenus par un même acte.

d'entrer un nouvel énoncé ou d'en prendre un dans la liste proposée par le simulateur. Une fois le choix réalisé, l'énoncé est envoyé au système évalué.

L'étape de validation / modification humaine des énoncés verbalisé par la stratégie automatique permet d'assurer la qualité des énoncés envoyés aux systèmes évalués, même si celle-ci est généralement bonne (voir le paragraphe sur les tests perceptifs). Appliquée à des évaluations comparatives, cette stratégie permet de placer les systèmes dans les mêmes conditions non seulement au niveau des comportements dialogiques mais aussi au niveau des énoncés fournis en entrée des systèmes. Ainsi le biais de la « génération automatique de langue naturelle » est contrôlé.

Un point à ajouter concernant ces stratégies de génération d'énoncés est qu'elles permettent de générer des énoncés pour toutes les valeurs des différents attributs à condition de disposer d'une façon de les verbaliser. L'utilisation de l'ensemble des verbalisations des valeurs du domaine utilisé pour l'interprétation permet ainsi de verbaliser toutes les valeurs des différents attributs pour lesquelles il n'y a pas d'occurrence dans le corpus annoté. Il est alors possible de « doper » les corpus, par exemple en testant le système sur toutes les valeurs de tous les attributs, le dopage étant ici considéré comme un moyen artificiel de faire augmenter la complétude d'un corpus par rapport à son domaine d'application.

### **Test perceptif**

L'objet de ce test a été d'évaluer l'aspect naturel ou non, le sens et la syntaxe des énoncés générés par la stratégie automatique de verbalisation. Ont été présentés aux sujets 50 énoncés, dont 25 étaient produits par la stratégie de verbalisation et 25 étaient des transcriptions automatiques d'énoncés humains (voir en annexe les tableaux C.2 et C.3). L'ordre de présentation des énoncés était aléatoire et pour chacun d'eux était demandé :

- Si l'énoncé semblait naturel (produit par un humain), non naturel (produit par une machine), ou moyennement naturel ;
- Si le sens des énoncés était compréhensible. Une note sur 100 était à associer aux énoncés où 100 était un énoncé dont le sens était entièrement compréhensible et 1 un énoncé n'ayant aucun sens ;
- Si la syntaxe des énoncés paraissait correcte ou incorrecte. Une troisième possibilité notée « Sans avis » était proposée.

Le test a été présenté à 25 sujets, dont 7 femmes et 18 hommes. L'âge des participants était majoritairement (70%) compris entre 20 et 30 ans et tous disaient utiliser un ordinateur plusieurs fois par semaine.

Les résultats détaillés en annexe dans le tableau C.1 montrent que la différence entre les énoncés produits par la stratégie de verbalisation les transcriptions automatiques d'énoncés humains n'est pas perceptible, tant du point de vue de l'aspect que du sens ou de la syntaxe. En effet que les énoncés aient réellement été produits par des humains ou par la stratégie de verbalisation, les réponses concernant l'aspect, le sens et la syntaxe sont identiques à moins de 5 points près.

De plus des accords inter-annotateur ont été calculés en appliquant le Kappa de Fleiss [Fleiss, 1971] aux résultats sur l'aspect et la syntaxe. Les coefficients de 0,25 pour l'aspect et 0,52 pour la syntaxe peuvent être interprétés d'après [Landis and Koch, 1977] comme « passable » et « bon ». Étant donné le caractère très subjectif de la notion d'aspect, le coefficient d'accord obtenu pour cette modalité semble suffisamment élevé pour valider la cohérence du test. Par ailleurs l'accord obtenu pour la modalité syntaxe est quant à lui plus élevé, confirmant la cohérence des résultats obtenus.

Enfin pour le sens des énoncés, la cohérence des résultats a été mesurée à l'aide de l'Alpha de Cronbach [Cronbach, 1951], puisque l'échelle des valeurs possibles était continue. Là encore la cohérence du résultat est confirmée par un coefficient de 0,95.

Les tests réalisés sur les énoncés produits par la stratégie de verbalisation automatique rendent notamment valide des évaluations automatiques de composants d'interprétation sémantique par cette stratégie. En effet puisque la différence entre des transcriptions automatiques d'énoncés produits par des humains et des énoncés produits par la stratégie présentée ci-dessus n'est pas perceptible, l'utilisation de cette stratégie plutôt que d'énoncés humains pour évaluer semble légitime. L'intérêt d'utiliser cette stratégie réside alors dans la taille du corpus à utiliser pour réaliser les évaluations. Dans le cas où l'évaluation a lieu avec des énoncés d'utilisateurs réels, tous les énoncés testés doivent être annotés. En utilisant la stratégie proposée, la taille du nombre d'énoncés annotés peut être considérablement réduite.

### 6.3.4 Diagnostic automatique

Ce module permet d'obtenir des informations sur les dialogues auxquels a pris part le simulateur. Trois phénomènes sont diagnostiqués par ce module, à savoir l'aboutissement des dialogues relativement à la tâche, la relaxation de critères de recherches et la désambiguïsation des critères de recherche. L'aboutissement des dialogue par rapport à la tâche, parfois nommé complétude de la tâche, est un critère d'évaluation considéré comme standard aujourd'hui pour l'évaluation des SDHM finalisés. Par ailleurs l'évaluation

face aux critères de relaxations de contraintes et de désambiguïsation est intéressant car ces phénomènes essentiels pour la réussite des dialogues entre humains commencent à être pris en compte par les SDHM développés aujourd'hui.

### **Abouti vs. Non abouti**

Un dialogue est dit abouti quand le simulateur a réussi à réaliser sa tâche, ou en d'autres termes à atteindre tous les buts de sa liste de buts. S'il n'y arrive pas, le dialogue est dit non abouti. La méthode de diagnostic de l'aboutissement repose sur le fait que le simulateur ne mette fin à l'interaction que s'il a réussi à réaliser sa tâche ou si le nombre de tours de parole qu'il a dû générer dépasse un seuil (voir 6.1.5). Un dialogue abouti est donc un dialogue dont le nombre de tours de parole est inférieur au nombre de tours de parole seuil.

Ce phénomène est particulièrement intéressant à diagnostiquer relativement à l'évaluation de systèmes de dialogue finalisés puisqu'il rend compte du fait que la tâche du simulateur ait été résolue ou pas.

### **Relaxation de critères**

Comme présenté précédemment, le simulateur permet qu'un de ses critères soit transformé en un autre en cas de requêtes sans solution. Ce phénomène est diagnostiqué grâce à la différence de comportements qu'il induit pour le simulateur. Les comportements du simulateur sont, pour chacun des dialogues auxquels il a pris part, enregistrés dans un fichier dont la lecture permet de savoir si le simulateur est passé dans le mode de relaxation, et si oui, pour combien de tours de parole.

Les faux positifs, s'il-en-est, sont ici dus à de mauvaises interprétations de la part du simulateur. De mauvaises interprétations peuvent mener à des comportements de relaxation pour le simulateur sans que ces derniers soient justifiés du point de vue du système évalué.

Cette mesure du nombre de tours de parole où le simulateur relâche des critères permet notamment d'expliquer certaines différences en terme de nombre de tours de parole pour résoudre la tâche.

### **Désambiguïsation**

La désambiguïsation de critères de recherche correspond à une demande de choix parmi un ensemble de valeurs proposée par le système évalué. Par exemple l'énoncé *Voulez-vous un restaurant chinois ou japonais ?* est



considéré comme une désambiguïsation. Le diagnostic de ce phénomène est réalisé d'après les informations contenues dans l'énoncé du système évalué.

Le diagnostic de ce phénomène est intéressant face à la génération d'ambiguïtés sur les critères de recherche en tant que phénomènes perturbateurs. En effet ces ambiguïtés telles qu'elles sont modélisées peuvent mener à différentes stratégies de résolution de la part du système. Le fait de diagnostiquer les désambiguïssations permet d'analyser la ou les différentes stratégies mises en œuvre par le système face à la génération d'une ambiguïté.

## 6.4 Coût de transformation du simulateur

La première version du simulateur a été réalisée afin d'interagir avec l'application PLANRESTO en utilisant des modèles facilement réutilisable pour d'autres tâches de recherche d'information. Une évaluation précise du coût de transfert vers une autre tâche a ensuite été évalué lors de la modification du simulateur pour l'évaluation de l'application DIALOGUEBOURSE.

La transformation, effectuée par une seule personne, a demandé 11 jours de travail. La répartition du temps nécessaire à la modification se répartie en :

- Deux journées pour annoter un corpus de verbalisation de 40 dialogues ;
- Trois jours pour modifier le module d'interprétation sémantique (analyse et annotation des verbalisations types du système et mise au format du thésaurus relatif à la tâche) ;
- Trois jours pour développer et implémenter le modèle de la tâche ;
- Trois jours pour finaliser le simulateur entier.

Le faible coût de transformation du simulateur de la tâche de recherche de restaurant vers la tâche de recherche d'information boursière démontre la facilité avec laquelle il est possible de modifier le simulateur pour interagir vers différentes tâches de recherche d'information.

Ce niveau de généralité par rapport à la tâche est particulièrement intéressant car le simulateur est modélisé par expertise et non par des méthodes stochastique. Ces dernières sont généralement aisément réutilisable puisqu'elles sont applicables à n'importe quelles données structurées de façon appropriée. Leur inconvénient réside alors dans la nécessité d'un gros corpus annoté pour l'apprentissage.

La méthode proposée est de définir les comportements du simulateur par expertise via des modèles suffisamment générique pour être aisément transformables et réutilisables, d'où le faible coût d'adaptation du simulateur à

une nouvelle tâche.

## 6.5 Conclusion

Ce chapitre a présenté la modélisation du simulateur d'utilisateurs réalisé pour évaluer des systèmes de dialogue avec le paradigme SIMDIAL. Ce simulateur a été implémenté dans langage Java pour deux tâches, l'une de recherche de restaurants à Paris et l'autre de recherche d'informations boursières.

Concernant la gestion de l'interaction les modèles choisis sont à la fois simple et génériques, ce qui permet une bonne maîtrise des comportements du simulateur ainsi qu'une flexibilité importante en terme de réutilisabilité. Par ailleurs les paramètres du modèle de l'utilisateur simulé fournissent une grande variabilité de comportements face auxquels adapter des mesures comparatives. En outre l'adjonction de phénomènes perturbateurs de façon décorrélée du modèle de l'interaction enrichie encore la variabilité des phénomènes observables. Enfin le diagnostic automatique de la complétude de la tâche associé à des mesures de performance sur la durée des dialogue en nombre de tours de parole permettent l'interprétation des comportements des systèmes évalués faces aux phénomènes observables en terme d'efficacité. Il reste à préciser que l'indépendance des mesures d'évaluation relativement aux systèmes évalués est un point important concernant la généralité du paradigme SIMDIAL.

Un autre point à noter concerne les méthodes employées pour générer des énoncés en langue naturelle. Une méthode semi-automatique de génération a été proposée qui permet le contrôle humain de la qualité des énoncés fournis aux systèmes évalués, minimisant ainsi le biais de la « génération automatique de la langue naturelle ». De plus la méthode proposée permet théoriquement l'évaluation automatique de composants d'interprétation sémantique face à des domaines complets sans nécessiter de gros corpus sur ces domaines.

Enfin du point de vue de la généralité le simulateur présenté est d'une part facilement adaptable à différentes tâches de recherches d'information et d'autre part permet théoriquement l'interaction avec différents systèmes réalisant la même tâche.

Le chapitre suivant expose les résultats obtenus en évaluant les applications PLANRESTO et DIALOGUEBOURSE avec le paradigme.



# Chapitre 7

## Deux exemples d'applications du paradigme SIMDIAL

Ce chapitre expose les résultats obtenus lors de l'évaluation des systèmes de recherche d'information PLANRESTO et DIALOGUEBOURSE. Seront notamment analysées les capacités des systèmes à résoudre les tâches telles que modélisées dans le simulateur, ainsi que leur résistance aux phénomènes perturbateurs étudiés.

L'objectif de ces expériences est, plus que d'obtenir des résultats d'évaluation sur les applications PLANRESTO et DIALOGUEBOURSE, de valider le paradigme SIMDIAL sur des systèmes réalisant différentes tâches. Il n'en reste pas moins vrai que les résultats obtenus sont très intéressants quant à d'éventuelles améliorations des systèmes évalués.

### 7.1 Évaluation de l'application PlanResto

#### 7.1.1 Observations

Lors de l'évaluation de l'application PLANRESTO, le comportement du simulateur variait d'un ensemble de dialogues à l'autre. Ces variations portaient sur :

- **Le nombre d'informations relatives à la tâche par énoncé.** Ce paramètre est considéré comme lié à l'expertise des utilisateurs. En effet un utilisateur novice aura tendance à ne verbaliser qu'une (voire aucune) information pertinente par tour de parole alors qu'un utilisateur expert fournira généralement toutes les informations pertinentes en un seul tour de parole. Le contrôle de ce paramètre permet donc de mesurer la différence d'efficacité entre experts et novices.

- **La stratégie directive ou non de l'utilisateur simulé.** Les différences générées par ce paramètre repose sur les capacités de gestion de l'initiative des systèmes évalués. D'une part ces stratégies étant rigides elles permettent l'observation d'erreurs que des humains auraient naturellement corrigées, et d'autre part elles rendent possible la comparaison d'efficacité en fonction de chacune d'elle.
- **L'ajout d'une hésitation par énoncé produit par le simulateur.** La robustesse de l'application évaluée face aux hésitations peut ainsi être contrôlée.
- **L'ajout d'une ambiguïté sur les critères de recherche par dialogue.** ce paramètre permet d'observer les stratégies de désambiguïsation de l'application évaluée, ainsi que mesurer l'efficacité de dialogues contenant des ambiguïtés.

Les paramètres que le simulateur permet de faire varier sont notamment représentatifs de comportements humains qui sont problématiques pour les systèmes de dialogue. Ainsi les résultats obtenus portent sur des phénomènes directement exploitables en terme d'ergonomie.

L'observation des dialogues auxquels a pris part le simulateur a porté sur les critères suivant :

- L'aboutissement ou non des dialogues ;
- La présence de relaxation de critères de recherche ;
- La présence de désambiguïsation ;
- Le nombre de tours de parole nécessaires à réaliser la tâche.

Il est par ailleurs à noter que d'autres observables sont possibles à ajouter, comme par exemple les erreurs d'interprétation.

Le simulateur a ainsi permis de générer 2400 dialogues présentés dans les tableaux 7.1 et 7.2, dont les différents cas ont été obtenus grâce à de simples modifications de paramètres. Dans ces tableaux, les termes « dialogues », « tours de parole » « relaxation de critère », « désambiguïsation », « hésitation » et « ambiguïté » ont été abrégés respectivement en « Dial », « TdP » « rel. », « dés. », « H » et « A ».

Dans le tableau 7.1, les dialogues présentés comme contenant des relaxations ou des désambiguïsations ne sont que des dialogues ayant aboutis. en effet l'ajout des dialogues non-aboutis à ces ensembles ne fournirait pas d'informations quant à ces derniers et biaiserait les calculs sur l'efficacité relati-

vement à ces paramètres. De plus les tours de paroles associés aux dialogues précédemment cités sont ceux qui contiennent des informations relatives à une relaxation ou à une désambiguïsation, dans l'objectif de pouvoir mesurer le nombre de tours de paroles associés à ces phénomènes.

				<b>Total</b>	<b>Aboutis</b>	<b>Avec rel.</b>	<b>Avec dés.</b>
		<b>H</b>	<b>A</b>	Dial – TdP	Dial – TdP	Dial – TdP	Dial – TdP
<b>Directif</b>	–	–		300 – 1901	277 – 1533	122 – 232	64 – 83
	✓	–		300 – 1965	278 – 1613	125 – 274	65 – 87
	–	✓		300 – 2432	240 – 1472	114 – 217	103 – 134
	✓	✓		300 – 2155	261 – 1531	119 – 208	106 – 128
<b>Non dir.</b>	–	–		300 – 1745	278 – 1393	118 – 144	43 – 49
	✓	–		300 – 1617	286 – 1393	112 – 143	41 – 45
	–	✓		300 – 2182	250 – 1382	116 – 164	84 – 102
	✓	✓		300 – 2176	243 – 1264	114 – 141	70 – 78
<b>Total</b>				2400 – 16173	2113 – 11581	940 – 1523	576 – 706

TAB. 7.1 – Dialogues générés et tours de paroles associés pour la tâche PLANRESTO

		<b>H</b>	<b>A</b>	<b>Aboutis</b>	<b>Avec rel.</b>	<b>Avec dés.</b>
<b>Directif</b>	–	–		5,53	1,90	1,30
	✓	–		5,80	2,19	1,34
	–	✓		6,13	1,90	1,30
	✓	✓		5,87	1,75	1,21
<b>Non dir.</b>	–	–		5,01	1,22	1,14
	✓	–		4,87	1,28	1,10
	–	✓		5,53	1,41	1,21
	✓	✓		5,20	1,24	1,11
<b>Total</b>				5,48	1,62	1,23

TAB. 7.2 – Tours de parole par dialogue pour la tâche PLANRESTO

### 7.1.2 Analyses

#### Sur l'aboutissement des dialogues

Le premier résultat est que 2113 dialogues sur 2400 ont abouti, soit 88%. Le diagnostic automatique d'aboutissement des dialogues permet notamment de n'analyser manuellement que les dialogues n'ayant pas abouti, soit 287 par rapport à 2400.

Un contrôle manuel effectué sur 100 dialogues a permis de valider la qualité du diagnostic d'aboutissement puisqu'aucun faux-positif ni faux-négatif n'a été relevé. Un faux-positif est ici un dialogue qui aurait été classé comme ayant abouti alors que la tâche n'aurait pas été réalisée et un faux-négatif un dialogue qui aurait été classé comme non abouti alors que la tâche y aurait été réalisée.

L'analyse manuelle des dialogues non aboutis permet de distinguer 2 types d'erreurs. L'un est relatif à la gestion de l'initiative mixte dans l'application PLANRESTO et l'autre est dû à des erreurs d'interprétation des énoncés du système évalué. Le premier représente environ 40% des dialogues non aboutis et le second 60%, soit environ 170 dialogues sur 2400.

L'erreur de gestion d'initiative mixte a lieu lors de la relaxation de critère quand le comportement du simulateur est directif. Le simulateur cherche dans ce cas à imposer le critère à relâcher tandis que l'application PLANRESTO propose des solutions avec un autre relâchement de critère. L'application PLANRESTO attend une réponse à ses propositions et ne comprend pas que le simulateur lui demande autre chose.

Les résultats obtenus démontrent que les modèles de l'interaction mis en œuvre dans le système évalué ainsi que dans le simulateur sont fortement cohérents, ou en d'autres termes que l'application PLANRESTO réalise assez bien la tâche que le simulateur cherche à effectuer via le dialogue. Les divergences observées entre les modèles résident dans la gestion de l'initiative-mixte face à une stratégie d'utilisateurs strictement directive.

#### Sur l'efficacité en fonction de la stratégie

Il apparaît que le nombre de tours de parole par dialogue abouti est plus faible de 11,68% lorsque la stratégie de l'utilisateur simulé est non directive, soit 5,82 tours de parole en moyenne pour la stratégie directive et 5,14 pour la stratégie non directive. Bien l'écart soit faible, ce résultat est observable sur l'ensemble des dialogues, quelque soit le nombre d'informations fournies par tour de parole et que les dialogues contiennent des relaxations ou pas.

Il est aussi intéressant de préciser que la différence entre stratégie directive et non directive sur le nombre de tours de parole contenant des relaxations est de 33,50%, ce qui est beaucoup plus significatif. Enfin la corrélation de ces deux données montre qu'environ 50% des 5% de différence vus plus haut sont dus à la relaxation de critère. En d'autres termes la moitié de la différence du nombre de tours de parole entre un utilisateur directif et non directif est expliquée par la relaxation de critère.

L'interprétation de ces résultats est intuitivement aisée. En effet un utilisateur directif impose son critère à relâcher et il y a assez peu de chance pour que ce dernier soit celui proposé par le système. S'ensuivent alors des tours de parole pour la désambiguïsation du critère choisi par le simulateur, ce qui peut même parfois mener à d'autres relaxations. Un utilisateur non directif se laissant guider, accepte directement une des solutions qui lui sont proposées, sans besoin de tours de parole supplémentaires.

La faible différence observée entre les stratégies sur le nombre de tours de parole montre que l'application PLANRESTO gère plutôt bien l'initiative mixte. En effet elle tolère généralement que l'utilisateur simulé ne réponde pas à ses propositions, ce qui est cohérent avec le « dialogue naturel » que permet la technologie ARTIMIS.

### **Sur le phénomène perturbateur hésitation**

Les résultats obtenus sur l'hésitation démontrent que l'application PLANRESTO est robuste face à ce phénomène. En effet la différence du nombre de tours de parole pour les dialogues auxquels ont été ajoutés des hésitations par rapport à ceux auxquels il n'en a pas été ajouté, est de moins de 2%, soit 5,43 tours de parole pour les dialogues ne contenant pas d'hésitations et 5,53 pour ceux qui en contenaient.

Il est à noter que des SDHM dont les modules d'interprétations sont basés sur des analyses syntaxiques des énoncés auraient été plus perturbés par ce phénomène que ne l'a été l'application PLANRESTO. En effet cette dernière interprète les énoncés par la technique d'îlots sémantiques, ce qui lui permet d'être robuste face à de nombreux phénomènes de l'oral spontané.

### **Sur le phénomène perturbateur ambiguïté**

Sur l'ambiguïté les résultats sont plus intéressants. En effet pour 1200 dialogues où étaient générées des ambiguïtés, seuls 442 ont été perçus comme contenant des désambiguïsations. Bien que ces 442 dialogues soient plus du



double de ceux contenant des désambiguïisations sans ajout d’ambiguïté, ce nombre reste faible.

L’analyse manuelle de dialogues auxquels avaient été ajoutés des ambiguïtés permet de fournir deux explications au fait que peu de désambiguïisations aient été générées en retour. La première est que l’application PLANRESTO permet de réaliser des requêtes avec des disjonctions de critères, ce qui transforme certaines des ambiguïtés générées en requêtes sur plusieurs valeurs d’un même critère. Par exemple en réponse à l’énoncé « *un restaurant traditionnel euh spécialité fruits-de-mer près de de Bastille pour cent dix euros* », l’application PLANRESTO répond « *Il n’y a pas de restaurant de type restaurant avec spécialité de fruits de mer ou restaurant traditionnel pour environ cent dix euros près de la Place de la Bastille. Par contre ...* ». La seconde explication provient du fait que l’application PLANRESTO n’applique pas toujours une stratégie de désambiguïisation lorsque deux critères différents sont présentés pour une même requête. En effet parfois l’un des deux critères est directement omis, comme dans par exemple lorsque le simulateur demande « *pour environ dix euros dans le huitième arrondissement euh euh pour plus de cinquante euros je recherche un restaurant de type algérien* » et que PLANRESTO répond « *Il n’y a pas de restaurant algérien pour environ dix euros dans le 8eme arrondissement. Par contre ...* ».

### 7.1.3 Exemples de dialogues

Trois dialogues ayant eu lieu entre l’application PLANRESTO et le simulateur d’utilisateurs sont présentés ci-dessous. Le premier (tab. 7.3) est comparable à notre avis à un dialogue entre un utilisateur réel et le système tant du point de vue de la résolution de la tâche que des verbalisations. Le second (tab. 7.4) contient des verbalisations que l’on a jugées trop distantes de verbalisations « humaines » même si la résolution de la tâche a été menée à bien. Le troisième (tab. 7.5) montre le problème de gestion de l’initiative mixte présenté ci-avant. Au huitième tour de parole une boucle d’interaction infinie est engagée où système et simulateur répètent tous les deux la même chose à chaque nouveau tour de parole. Le simulateur a mis fin au dialogue quand le nombre maximum de tour de paroles pour un dialogue a été dépassé. Un autre comportement pourrait être par exemple d’effectuer une nouvelle recherche plutôt que de clore le dialogue.

PlanResto	:Bienvenue sur PlanResto. Ce service vous permet de trouver un restaurant à Paris et d'être mis en relation téléphonique avec lui. Vous pouvez à tout moment obtenir une aide en disant mode d'emploi général. Que désirez-vous ?
Simulateur	:je cherche un restaurant dans le cinquième arrondissement
PlanResto	:Vous recherchez un restaurant dans le 5eme arrondissement. J'en ai trouvé plus de 10. Vous pouvez donner plus de précisions, par exemple la spécialité, consulter les solutions ou effectuer une nouvelle recherche.
Simulateur	:je recherche un restaurant antillais
PlanResto	:J'ai trouvé deux restaurants antillais dans le 5eme arrondissement. Le premier restaurant est le restaurant "La Barbacane". Souhaitez-vous plus d'informations, le restaurant suivant ou effectuer une nouvelle recherche ?
Simulateur	:plus d'informations
PlanResto	:Le restaurant "La Barbacane" est situé 13 rue Cardinal Lemoine. Son numéro de téléphone est le 01 43 26 37 01. Voulez-vous d'autres informations, être mis en relation avec ce restaurant, le restaurant suivant ou effectuer une nouvelle recherche ?
Simulateur	:quitter
PlanResto	:Voulez-vous réellement quitter PlanResto ?
Simulateur	:oui

TAB. 7.3: Dialogue entre PLANRESTO et le simulateur ressemblant à un dialogue humain-machine

PlanResto	:Bienvenue sur PlanResto. Ce service vous permet de trouver un restaurant à Paris et d'être mis en relation téléphonique avec lui. Vous pouvez à tout moment obtenir une aide en disant mode d'emploi général. Que désirez-vous ?
Simulateur	:restaurant pour plus de cinquante euros Un restaurant traditionnel
Suite à la page suivante	

PlanResto :Vous recherchez un restaurant traditionnel pour plus de cinquante euros. J'en ai trouvé plus de 10. Vous pouvez donner plus de précisions, par exemple le lieu, consulter les solutions ou effectuer une nouvelle recherche.

Simulateur :Dans le troisième arrondissement

PlanResto :Il n'y a pas de restaurant traditionnel pour plus de cinquante euros dans le 3eme arrondissement. Par contre, il y a des restaurants traditionnels pour plus de cinquante euros dans le 2eme arrondissement, des restaurants traditionnels pour plus de cinquante euros dans le premier arrondissement et des restaurants traditionnels pour plus de cinquante euros dans le 4eme arrondissement. Quelle proposition vous intéresse ?

Simulateur :Un restaurant pour plus de cinquante euros je recherche  
Un restaurant japonais Dans le troisième arrondissement

PlanResto :Il n'y a pas de restaurant de type restaurant japonais ou restaurant traditionnel pour plus de cinquante euros dans le 3eme arrondissement. Par contre, il y a des restaurants de type restaurant japonais ou restaurant traditionnel pour plus de cinquante euros dans le premier arrondissement, des restaurants de type restaurant japonais ou restaurant traditionnel pour plus de cinquante euros dans le 2eme arrondissement et des restaurants de type restaurant japonais ou restaurant traditionnel pour plus de cinquante euros dans le 4eme arrondissement. Quelle proposition vous intéresse ?

Simulateur :je désire manger japonais

PlanResto :Vous recherchez un restaurant de type restaurant japonais ou restaurant traditionnel pour plus de cinquante euros dans le premier arrondissement. J'en ai trouvé 8. Vous pouvez donner plus de précisions, par exemple le lieu, consulter les solutions ou effectuer une nouvelle recherche.

Simulateur :consulter

PlanResto :Le premier restaurant est le restaurant "Kinugawa". Il propose des repas à environ cent sept euros le midi et des repas à environ cent sept euros le soir. Voulez-vous plus d'informations, les restaurants suivants ou effectuer une nouvelle recherche ?

Suite à la page suivante

<p>Simulateur :au revoir  PlanResto :Voulez-vous réellement quitter PlanResto ?  Simulateur :oui</p>
--

TAB. 7.4: Dialogue avec l'application PLANRESTO où les énoncés du simulateur ne semblent pas humains

<p>PlanResto :Bienvenue sur PlanResto. Ce service vous permet de trouver un restaurant à Paris et d'être mis en relation téléphonique avec lui. Vous pouvez à tout moment obtenir une aide en disant mode d'emploi général. Que désirez-vous ?</p> <p>Simulateur :Riquet moins de cinq euros</p> <p>PlanResto :Il n'y a pas de restaurant pour moins de cinq euros près du métro Riquet. Par contre, il y a des restaurants pour environ neuf euros le midi près du métro Riquet, des restaurants pour environ onze euros le midi près du métro Riquet et des restaurants pour environ quatorze euros le midi près du métro Riquet. Quelle proposition vous intéresse ?</p> <p>Simulateur :moins de cinq euros euh sud-ouest dix-huitième arrondissement</p> <p>PlanResto :Je ne peux pas effectuer une recherche sur un lieu et un arrondissement ensemble, j'ignore l'arrondissement. Il n'y a pas de restaurant du Sud-Ouest pour moins de cinq euros près du métro Riquet. Par contre, il y a des restaurants du Sud-Ouest pour environ vingt-trois euros le midi près du métro Ourcq, des restaurants du Sud-Ouest pour environ soixante-et-un euros le midi près du métro Louis Blanc et des restaurants du Sud-Ouest pour environ quarante-six euros le midi près du Parc des Buttes-Chaumont. Quelle proposition vous intéresse ?</p> <p>Simulateur :euh sud-ouest prix de moins de dix euros dix-huitième arrondissement</p>
---

Suite à la page suivante

PlanResto :Vous recherchez un restaurant du Sud-Ouest près du métro Riquet. Je ne peux pas effectuer une recherche sur un lieu et un arrondissement ensemble, j'ignore l'arrondissement. Un restaurant pour moins de cinq euros ou un restaurant pour moins de dix euros?

Simulateur :dix-huitième arrondissement

PlanResto :Je ne peux pas effectuer une recherche sur un lieu et un arrondissement ensemble, j'ignore l'arrondissement. Il n'y a pas de restaurant du Sud-Ouest près du métro Riquet. Par contre, il y a des restaurants du Sud-Ouest près du métro Louis Blanc, des restaurants du Sud-Ouest près du Parc des Buttes-Chaumont et des restaurants du Sud-Ouest près du métro Ourcq. Quelle proposition vous intéresse ?

Simulateur :prix de moins de dix euros je cherche un restaurant africain dix-huitième arrondissement

PlanResto :Je ne peux pas effectuer une recherche sur un lieu et un arrondissement ensemble, j'ignore l'arrondissement. Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros près du métro Riquet. Par contre, il y a des restaurants du Sud-Ouest pour environ vingt-trois euros le midi près du métro Ourcq, des restaurants du Sud-Ouest pour environ soixante-et-un euros le midi près du métro Louis Blanc et des restaurants du Sud-Ouest pour environ quarante-six euros le midi près du Parc des Buttes-Chaumont. Quelle proposition vous intéresse ?

Simulateur :je cherche Un restaurant africain Un restaurant pour moins de dix euros Ourcq

PlanResto :Vous recherchez un restaurant du Sud-Ouest pour moins de dix euros. Un restaurant près du métro Ourcq ou un restaurant près du métro Riquet ?

Simulateur :Un restaurant africain

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

Suite à la page suivante

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

PlanResto :Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse ?

Simulateur :Un restaurant africain

Suite à la page suivante

PlanResto	:Je ne vous ai pas compris. Pouvez-vous répéter? Il y a des restaurants du Sud-Ouest pour environ quinze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ onze euros le midi. Quelle proposition vous intéresse?
Simulateur	:Un restaurant africain
PlanResto	:Il n'y a pas de restaurant du Sud-Ouest pour moins de dix euros. Par contre, il y a des restaurants du Sud-Ouest pour environ onze euros le midi, des restaurants du Sud-Ouest pour environ quatorze euros le midi et des restaurants du Sud-Ouest pour environ quinze euros le midi. Quelle proposition vous intéresse?
Simulateur	:au revoir
PlanResto	:Voulez-vous réellement quitter PlanResto?
Simulateur	:oui

TAB. 7.5: Dialogue où l'application PLANRESTO et le simulateur bouclent de façon infinie

En conclusion, l'évaluation de l'application PLANRESTO présentée ci-dessus démontre que le paradigme SIMDIAL est pertinent pour évaluer des SDHM. En effet les résultats obtenus sont cohérents avec les capacités de « dialogue naturel » de l'application PLANRESTO même si l'inter-compréhension nécessaire à la résolution de la tâche du simulateur nécessite parfois plus de tours de parole que ne le ferait un utilisateur humain.

## 7.2 Évaluation de l'application DialogueBourse

Cette expérience a pour objectif d'une part de démontrer qu'il est possible d'évaluer d'autres application que PLANRESTO avec le paradigme SIMDIAL, et d'autre part de quantifier la difficulté de transformation du simulateur de l'application PLANRESTO vers DIALOGUEBOURSE. Les résultats obtenus sur ce derniers point sont présentés en section 6.4.

### 7.2.1 Observations

Pour l'évaluation de l'application DIALOGUEBOURSE le comportement du simulateur variait en fonction de trois des paramètres utilisés pour l'évaluation de l'application PLANRESTO. En effet comme précisé dans le premier paragraphe de la section 6.2.2, l'ajout d'ambiguïtés n'a pas été testé avec l'application DIALOGUEBOURSE. Les variations portaient donc sur :

- **Le nombre d'informations relatives à la tâche par énoncé.** Ce paramètre est, comme explicité précédemment, considéré comme une information sur l'expertise des utilisateurs.
- **La stratégie directive ou non de l'utilisateur simulé.** De même que pour l'application PLANRESTO, les différences générées par ce paramètre repose sur les capacités de gestion de l'initiative des systèmes évalués.
- **L'ajout d'une hésitation par énoncé produit par le simulateur.** La robustesse de l'application DIALOGUEBOURSE face aux hésitations peut ainsi être contrôlée.

Concernant les critères d'observation des dialogues, les mêmes que ceux utilisés pour l'application PLANRESTO ont été employés, augmentés pour chacun des dialogues du nombre d'informations que le simulateur cherchait à obtenir. Ce dernier critère a été ajouté car il impacte considérablement le nombre de tours de parole d'un dialogue et que le modèle de la tâche pour l'application DIALOGUEBOURSE diffère notamment de celui utilisé pour PLANRESTO sur ce point. En effet lors d'interaction avec l'application PLANRESTO seules une ou deux informations pouvaient être obtenues alors qu'avec DIALOGUEBOURSE leur nombre peut être de six ou sept. Ainsi l'application PLANRESTO fournissait au simulateur les informations attendues en un voire deux tours de paroles, alors qu'avec l'application DIALOGUEBOURSE ce nombre de tours de parole peut être de 6 ou 7. Les résultats en terme d'efficacité de résolution de la tâche seraient alors considérablement différent entre PLANRESTO et DIALOGUEBOURSE si le nombre d'informations demandées n'était pas pris en compte.

De plus l'application DIALOGUEBOURSE ne relâchant pas de contraintes sur les critères de recherches, aucune observation de ce phénomène n'a été réalisée. Enfin aucune désambiguïstation sur les critères de recherche n'a été observée non plus à cause du petit nombre de critères dans le modèle de la tâche.



Les résultats obtenus portent donc sur :

- L’aboutissement ou non des dialogues ;
- Le nombre de tours de parole nécessaires à réaliser la tâche ;
- Le nombre d’information que le simulateur cherchait à obtenir.

Ainsi 1200 dialogues ont été générés suivant les paramètres présentés ci-dessus. Il est à noter que les mêmes conventions sont utilisées dans les tableaux 7.6 et 7.7 que celles précisées en 7.1.1 et que le nombre d’informations que le simulateur cherchait à obtenir est noté « Nb Info ».

	H	Total	Aboutis	Nb Infos
		Dial – TdP	Dial – TdP	
Directif	–	300 – 1987	278 – 1522	1048
	✓	300 – 1946	279 – 1523	1040
Non Dir.	–	300 – 2294	275 – 1729	1079
	✓	300 – 2134	284 – 1772	1105
<b>Total</b>		1200 – 8361	1116 – 6546	4272

TAB. 7.6 – Dialogues générés et tours de paroles associés pour la tâche DIALOGUEBOURSE

	H	TdP par Dial	Nb Infos par Dial	TdP par Nb Infos
Directif	–	5,47	3,77	1,45
	✓	5,46	3,73	1,46
Non Dir.	–	6,29	3,92	1,60
	✓	6,24	3,89	1,60
<b>Total</b>		5,87	3,83	1,53

TAB. 7.7 – Ratios pour les dialogues aboutis de la tâche DIALOGUEBOURSE

## 7.2.2 Analyses

### Sur l’aboutissement des dialogues

Sur 1200 dialogues auxquels ont pris part le simulateur et l’application DIALOGUEBOURSE, 1116 ont abouti en terme de réalisation de la tâche,

soit 93%. Ce résultat est comparable à celui obtenu lors de l'évaluation de l'application PLANRESTO, pour laquelle 88% des dialogues avaient aboutis. De même, aucun faux-positif ni faux-négatif n'a été relevé dans l'analyse manuelle de dialogues, ce qui tend à démontrer la fiabilité du diagnostic d'aboutissement.

Concernant les 84 dialogues non aboutis, une analyse manuelle a permis de quantifier à 35 les dialogues contenant des erreurs d'interprétation des énoncés de l'application DIALOGUEBOURSE par le simulateur, soit environ 40% des erreurs observées.

Les 49 autres dialogues non aboutis sont dus à un dysfonctionnement de l'application DIALOGUEBOURSE. En effet cette dernière interrompt parfois le dialogue en cours en annonçant qu'elle le reprend depuis le début. Le simulateur reprend alors lui aussi le dialogue depuis le début, ce qui mène régulièrement au dépassement du seuil de quinze tours de parole pour résoudre la tâche (voir l'exemple dialogue 7.10).

Comme pour l'application PLANRESTO, les résultats en terme de résolution de la tâche démontrent une forte cohérence entre les modèles de l'interaction du simulateur et de l'application DIALOGUEBOURSE, les quelques dialogues non aboutis n'étant pas même dus à des divergences dialogiques mais à des problèmes linguistiques ou techniques.

### **Sur l'efficacité en fonction de la stratégie**

Les résultats obtenus en fonction de la stratégie de l'utilisateur sont très intéressants puisque contre-intuitifs. En effet l'utilisateur simulé directif a ici été plus efficace que l'utilisateur simulé non directif. En effet le nombre de tours de parole moyen de l'utilisateur simulé directif pour obtenir une information est de 1,46 alors que l'utilisateur simulé non directif doit en moyenne réaliser 1,60 tours de parole.

La différence entre les résultats obtenus par l'application PLANRESTO et l'application DIALOGUEBOURSE est interprétable en termes de capacités dialogiques et de modèles de la tâche. D'une part l'application DIALOGUEBOURSE ne relâche pas de contrainte sur les critères de recherche et donc la stratégie non directive ne gagne pas en efficacité sur ce point. D'autre part le modèle de la tâche de recherche d'information boursière est hiérarchisé sur un niveau de plus que celui de la recherche de restaurant, ce qui rend la stratégie non directive moins efficace. En effet l'application DIALOGUEBOURSE propose d'abord à l'utilisateur de choisir s'il veut des informations sur ses comptes, sur des actions ou sur des indices. L'utilisateur simulé non

directif répond alors au système tandis que l'utilisateur simulé directif fournit directement le nom du compte, de l'action ou de l'indice qui l'intéresse.

### Sur le phénomène perturbateur hésitation

Comme pour l'application PLANRESTO, les résultats obtenus sur l'hésitation démontrent que l'application DIALOGUEBOURSE est robuste face à ce phénomène. La différence du nombre de tours de parole pour les dialogues auxquels ont été ajoutés des hésitations par rapport à ceux auxquels il n'en a pas été ajouté est encore de moins de 2%, soit 5,88 tours de parole pour les dialogues ne contenant pas d'hésitations et 5,85 pour ceux qui en contenaient.

### 7.2.3 Exemples de dialogues

De même que pour l'application PLANRESTO sont présentés trois dialogues ayant eu lieu entre le simulateur d'utilisateurs et l'application DIALOGUEBOURSE. Les deux premiers (tab. 7.8 et 7.9) sont comparables à des dialogues entre un utilisateur réel et le système. Le troisième (tab. 7.10) est un dialogue non abouti pour la raison précédemment expliquée, à savoir des reprises à zéro du dialogue par le système, ce qui mène à un dialogue dépassant les quinze tours de paroles.

DialogueBourse :Bonjour Monsieur Allemandou. Bienvenue sur Dialogue Bourse. Vous pouvez consulter votre portefeuille ou votre carnet d'ordres, obtenir une information boursière ou effectuer une transaction sur les actions cotées à la Bourse de Paris. Que désirez-vous ?

Simulateur :Des informations sur l'action Clarins

DialogueBourse :L'action Clarins cotait 57 euros 30 à 17 heures 30. Vous pouvez également demander des informations plus détaillées ou formuler une autre demande. Que désirez-vous ?

Simulateur :Quel est l'avis des analystes

DialogueBourse :Pour l'action Clarins, l'opinion de nos analystes est de conserver depuis le 24 octobre 2000. Que désirez-vous ?

Simulateur :Quel est le volume d'action échangé

DialogueBourse :Le volume échangé depuis le début de la séance pour l'action Clarins s'élève à trente-sept mille cinq cent cinquante-six actions. Que désirez-vous ?

Simulateur :Je veux le dividende

Suite à la page suivante

DialogueBourse	:Le dividende prévu de l'action Clarins était 1 euro 52 pour l'exercice 2000. Le dividende prévu était 1 euro 31 pour l'exercice 1999. Voulez-vous une autre information ?
Simulateur	:Quel est sa variation
DialogueBourse	:Le cours de l'action Clarins a baissé de 0,17 pour cent par rapport à la dernière séance. Je ne connais pas son évolution depuis le 1er janvier. Son cours le plus bas depuis le 1er janvier 2006 est 44 euros 26. Son cours le plus haut depuis le 1er janvier 2006 est 58 euros 50. Que désirez-vous ?
Simulateur	:Je veux quitter
DialogueBourse	:J'ai compris que vous ne désirez plus d'information. Pouvez-vous le confirmer ?
Simulateur	:Oui

TAB. 7.8: Dialogue entre DIALOGUEBOURSE et le simulateur ressemblant à un dialogue humain-machine

DialogueBourse	:Bonjour Monsieur Allemandou. Bienvenue sur Dialogue Bourse. Vous pouvez consulter votre portefeuille ou votre carnet d'ordres, obtenir une information boursière ou effectuer une transaction sur les actions cotées à la Bourse de Paris. Que désirez-vous ?
Simulateur	:Mon compte titre
DialogueBourse	:La valorisation globale de votre compte-titres ordinaire était 60027 euros 14 le 17 mai 2006. Vous pouvez demander le détail des lignes, le disponible espèce ou RD ou formuler une autre demande. Que désirez-vous ?
Simulateur	:Quel est le solde S.R.D.
DialogueBourse	:Votre solde disponible SRD pour votre compte-titres ordinaire est 10000 euros. Votre plafond d'autorisation d'encours SRD pour votre compte-titres ordinaire est 10000 euros. Que désirez-vous ?
Simulateur	:Quel est le solde
DialogueBourse	:La position espèce de votre compte-titres ordinaire est 10000 euros. Que désirez-vous ?
Simulateur	:Quitter
Suite à la page suivante	

DialogueBourse	:J'ai compris que vous ne désirez plus d'information. Pouvez-vous le confirmer ?
Simulateur	:Oui je confirme

TAB. 7.9: Dialogue entre DIALOGUEBOURSE et le simulateur ressemblant à un dialogue humain-machine

DialogueBourse	:Bonjour Monsieur Allemandou. Bienvenue sur Dialogue Bourse. Vous pouvez consulter votre portefeuille ou votre carnet d'ordres, obtenir une information boursière ou effectuer une transaction sur les actions cotées à la Bourse de Paris. Que désirez-vous ?
Simulateur	:L'action Monsieur-Bricolage euh
DialogueBourse	:L'action Monsieur-Bricolage cotait 17 euros 10 à 17 heures 18. Vous pouvez également demander des informations plus détaillées ou formuler une autre demande. Que désirez-vous ?
Simulateur	:Euh quel est votre opinion
DialogueBourse	:Je ne connais pas l'opinion de nos analystes pour l'action Monsieur-Bricolage. Que désirez-vous ?
Simulateur	:Quel est euh le volume d'action échangé
DialogueBourse	:Le volume échangé depuis le début de la séance pour l'action Monsieur-Bricolage s'élève à mille cinq cent trois actions. Que désirez-vous ?
Simulateur	:Le dividende euh
DialogueBourse	:Je ne connais pas le dividende de l'action Monsieur-Bricolage. Que désirez-vous ?
Simulateur	:Euh le rendement
DialogueBourse	:Désolé, j'ai un problème pour répondre à cette question. Je repars à zéro. Veuillez reformuler votre requête. Voulez-vous une autre information ?
Simulateur	:Pour euh Monsieur-Bricolage
DialogueBourse	:L'action Monsieur-Bricolage cotait 17 euros 10 à 17 heures 18. Vous pouvez également demander des informations plus détaillées ou formuler une autre demande. Que désirez-vous ?
Simulateur	:Euh quel est l'avis des analystes

Suite à la page suivante

DialogueBourse :Je ne connais pas l'opinion de nos analystes pour l'action Monsieur-Bricolage. Que désirez-vous ?

Simulateur :Quel est le volume échangé euh

DialogueBourse :Le volume échangé depuis le début de la séance pour l'action Monsieur-Bricolage s'élève à mille cinq cent trois actions. Que désirez-vous ?

Simulateur :Le euh dividende

DialogueBourse :Désolé, j'ai un problème pour répondre à cette question. Je repars à zéro. Veuillez reformuler votre requête. Voulez-vous une autre information ?

Simulateur :euh Monsieur-Bricolage

DialogueBourse :L'action Monsieur-Bricolage cotait 17 euros 10 à 17 heures 18. Vous pouvez également demander des informations plus détaillées ou formuler une autre demande. Que désirez-vous ?

Simulateur :Euh Quel est votre opinion

DialogueBourse :Je ne connais pas l'opinion de nos analystes pour l'action Monsieur-Bricolage. Que désirez-vous ?

Simulateur :Le volume échangé je voudrais euh

DialogueBourse :Le volume échangé depuis le début de la séance pour l'action Monsieur-Bricolage s'élève à mille cinq cent trois actions. Que désirez-vous ?

Simulateur :euh Quel est le rendement

DialogueBourse :Je ne connais pas le rendement brut, avoir fiscal inclus, de l'action Monsieur-Bricolage. Que désirez-vous ?

Simulateur :et euh la variation

DialogueBourse :Le cours de l'action Monsieur-Bricolage a augmenté de 0,06 pour cent par rapport à la dernière séance. Je ne connais pas son évolution depuis le 1er janvier. Par contre, le cours de l'action Monsieur-Bricolage a augmenté de 0,06 pour cent par rapport à la dernière séance. Son cours le plus bas depuis le 1er janvier 2006 est 14 euros 01. Son cours le plus haut depuis le 1er janvier 2006 est 18 euros 17. Que désirez-vous ?

Simulateur :euh ce sera tout merci

DialogueBourse :J'ai compris que vous ne désirez plus d'information. Pouvez-vous le confirmer ?

Simulateur :euh oui

Suite à la page suivante

---

TAB. 7.10: Dialogue non abouti pour cause de reprise à zéro de DIALOGUEBOURSE

### 7.3 Conclusion

L'application du paradigme SIMDIAL pour évaluer les applications PLAN-RESTO et DIALOGUEBOURSE a démontré d'une part que des évaluations technologiques quantitatives sont réalisables par simulation déterministe d'utilisateurs, et d'autre part que les phénomènes et stratégies modélisés sont pertinents à étudier face aux systèmes de dialogues.

Les stratégies directives et non directives de l'utilisateur simulé fournissent en effet des résultats différents en terme d'efficacité en fonction des stratégies et capacités des systèmes évalués. Leur analyse peut permettre d'affiner les stratégies des systèmes de dialogue en fonction des comportements des utilisateurs.

Les phénomènes perturbateurs étudiés dont notamment l'ajout d'ambiguïté sur les critères de recherche permettent d'analyser les stratégies mises en œuvre par les systèmes en fonction des informations sur lesquelles portent de tels phénomènes ainsi que du contexte dans lequel ils apparaissent.

Enfin le contrôle de cohérence entre les modèles de l'interaction du simulateur et ceux des systèmes évalués est un bon moyen de test technologique pour des systèmes complexes comme les SDHM.

# Chapitre 8

## Conclusion et Perspectives

### 8.1 Conclusion

L'évaluation des systèmes de dialogue est un défi actuel de la communauté de recherche. En effet les récentes avancées dans le domaine du dialogue homme-machine ont démontré la nécessité d'évaluer les interfaces dialoguantes pour mieux les faire évoluer.

La première partie de cette thèse, principalement introductive, présentait le dialogue homme-machine, la problématique de son évaluation, ainsi qu'un état de l'art de cette activité. Ensuite le paradigme SIMDIAL a été détaillé, au travers des choix qui ont mené à sa réalisation, de la conception du simulateur et enfin des deux expériences réalisées.

Le paradigme SIMDIAL répond d'une part à des besoins industriels en ce qu'il permet d'automatiser l'évaluation de SDHM sur des tâches et phénomènes précis, et d'autre part il s'intègre dans les recherches menées sur les SDHM par les conceptions tant méthodologiques qu'appliquées qui ont permis sa réalisation.

Enfin il est à noter que ce paradigme a pour objectif d'être à la fois très générique et peu coûteux en ressources. Chacun de ses différents modules pourrait certainement être amélioré, notamment de façon dépendante des objectifs et fonctionnalités nécessaires à tester pour améliorer un système mais il ne s'agit évidemment pas de créer un simulateur aussi complexe qu'un système de dialogue.

#### 8.1.1 Méthodes et modèles employés

Sur les méthodes et modèles employés pour la réalisation du paradigme SIMDIAL le premier point à noter est que l'approche utilisée pour modéliser



les comportements dialogiques du simulateur est déterministe, autrement dit définie par expertise. Comme avec une approche stochastique, l'approche choisie permet d'évaluer les systèmes dans les cas d'utilisation standards, pour peu qu'ils aient été définis.

Ce qu'elle permet de plus est l'évaluation face à des comportements marginaux peu représentés dans les corpus, notamment grâce à la notion de phénomène perturbateur. Enfin elle est aisément transposable à d'autres tâches de part sa conception basée sur des modèles génériques (voir 6.4).

Concernant la verbalisation des énoncés du simulateur, l'approche développée est hybride : par expertise et avec corpus. L'avantage majeur de cette approche est de fournir de bons résultats de verbalisations sans nécessiter de gros corpus. Par exemple le premier corpus de test ne contenait que 28 dialogues et les énoncés générés étaient déjà de bonne qualité. De plus la distinction entre verbalisation et choix du contenu sémantique de la réponse autorise la modification des contenus indépendamment de la verbalisation, ce qui permet notamment l'évaluation semi-automatique.

Par ailleurs le test perceptif effectué sur l'aspect, le sens et la syntaxe des énoncés générés automatiquement par la stratégie, ont démontré que les différences ne sont presque pas perceptibles entre ces énoncés et des transcriptions automatiques d'énoncés humains.

Enfin la modélisation des phénomènes perturbateurs de façon décorrélée de l'algorithme de génération de la réponse semble un bon compromis entre, d'une part, la possibilité de générer « tous » les phénomènes perturbateurs, et d'autre part, la difficulté d'adaptation des comportements du simulateur à ces phénomènes. Il est clair que les modèles utilisés pour le simulateur ne permettent pas de représenter correctement tous les phénomènes perturbateurs. Par exemple la définition d'un critère de recherche par négation, comme dans l'énoncé « *Je veux un restaurant mais pas un japonais* », n'est pas réalisable sans modifier le modèle de la tâche. En revanche l'ajout des phénomènes perturbateurs de l'oral spontané définis par le GDR-I3 [Antoine et al., 2002a] est aisée, ainsi que des phénomènes d'ordre dialogique comme par exemple la génération d'ambiguïtés, la contestation d'un énoncé ou la demande de reformulation.

### 8.1.2 Propriétés du paradigme SIMDIAL

Concernant les propriétés du paradigme SIMDIAL, la méthode de la simulation d'utilisateurs pour obtenir des corpus de dialogues ouvre des perspectives quant aux méthodes de développement des SDHM. En effet

l'approche proposée permet théoriquement de comparer différents systèmes sur une même tâche en les mettant dans les mêmes conditions contrôlées : scénarios, phénomènes perturbateur, critères d'évaluation. L'évaluateur humain a alors toute latitude de mettre en œuvre sa propre métrique d'évaluation comparative.

Par ailleurs il devient possible d'accéder à de gros ensembles de données générées à partir du SDHM en développement. Évidemment les dialogues obtenus ne sont pas de « vrais » dialogues humain-machine et les corpus générés sont principalement utilisables pour du test. Un apprentissage sur ces données amènerait le biais de n'apprendre que les stratégies spécifiées de façon déterministe. En revanche ces corpus permettent d'obtenir des données à observer et analyser sans avoir à réaliser de tests avec des utilisateurs réels. En d'autres termes cette méthode permet d'automatiser certaines étapes de tests du cycle de développement d'un SDHM. Il est clair que des résultats relatifs aux utilisateurs comme la satisfaction ou l'utilisabilité ne peuvent être obtenus directement avec cette méthode. Ainsi les paradigmes d'évaluation confrontant les systèmes de dialogue aux utilisateurs afin d'obtenir des informations subjectives sont complémentaires à l'approche proposée.

Toujours dans l'idée de réduire le coût de la phase d'évaluation d'un SDHM, le diagnostic automatique de phénomènes permet de restreindre significativement les analyses manuelles. Les tests réalisés avec l'application PLANRESTO et DIALOGUEBOURSE ont permis de diagnostiquer le non-aboutissement et ainsi de n'analyser manuellement que 371 dialogues sur 3600, soit environ 10%. Ont aussi été diagnostiqués les dialogues et tours de parole contenant des relaxations de critères et des désambiguïisations. Ces diagnostics permettent d'obtenir des informations relatives à ces phénomènes, comme le fait que l'application PLANRESTO n'utilise pas toujours la même stratégie de désambiguïisation, ou le fait que la différence d'efficacité observée entre un utilisateur simulé directif et non-directif est très corrélée avec la présence de relaxation dans les dialogues.

Une restriction à apporter sur le diagnostic reste la difficulté à définir précisément ce qui doit être diagnostiqué. Le problème des faux positifs en est un exemple. Bien que les cas traités pour les applications évaluées valident la définition des diagnostics réalisés, le problème n'en reste pas moins présent pour la définition d'autres diagnostics automatiques. Le problème de la définition précise de ce que l'on souhaite diagnostiquer existe aussi pour le diagnostic manuel, avec la souplesse que la compréhension humaine apporte face à la rigidité des systèmes automatiques.

Une quatrième possibilité offerte par le paradigme *SIMDIAL* est l'observation de corrélations entre mesures de performance sur les dialogues, stratégies d'utilisateurs simulés et stratégies de systèmes. Comme observé dans [Schatzmann et al., 2005], cette propriété est générale à toutes les méthodes d'évaluation par simulation d'utilisateurs déterministe.

Les limites de cette propriété sont celles imposées par les modèles utilisés, tant au niveau des stratégies d'utilisateurs simulés que des systèmes évalués. En d'autres termes il n'est possible d'obtenir d'informations que sur les stratégies modélisées.

La dernière propriété à noter est que le paradigme *SIMDIAL* est modifiable à faible coût pour être appliqué à d'autres tâches. En effet le simulateur est fondé sur l'analyse de capacités généralement admises pour les SDHM et non précisément celles de SDHM en particulier. Par exemple le transfert du simulateur de l'application *PLANRESTO* vers *DIALOGUEBOURSE* a été effectué en 11 jours par une personne, soit une journée d'annotation pour vingt dialogues, trois pour modifier les modules d'interprétation et de verbalisation, trois pour implémenter la tâche et trois pour affiner le simulateur dans son ensemble.

En somme le paradigme *SIMDIAL* permet une évaluation automatique des comportements dialogiques de systèmes de dialogue. Les critères d'évaluation exploités sont la résolution de la tâche et la durée des dialogues en tours de parole, associés à différentes stratégies et phénomènes perturbateurs à des fins de diagnostic. Par ailleurs le paradigme utilise le langage naturel pour interagir avec les systèmes évalués et n'exploite pas de traces de ces derniers pour réaliser ses mesures, ce qui le rend théoriquement interfaçable avec différents SDHM sur une tâche précise et permet ainsi des évaluations comparatives. Enfin il est aisément transposable à différentes tâches.

## 8.2 Perspectives

Un premier point pourrait-être de définir plus finement les comportements générés par le simulateur, ce qui demande de complexifier l'algorithme de génération de contenu de la réponse ainsi que les modèles associés. Outre le coût associé il faudrait alors choisir sur quelles notions affiner les comportements, comme par exemple l'implémentation d'autres stratégies utilisateurs ou la modélisation de phénomènes dialogiques complexes comme la négation

ou les critères disjoints.

Par ailleurs il serait aussi très intéressant de continuer à modéliser les phénomènes perturbateurs de façon indépendante de l'interaction. Cela permettrait d'une part de les générer aisément, et d'autre part de faire que les systèmes de dialogue se comportent spécifiquement face à ces phénomènes. La modélisation de tels phénomènes de façon indépendante de l'interaction doit notamment être menée en collaboration avec différentes communautés de recherche telles que la linguistique, l'ergonomie ou la psychologie expérimentale afin d'élargir la notion de phénomènes perturbateurs.

De plus il serait intéressant de chercher le bon niveau de granularité des modèles, afin d'obtenir un niveau d'évaluation suffisamment fin tout en préservant leurs capacités de réutilisation. En effet la généralité par rapport à la tâche et aux modes d'interaction est dépendante des possibilités de généralisation des modèles aux nouveaux cas à traiter.

Sur ce point il reste à noter qu'il est difficile de trouver un « bon » compromis entre généralité et précision des modèles. En effet le besoin de généralité pour les systèmes de dialogue s'oppose généralement à la tentation d'être toujours plus précis dans la modélisation des tâches et du dialogue. La définition de structures de représentation permettant à la fois la précision des descriptions et une généralité importante est un défi de la communauté de recherche.

Une autre évolution possible porte sur le module d'interprétation sémantique du simulateur. En effet ce composant est aujourd'hui doté de capacités minimales afin de permettre les évaluations souhaitées, à savoir la méthode d'îlots sémantiques par égalité textuelle sur des portions d'énoncés.

La prise en compte de variantes flexionnelles voire de scores de confiance associés à l'interprétation des énoncés pourraient être envisagée, ou encore l'adaptation d'un module plus performant pré-existant au simulateur. Ce dernier point reste dépendant des différences potentielles de formalismes de représentation.

De plus le module de verbalisation des énoncés pourrait aussi être exploité à des fins d'évaluation automatique de composants d'interprétation sémantique. En effet le peu de différence existant entre les énoncés générés automatiquement et ceux transcrits par des systèmes de reconnaissance vocale depuis des énoncés humains rend possible de telles évaluations.

Enfin les SDHM et plus particulièrement leurs évaluations sont fondés sur la définition d'un « bon » comportement face aux utilisateurs. Or la définition

de comportements de référence pour une tâche comme le dialogue naturel est éminemment complexe. En effet le « bon » comportement dialogique n'est d'une part pas nécessairement unique et d'autre part est fortement lié au contexte de déroulement du dialogue. Le « bon » comportement face à un utilisateur lambda ne sera pas le même que face à un utilisateur epsilon, de même que le « bon » comportement face à un utilisateur utilisant son téléphone fixe ne sera pas le même que celui face à un utilisateur en situation de mobilité.

Une perspective intéressante serait donc de s'appuyer sur la notion d'instance de modèle utilisateur plutôt que sur un modèle unique de tous les utilisateurs. Dans cette optique, la distinction en *utilisateur* et *usager* proposée par Stéphane Chaudiron semble une bonne dichotomie entre les côtés sociaux et donc généraux des usages et les capacités individuelles des utilisateurs.

Les systèmes pourraient alors être guidés à la fois par un modèle général de comportement associé aux usages en vigueur pour la tâche considérée et par un modèle dynamique des capacités personnelles de l'utilisateur. Cette idée d'adaptabilité des systèmes face aux utilisateurs est l'objet de recherches récentes. L'idée de décorrélérer le modèle d'un usager générique de celui d'utilisateurs individuels semble être un niveau de granularité intéressant entre généralité et précision des modèles.

En conclusion, le paradigme SIMDIAL est très attractif pour tester et comparer les systèmes de dialogue à moindre coût. Son principe a été énoncé dans cette thèse et il a été appliqué à l'évaluation de différents paramètres sur deux tâches. Il serait aisé d'inclure dans le paradigme d'autres mesures sur des phénomènes dialogiques ou linguistiques tels que les reformulations, les répétitions ou encore les auto-corrrections.

L'approche choisie explore ainsi plus avant l'évaluation de système de dialogue par simulation déterministe d'utilisateurs, voie précédemment ouverte par les études menées par Lopez et al. [López-Cózar et al., 2003].

# Troisième partie

## Annexes



# Annexe A

## Tâche PlanResto

Spécialités	246	3,90%
Lieux	481	7,63%
Prix	102	1,62%
Restaurants	5473	86,85%

TAB. A.1 – Répartition des valeurs du thésaurus relatif à la tâche PLANRESTO

Inform	672	84,00%
Request	77	9,63%
Accept	25	3,12%
Reject	23	2,88%
Open	3	0,37%
Close	0	0,00%

TAB. A.2 – Répartition des annotations dialogiques de verbalisation pour le domaine de la tâche PLANRESTO

Comme précisé dans la section 6.1.3, les actes de dialogues sont ici représentatifs de la forme de surface des énoncés et pas de la pragmatique du discours, ce qui explique le taux d'« Inform » dans le corpus. Dans le cas où les actes auraient représenté la pragmatique du discours, les taux d'« Inform » et de « Request » auraient été inversés.



restaurant	2	0,20%
specialite	130	13,58%
lieu	180	18,80%
prix <sup>1</sup>	149	15,60%
info	86	8,99%
meta	129	13,48%
liste <sup>2</sup>	78	8,15%
bruit	94	9,82%
null	106	11,08%

TAB. A.3 – Répartition des annotations sémantiques de verbalisation pour le domaine de la tâche PLANRESTO

---

<sup>1</sup>Sont regroupés sous cette notation les trois attributs prixInf, prixEgal et prixSup.

<sup>2</sup>Sont regroupés sous cette notation les trois attributs listeNav, listeNavNum et listeProp.

## Annexe B

### Tâche DialogueBourse

Compte	2	0,18%
Indice	10	0,88%
Action	1080	95,49%
InfoCompte	13	1,15%
InfoIndice	7	0,62%
InfoAction	19	1,68%

TAB. B.1 – Répartition des valeurs du thésaurus relatif à la tâche DIALOGUEBOURSE

Inform	240	76,92%
Request	50	16,03%
Accept	10	3,21%
Reject	11	3,52%
Open	1	0,32%
Close	0	0,00%

TAB. B.2 – Répartition des annotations dialogiques de verbalisation pour le domaine de la tâche DIALOGUEBOURSE

---

<sup>1</sup>Sont regroupés sous cette notation les trois attributs listeNav, listeNavNum et listeProp.

Compte	48	10,74%
Indice	9	2,02%
Action	106	23,71%
InfoCompte	51	11,41%
InfoIndice	14	3,13%
InfoAction	113	25,28%
meta	28	6,26%
liste <sup>1</sup>	3	0,67%
bruit	53	11,86%
null	22	4,92%

TAB. B.3 – Répartition des annotations sémantiques de verbalisation pour le domaine de la tâche DIALOGUEBOURSE

## Annexe C

### Détails sur le test perceptif de la verbalisation automatique

		Transcriptions d'énoncés humains		Énoncés produits automatiquement	
Aspect naturel	Oui	421	50,66%	410	49,34%
	Moyen	138	49,29%	142	50,71%
	Non	66	47,48%	73	52,52%
Syntaxe correcte	Oui	306	48,49%	325	51,51%
	Non	306	52,04%	282	47,96%
	Sans avis	13	41,94%	18	58,06%
Sens (moyenne)		78,17 / 100		80,10 / 100	

TAB. C.1 – Résultats du test perceptif

un restaurant pour vingt-cinq euros
je désire un hum restaurant africain entre pour le prix de cinq euros dans le vingtième arrondissement s'il vous plait
je recherche un restaurant antillais
plus de cinq euros dans le vingtième arrondissement s'il vous plait
je veux consulter les solutions
un restaurant traditionnel
un restaurant près du Louvre euh sud-ouest
un restaurant pour plus de cinquante euros
un restaurant gastronomique un prix de moins de cinquante euros près du treizième arrondissement
moins de cinq euros restaurant traditionnel louvre
près du treizième arrondissement pour plus de cinq euros
prix de moins de dix euros de spécialités rapide
dans le vingtième arrondissement s'il vous plait
sixième arrondissement pour moins de vingt euros spécialité hamburger
un restaurant de spécialités savoyardes pour plus de cinq euros
je cherche un restaurant à Paris proche du Louvre.
pour moins de quinze euros spécialité hamburger Nation
dix-septième arrondissement vingt-cinq euros euh sud-ouest
je voudrais un restaurant près de Bercy
je cherche un restaurant dans le cinquième arrondissement trente euros
dix-septième arrondissement je désire manger japonais
dix-septième arrondissement de spécialités rapide prix environ trente euros
restaurant européen plus de cinq euros le neuvième arrondissement
un restaurant traditionnel
dans le vingtième arrondissement s'il vous plait moins de cinq euros
un restaurant pour moins de dix euros

TAB. C.2 – Énoncés produits par la stratégie de verbalisation

je cherche un restaurant euh le deuxième arrondissement
je cherche un restaurant thaïlandais
je cherche un restaurant indien dans le onzième
euh je cherche un restaurant dans le quinzième une brasserie
voudrais un restaurant à cinquante euros
je cherche un restaurant antillais et dans le dixième arrondissement
dans le quinzième arrondissement pour trente deux euros vingt-cinq euros
un restaurant près de la place de la Bastille
euh oui bonjour je recherche une brasserie dans le
quinzième arrondissement pour cent trente-cinq euros
euh j'aurais consulter les solutions
alors je recherche un restaurant de traditionnelle
près de la Tour-Eiffel pour trente euros
un restaurant un euh le dixième arrondissement
un troisième arrondissement de la cuisine étrangère
sixième arrondissement type végétarien pour vingt euros
restaurant près tour Eiffel et traditionnel
je cherche un restaurant près de Bastille à moins de quinze euros
je recherche un restaurant antillais pour moins
de vingt euros près du treizième arrondissement
euh un restaurant dans le cinquième arrondissement
je voudrais un restaurant italien Nation
plus d'un restaurant indien euh rue euh pour plus de cinq euros
je veux consulter les solutions
je cherche un restaurant africain dans le troisième arrondissement
je cherche un restaurant algérien un de quinze euros
voudrais un restaurant dans le deuxième arrondissement
dans le troisième arrondissement pour un montant d'environ quinze euros

TAB. C.3 – Transcriptions automatiques d'énoncés produits par des humains



# Bibliographie

- [Aist et al., 2006] Aist, G., Michalak, P., Ferguson, G., and Allen, J. (2006). Challenges in evaluating spoken dialogue systems that reason and learn. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Allwood, 1977] Allwood, J. (1977). *Logic, Pragmatics and Grammar*, chapter A Critical Look at Speech Act Theory, pages 53–99. Lund. Studentlitteratur.
- [Allwood, 1987] Allwood, J. (1987). A semantic analysis of understanding. In *Proceedings of the 10th Scandinavian Conference of Linguistics*.
- [Allwood, 1995] Allwood, J. (1995). *Dialogue and Instruction : Modeling Interaction in Intelligent Tutoring Systems*, chapter Reasons for Management in Spoken Dialogue, pages 241–250. Springer-Verlag.
- [Antoine, 2001] Antoine, J.-Y. (2001). Méthodologie d'évaluation par défi. Technical report, VALORIA - EQUIPAGE - LN.
- [Antoine et al., 2002a] Antoine, J.-Y., Bousquet, C., Goulian, J., Jamoussi, S., Kurdi, M., Rosset, S., Vigouroux, N., and Villaneau, J. (2002a). Quelques problèmes posés à la compréhension de parole : typologie de phénomènes étudiés dans le cadre des campagnes d'évaluation par défi du gdr-i3 du cnrs. Technical report, GT 5.5 - GDR I3 - CNRS.
- [Antoine et al., 2002b] Antoine, J.-Y., Bousquet, C., Goulian, J., Kurdi, M., Rosset, S., Vigouroux, N., and Villaneau, J. (2002b). Predictive and objective evaluation of speech understanding : the "challenge" evaluation campaign of the i3 speech workgroup of the french cnrs. In *Proceedings of the 3rd Language Resource and Evaluation Conference*.
- [Antoine and Caelen, 1999] Antoine, J.-Y. and Caelen, J. (1999). Pour une évaluation objective, prédictive et générique de la compréhension de chm orale : le paradigme dcr (demande, contrôle, résultat). In *Langues*.
- [Antoine et al., 2000] Antoine, J.-Y., Siroux, J., Caelen, J., Villaneau, J., Goulian, J., and Ahafhaf, M. (2000). Obtaining predictive results with an



- objective evaluation of spoken dialogue systems : experiments with the dcr assessment paradigm. In *Proceedings of the 2nd Language Resource and Evaluation Conference*.
- [Antoine et al., 1997] Antoine, J.-Y., Zeiliger, J., and Caelen, J. (1997). Rqa methodology : towards a qualitative evaluation of speech understanding and spoken dialog systems. In *Proceedings of the Systematic Analysis of Language Transcripts Workshop on Evaluation*.
- [Antoine et al., 1998] Antoine, J.-Y., Zeiliger, J., and Caelen, J. (1998). Dqr test suites for a qualitative evaluation of spoken dialogue systems : from speech understanding to dialogue strategy. In *Proceedings of the 1st Language Resource and Evaluation Conference*.
- [Araki et al., 1997] Araki, M., Watanabe, T., and Doshita, S. (1997). Evaluating dialogue strategies for recovering from misunderstandings. In *Proceedings of the IJCAI 1997 Workshop on Collaboration Cooperation and Conflict in Dialogue Systems*.
- [ATIL-CNRS, 2003] ATIL-CNRS (2003). Ananas - annotation anaphorique pour l'analyse sémantique de corpus. Technical report, Programme Interdisciplinaire "Société de l'information".
- [aust and Schroer, 1998] aust, H. and Schroer, O. (1998). An overview of the philips dialog system. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.
- [Austin, 1962] Austin, J. L. (1962). *How to do Things with Words*. Harvard University Press.
- [Austin, 1969] Austin, J. L. (1969). *Quand dire c'est faire*. Éditions du Seuil. Traduction de [Austin, 1962].
- [Bernsen and Dybkjaer, 2000] Bernsen, N. O. and Dybkjaer, L. (2000). A methodology for evaluating spoken language dialogue systems and their components. In *Proceedings of the 2nd Language Resource and Evaluation Conference*.
- [Bernsen et al., 1999] Bernsen, N. O., Dybkjaer, L., and Heid, U. (1999). Current practice in the development and evaluation of spoken language dialogue systems. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.
- [Blasband et al., 1999] Blasband, M., Bevan, N., King, M., Maegaard, B., des Tombe, L., Krauwer, S., Manzi, S., and Underwood, N. (1999). Expert advisory group on language engineering standards / evaluation working group final report 2.

- [Bobrow et al., 1977] Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. S., and Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial Intelligence*, 8(2) :155–173.
- [Bodic, 2005] Bodic, L. L. (2005). *Approche de l'évaluation des systèmes interactifs multimodaux par simulation comportementale située*. PhD thesis, Université de Bretagne Occidentale.
- [Bonneau-Maynard and Devillers, 2000] Bonneau-Maynard, H. and Devillers, L. (2000). A framework for evaluating contextual understanding. In *Proceedings of the 6<sup>th</sup> International Conference on Spoken Language Processing*, volume 3, pages 638–641.
- [Bonneau-Maynard and Rosset, 2003] Bonneau-Maynard, H. and Rosset, S. (2003). A semantic representation for spoken dialogs. In *Proceedings of the 8<sup>th</sup> European Conference on Speech Communication and Technology*.
- [Bonneau-Maynard et al., 2005] Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., Mostefa, D., and the MEDIA consortium (2005). Semantic annotation of the French MEDIA dialog corpus. In *Proceedings of the 9<sup>th</sup> European Conference on Speech Communication and Technology*, pages 3457–3460.
- [Caelen et al., 1997] Caelen, J., Zeiliger, J., Bessac, M., Siroux, J., and Pérennou, G. (1997). Les corpus pour l'évaluation du dialogue homme-machine. In *Actes des Journées Scientifiques et Techniques*.
- [Camus, 2001] Camus, S. (2001). Expertise applicative de la technologie de dialogue artificiel. Rapport de stage.
- [Carberry, 1990] Carberry, S. (1990). *ACL-MIT Press Series in Natural Language Processing*, chapter Plan Recognition in Natural Language Dialogue. Bradford Book.
- [Charnay, 1999] Charnay, L. (1999). *Dialogue et Explication dans les Systèmes à Base de Connaissances, ADex, un modèle informatique pour l'énonciation*. PhD thesis, Paris XI. notes et documents LIMSI n°99-19.
- [Chaudiron, 2004] Chaudiron, S. (2004). *L'évaluation des systèmes de traitement de l'information*. Hermès.
- [Chaudiron, 2001] Chaudiron, S. (HDR, Paris X, 2001). évaluation des systèmes de traitement de l'information textuelle.
- [Chaudiron and Ihadjadene, 2002] Chaudiron, S. and Ihadjadene, M. (2002). Quelle place pour l'utilisateur dans l'évaluation des sri ? In Editions, A., editor, *Recherches récentes en sciences de l'information : convergences et dynamiques, actes du colloque MICS-LERASS*, pages 211–231.

- [Chung, 2004] Chung, G. (2004). Developing a flexible spoken dialog system using simulation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20 :37–46.
- [Cronbach, 1951] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 :297–334.
- [Cuayáhuitl, 2005] Cuayáhuitl, H. (2005). *Spoken Dialogue Management using Hierarchical Reinforcement Learning and Dialogue Simulation*. PhD thesis, Institute for Communicating and Collaborative Systems - University of Edinburgh.
- [des Tombe et al., 1996] des Tombe, L., Douglas, S., King, M., Krauwer, S., Maegaard, B., Manzi, S., Paggio, P., Rohde, G., and Underwood, M. T. N. (1996). a testbed study of evaluation methodologies : Authoring aids - final report.
- [Devillers, 2004] Devillers, L. (2004). évaluation des systèmes de dialogue oral. In Chaudiron, S., editor, *Évaluation des systèmes de traitement de l'information*. Hermès Sciences.
- [Devillers, 2006] Devillers, L. (2006). Les émotions dans les interactions homme-machine : perception, détection et génération. HDR, Paris XI.
- [Devillers and Bonneau-Maynard, 1998] Devillers, L. and Bonneau-Maynard, H. (1998). Evaluation of dialog strategies for a tourist information retrieval system. In *Proceedings of the 5th International Conference on Spoken Language Processing*.
- [Devillers et al., 2002] Devillers, L., Bonneau-Maynard, H., and Paroubek, P. (2002). Méthodologie d'évaluation des systèmes de dialogue parlé : réflexions et expériences autour de la compréhension. *Traitement Automatique des Langues*, 43(2) :155–184.
- [Devillers et al., 2003] Devillers, L., Maynard, H., Paroubek, P., and Rosset, S. (2003). The peace slds understanding evaluation paradigm of the french media campaign. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics Workshop on Evaluation Initiatives in NLP*.
- [Devillers et al., 2004] Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet, C., Vigouroux, N., Béchet, F., Romary, L., Antoine, J.-Y., Villaneau, J., Vergnes, M., and Goulian, J. (2004). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *Proceedings of the 4th Language Resource and Evaluation Conference*.

- [Devillers et al., 2005] Devillers, L., Vidrascu, L., and Lamel, L. (2005). Emotion detection in real-life spoken dialogs recorded in call center. *Journal of Neural Networks, numéro spécial "Emotion and Brain"*, 18(4) :407–422.
- [DISC Consortium, 1999] DISC Consortium (1999). Disc - final report.
- [Dybkjaer et al., 1998] Dybkjaer, L., Bernsen, N. O., Carlson, R., Chase, L., Dahlbäck, N., Failenschmid, K., Heid, U., Heisterkamp, P., Jönsson, A., Kamp, H., Karlsson, I., Kuppevelt, J., Lamel, L., Paroubek, P., and Williams, D. (1998). The disc approach to spoken language systems development and evaluation. In *Proceedings of the 1st Language Ressource and Evaluation Conference*.
- [Dybkjær et al., 2004] Dybkjær, L., Bernsen, N. O., and Minker, W. (2004). New challenges in usability evaluation - beyond task-oriented spoken dialogue systems. In *Proceedings of the 8th International Conference on Spoken Language Processing*.
- [Eckert et al., 1997] Eckert, W., Levin, E., and Pieraccini, R. (1997). User modeling for spoken dialogue system evaluation. In *Proceedings of the IEEE Automatic Speech Recognition Workshop*.
- [Eckert et al., 1998] Eckert, W., Levin, E., and Pieraccini, R. (1998). Automatic evaluation of spoken dialogue systems. Technical report, AT&T.
- [Edlund et al., 2006] Edlund, J., Heldner, M., and Gustafson, J. (2006). Two faces of spoken dialogue systems. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Edmonds, 1993] Edmonds, P. (1993). A computational model of collaboration on reference in direction-giving dialogues. Master's thesis, University of Toronto.
- [Filisko and Seneff, 2004] Filisko, E. and Seneff, S. (2004). Error detection and recovery in spoken dialogue systems. In *Proceedings of HLT-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems*.
- [FIPA Consortium, 2002] FIPA Consortium (2002). Fipa communicative act library specification.
- [Fischer, 2001] Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User Adapted Interaction*, 11 :65–86.
- [Fleiss, 1971] Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 :378–382.

- [Georgila et al., 2005a] Georgila, K., Henderson, J., and Lemon, O. (2005a). Learning user simulations for information state update dialogue systems. In *Proceedings of the 9<sup>th</sup> European Conference on Speech Communication and Technology*, pages 893–896.
- [Georgila et al., 2006] Georgila, K., Henderson, J., and Lemon, O. (2006). User simulation for spoken dialogue systems : Learning and evaluation. In *Proceedings of the 9th International Conference on Spoken Language Processing*.
- [Georgila et al., 2005b] Georgila, K., Lemon, O., and Henderson, J. (2005b). Automatic annotation of communicator dialogue data for learning dialogue strategies and user simulations. In *Proceedings of DIALOR'05 Ninth Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*.
- [Gibbon et al., 1997] Gibbon, D., Moore, R., and Winsky, R. (1997). *Handbook of Standards and Resources for Spoken Language Resources*. Walter de Gruyter.
- [Glass, 1999] Glass, J. (1999). Challenges for spoken dialogue systems. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.
- [Glass and Seneff, 2003] Glass, J. and Seneff, S. (2003). Flexible and personalizable mixed-initiative dialogue systems. In *Proceedings of HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*.
- [Glass et al., 2004] Glass, J., Weinstein, E., Cyphers, S., Polifroni, J., Chung, G., and Nakano, M. (2004). A framework for developing conversational user interfaces. In *Proceedings of the 5th International Conference on Computer-Aided Design of User Interfaces*.
- [Glass and Zue, 2000] Glass, J. and Zue, V. (2000). Conversational interfaces : Advances and challenges. In *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, volume 88, pages 1166–1180.
- [Grice, 1957] Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66 :377–388.
- [Grice, 1975] Grice, H. P. (1975). *Syntax and Semantics : Speech Acts*, volume 3, chapter Logic and conversation, pages 41–58. Academic Press.
- [Grosz and Sidner, 1986] Grosz, B. and Sidner, C. (1986). Attention, intentions, and the structure of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*.
- [Grosz and Sidner, 1990] Grosz, B. and Sidner, C. (1990). *Intentions in Communication*, chapter Plans for Discourse, pages 417–444. MIT Press.

- [Henderson et al., 2005] Henderson, J., Lemon, O., and Georgila, K. (2005). Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In *Proceedings of IJCAI 2005 workshop on Knowledge and Reasoning in Practical Dialogue Systems*.
- [Hirschman et al., 1992] Hirschman, L., Bates, M., Dahl, D., Fisher, W., Garofolo, J., Pallett, D., Hunicke-Smith, K., Price, P., Rudnicky, A., and Pao, C. (1992). Multi-site data collection for a spoken language corpus. In *Proceedings of the 2nd International Conference on Spoken Language Processing*.
- [Hirschman et al., 1993] Hirschman, L., Bates, M., Dahl, D., Fisher, W., Garofolo, J., Pallett, D., Hunicke-Smith, K., Price, P., Rudnicky, A., and Tzoukermann, E. (1993). Multi-site data collection and evaluation in spoken language understanding. In *Proceedings of the HLTC 1993 Workshop on Human Language Technology*.
- [Hirschman et al., 1990] Hirschman, L., Dahl, D., McKay, D., Norton, L., and Linebarger, M. (1990). Beyond class a : A proposal for automatic evaluation of discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 109–113.
- [Hirschman et al., 1996] Hirschman, L., Thompson, H., Sundheim, B., Hutchins, J., Black, E., King, M., Pallett, D. D., Fourcin, A., Pols, L. C. W., Oviatt, S., Steeneken, H. J., and Kanai, J. (1996). *Survey of the State of the Art in Human Language Technology*, chapter Evaluation, pages 475–518. Cambridge University Press.
- [Horchani et al., 2007] Horchani, M., Nigay, L., and Panaget, F. (2007). A platform for output dialogic strategies in natural multimodal dialogue systems. In *Proceedings of the International Conference on Intelligent User Interfaces*.
- [James et al., 2000] James, F., Rayner, M., and Hockey, B. A. (2000). "do that again" : Evaluating spoken dialogue interfaces. Technical report, RIACS.
- [Jokinen, 2003] Jokinen, K. (2003). Natural interaction in spoken dialogue systems. In *Proceedings of HCI 2003 Workshop on Ontologies and Multilinguality in User Interfaces*.
- [Jung et al., 2006] Jung, S., Lee, C., and Lee, G. G. (2006). Dialog studio : An example based spoken dialog system development workbench. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.

- [Kamm et al., 1999] Kamm, C., Walker, M., and Litman, D. (1999). Evaluating spoken language systems. In *Proceedings of the Applied Voice Input/Output Society Conference*.
- [Kaplan, 1981] Kaplan, J. (1981). Appropriate responses to inappropriate questions. In *Elements of Discourse Understanding*. Cambridge University Press.
- [King et al., 1996] King, M., Maegaard, B., Schütz, J., des Tombe, L., Bech, A., Neville, A., Arppe, A., Balkan, L., Colin Brace, Bunt, H., Carlson, L., Douglas, S., Höge, M., Krauwer, S., Manzi, S., Mazzi, C., Sielemann, A. J., and Steenbakkens, R. (1996). Expert advisory group on language engineering standards / evaluation working group final report i.
- [L. Lamel, 1995] L. Lamel, M. Adda-Decker, J. G. . . E. M. S. . (1995). Issues in large vocabulary, multilingual speech recognition. In *Proceedings of the 4<sup>th</sup> European Conference on Speech Communication and Technology*.
- [Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 :159–174.
- [Larsson and Traum, 2000] Larsson, S. and Traum, D. (2000). Information state and dialogue management in the trindi dialogue move engine toolkit. In *Natural Language Engineering Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, pages 323–340. Cambridge University Press.
- [Levin et al., 2000] Levin, E., Pieraccini, R., and Eckert, W. (2000). A stochastic model of human - machine interaction for learning dialogue strategies. In *IEEE Transactions on Speech and Audio Processing*.
- [Lin and Lee, 2001] Lin, B.-S. and Lee, L.-S. (2001). Computer-aided analysis and design for spoken dialogue systems based on quantitative simulations. In *IEEE Transactions on Speech and Audio Processing*, volume 9, pages 534–548.
- [Litman and Forbes, 2003] Litman, D. and Forbes, K. (2003). Recognizing emotions from student speech in tutoring dialogues. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.
- [Litman and Pan, 1999] Litman, D. and Pan, S. (1999). Empirically evaluating an adaptable spoken dialogue system. In *Proceedings of the 7th International Conference on User Modeling*.
- [Litman et al., 1998] Litman, D., Pan, S., and Walker, M. A. (1998). Evaluating response strategies in a web-based spoken dialogue agent. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*.

- [Lonsdale and Madsen, 2006] Lonsdale, D. and Madsen, R. (2006). Unifying language modeling capabilities for flexible interaction. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [López-Cózar et al., 2003] López-Cózar, R., De La Torre, A., Segura, J., and Rubio, A. (2003). Assessment of dialogue systems by means of a new simulation technique. *Speech Communication*, 40 :387–407.
- [López-Cózar et al., 2002] López-Cózar, R., Torre, A. D. L., Segura, J., Rubio, A., and Sánchez, V. (2002). Testing dialogue systems by means of automatic generation of conversations. *Interacting with Computers*, 14(5) :521–546.
- [Luzzati, 1993] Luzzati, D. (1993). *Linguistique et traitements automatiques des langues*, chapter Dialogue homme-machine, pages 267–291. C. Fuchs.
- [Luzzati, 1995] Luzzati, D. (1995). *Le dialogue verbal homme-machine*.
- [Luzzati, 2005] Luzzati, D. (2005). Le dialogue oral spontané : quels objets pour quels corpora? In *Actes de Sur les outils informatiques pour des études empiriques*.
- [Luzzati, 2007] Luzzati, D. (2007). Travailler sur corpus, d’hier à aujourd’hui, progrès quantitatifs et progrès qualitatifs. *M.L.M.S.*, 13 :1–0.
- [Makhoul et al., 1989] Makhoul, J., Jelinek, F., Rabiner, L., Weinstein, C., and Zue, V. (1989). White paper on spoken language systems. In *Proceedings of the HLTC 1989 Speech and Natural Language Workshop*.
- [Mariani, 1998] Mariani, J. (1998). The aupelf-uref evaluation-based language engineering action and related projects. In *Proceedings of the 1<sup>st</sup> Language Ressource and Evaluation Conference*, volume 1, pages 123–128.
- [Mariani et al., 2000] Mariani, J., Masson, N., Néel, F., and Chibout, K. (2000). *Ressources et Évaluations en Ingénierie de la Langue*. AUF et De Boeck Université.
- [Martinovska and Vaswani, 2006] Martinovska, B. and Vaswani, A. (2006). Activity-based dialogue analysis as evaluation method. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Maynard et al., 2004] Maynard, H., McTait, K., Mostefa, D., Devillers, L., Rosse, S., Paroubek, P., Bousquet, C., Choukri, K., Goulian, J., Antoine, J.-Y., Béchet, F., Bontron, O., Charnay, L., Romary, L., Vergnes, M., and Vigouroux, N. (2004). Constitution d’un corpus de dialogue oral pour l’évaluation automatique de la compréhension hors et en contexte du dialogue. In *Actes de la 25<sup>ème</sup> édition des Journées d’Étude sur la Parole*.



- [McTear, 1998] McTear, M. (1998). Modelling spoken dialogues with state transition diagrams : Experiences with the cslu toolkit. In *Proceedings of the 5<sup>th</sup> International Conference on Spoken Language Processing*.
- [Minker, 1999] Minker, W. (1999). *Compréhension automatique de la parole spontanée*. L'Harmattan.
- [Minker and Bennacef, 2000] Minker, W. and Bennacef, S. (2000). *Parole et dialogue homme machine*. Eyrolles.
- [Minsky, 1975] Minsky, M. (1975). A framework for representing knowledge. In Winston, P., editor, *Psychology of Computer Vision*, pages 211–277. McGraw-Hill.
- [Moor, 1976] Moor, J. (1976). An analysis of the turing test. *Philosophical Studies*, 30 :249–257.
- [Nguyen and Caelen, 2003] Nguyen, H. and Caelen, J. (2003). Generic manager for spoken dialogue systems. In *Proceedings of the 7<sup>th</sup> Workshop on the Semantics and Pragmatics of Dialogue*, pages 201–202.
- [on the France Telecom 3000 Voive Agency Corpus, 2007] on the France Telecom 3000 Voive Agency Corpus, S. L. U. S. (2007). Géraldine damnati and frédéric béchet and renato de mori. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 48–55.
- [Paek, 2001] Paek, T. (2001). Empirical methods for evaluating dialog systems. In *Proceedings of ACL/EACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*.
- [Paek, 2006] Paek, T. (2006). Reinforcement learning for spoken dialogue systems : Comparing strengths and weaknesses for practical deployment. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Pearlman, 2000] Pearlman, J. (2000). Sls-lite : Enabling spoken language systems design for non-experts. Master's thesis, Massachusetts Institute of Technology.
- [Pietquin, 2004] Pietquin, O. (2004). *A Framework for Unsupervised Learning of Dialogue Strategies*. PhD thesis, Faculté Polytechnique de Mons.
- [Pietquin and Dutoit, 2006] Pietquin, O. and Dutoit, T. (2006). A probabilistic framework for dialog simulation and optimal strategy learning. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 14.
- [Polifroni and Seneff, 2000] Polifroni, J. and Seneff, S. (2000). Galaxy-ii as an architecture for spoken dialogue evaluation. In *Proceedings of the 2<sup>nd</sup> Language Ressource and Evaluation Conference*.

- [Power, 1979] Power, R. (1979). The organization of purposeful dialogues. *Linguistics*, 17 :107–152.
- [Price et al., 1992] Price, P., Hirschman, L., Shriberg, E., and Wade, E. (1992). Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the HLTC 1992 Speech and Natural Language Workshop*.
- [Rajman et al., 2003] Rajman, M., Rajman, A., Seydoux, F., and Trutnev, A. (2003). Prototypage rapide et évaluation de modèles de dialogues finalisés. In *Actes de la 10ème Conférence sur le Traitement Automatique des Langues Naturelles*.
- [Reichman, 1981] Reichman, R. (1981). *Plain Speaking : A Theory and Grammar of Spontaneous Discourse*. PhD thesis, Harvard University, Cambridge, Massachusetts.
- [Rosset, 2001] Rosset, S. (2001). *Stratégies et gestionnaire de dialogue pour des systèmes d'interrogation de bases de données à reconnaissance vocale*. PhD thesis, LIMSI.
- [Rosset et al., 2005] Rosset, S., Galibert, O., and Illouz, G. (2005). Human-computer dialog system in an open domain., edinburgh, july 2005. In *Proceedings of Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*.
- [Rouillard and Caelen, 2001] Rouillard, J. and Caelen, J. (2001). Le système halpin : recherche documentaire en langue naturelle et dialogue multimodal. *Revue RIHM*, 2(2) :55–82.
- [Sabah, 1997] Sabah, G. (1997). Intelligence artificielle et sciences cognitives. In *InCognito*. Association In Cognito.
- [Sadek, 1991] Sadek, D. (1991). *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication*. PhD thesis, Rennes I.
- [Sadek, 1999] Sadek, D. (1999). Design considerations on dialogue systems : From theory to technology - the case of artimis -. In *Proceedings of the European Speech Communication Association*.
- [Sadek et al., 1997] Sadek, D., Bretier, P., and Panaget, F. (1997). Artimis : Natural dialogue meets rational agency. In *Proceedings of the 15<sup>th</sup> International Joint Conference on Artificial Intelligence*, pages 1030–1035.
- [Sadek et al., 1996] Sadek, D., Ferrieux, A., Cozannet, A., Bretier, P., Panaget, F., and Simonin, J. (1996). Effective human-computer cooperative spoken dialogue : The ags demonstrator. In *Proceedings of the 4<sup>th</sup> International Conference on Spoken Language Processing*.

- [Schatzman et al., 2005] Schatzman, J., Stuttle, M. N., Weilhammer, K., and Young, S. (2005). Effects of the user model on simulation-based learning of dialogue strategies. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.
- [Schatzmann et al., 2005] Schatzmann, J., Geogila, K., and Young, S. (2005). Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*.
- [Scheffler and Young, 2000] Scheffler, K. and Young, S. (2000). Probabilistic simulation of human - machine dialogues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1217–1220.
- [Scheffler and Young, 2001] Scheffler, K. and Young, S. (2001). Corpus - based dialogue simulation for automatic strategy learning and evaluation. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.
- [Searle, 1969] Searle, J. (1969). *Speech Act, an Essay in the Philosophy of Language*. Cambridge University Press.
- [Searle, 1972] Searle, J. (1972). *Les actes de langage*. Éditions Hermann. Traduction de [Searle, 1969].
- [Searle, 1982] Searle, J. (1982). *Sens et Expression, Etude de théorie des Actes de langage*. Éditions de Minuit.
- [Sidner and Rich, 2006] Sidner, C. and Rich, C. (2006). Do dialogue systems really make a difference? evaluating a dialogue system for use with consumer products. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Singh, 1992] Singh, M. (1992). A critical examination of the cohen-levesque theory of intentions. In *Proceedings of the 10th European Conference on Artificial Intelligence*.
- [Smith and Hipp, 1994] Smith, R. and Hipp, R. (1994). *Spoken Natural Language Dialog Systems : a Practical Approach*. Oxford University Press.
- [Sperber and Wilson, 1986] Sperber, D. and Wilson, D. (1986). *Relevance, communication and cognition*. Basil Blackwell.
- [Sperber and Wilson, 1989] Sperber, D. and Wilson, D. (1989). *La pertinence, communication et cognition*. Éditions de Minuit. Traduction de [Sperber and Wilson, 1986] par A. Gerschenfeld et D. Sperber.

- [Traum and Larsson, 2003] Traum, D. and Larsson, S. (2003). The information state approach to dialogue management. In Smith and Kuppevelt, editors, *Current and New Directions in Discourse & Dialogue*, pages 325–353. Kluwer Academic Publishers.
- [Traum, 1999] Traum, D. R. (1999). *Foundations of Rational Agency*, chapter Speech Acts for Dialogue Agents, pages 169–201. Kluwer Academic Publishers.
- [Traum, 2000] Traum, D. R. (2000). 20 questions on dialogue acts taxonomies. *Journal of Semantics*, 17 :7–30.
- [Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49 :433–460.
- [Villaneau, 2000] Villaneau, J. (2000). Un système basé sur les types logiques pour la compréhension de la parole. In *Acte de la 7ème Conférence sur le Traitement Automatique des Langues Naturelles*.
- [Walker et al., 1997a] Walker, M., Hindle, D., Fromer, J., Fabrizio, G. D., and Mestel, C. (1997a). Evaluating competing agent strategies for a voice email agent. In *Proceedings of the 5th European Conference on Speech Communication and Technology*.
- [Walker et al., 2000] Walker, M., Hirschman, L., and Aberdeen, J. (2000). Evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 2nd Language Resource and Evaluation Conference*.
- [Walker et al., 1997b] Walker, M., Litman, D., Kamm, C., and Abella, A. (1997b). Evaluating interactive dialogue systems : Extending component evaluation to integrated system evaluation. In *Interactive Spoken Dialog Systems : Bridging Speech and NLP Together in Real Applications*. Association for Computational Linguistics.
- [Walker et al., 1997c] Walker, M., Litman, D., Kamm, C., and Abella, A. (1997c). Paradise : A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 271–280.
- [Walker et al., 1998] Walker, M., Litman, D., Kamm, C., and Abella, A. (1998). Evaluating spoken dialogue agents with paradise : Two case studies. In *Computer Speech and Language*. Academic Press.
- [Walker et al., 2001] Walker, M., Passonneau, R., and Boland, J. (2001). Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 512–522.

- [Walker, 1994] Walker, M. A. (1994). Experimentally evaluating communicative strategies : The effect of the task. In *Proceedings of the 12<sup>th</sup> National Conference on Artificial Intelligence*, volume 1, pages 86–93.
- [Walker and Moore, 1997] Walker, M. A. and Moore, J. D. (1997). Empirical studies in discourse. *Computational Linguistics*, 23(1) :1–12.
- [Ward, 2006] Ward, N. G. (2006). Evaluating real-time responsiveness in dialog. In *Proceedings of InterSpeech 2006 Satellite Workshop : Dialogue on Dialogues, Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*.
- [Watanabe et al., 1998] Watanabe, T., Araki, M., and Doshita, S. (1998). Evaluating dialogue strategies under communication errors using computer-to-computer simulation. *Institute of Electronics, Information and Communication Engineers Transactions on Information and Systems*, 81-part 9 :1025–1033.
- [Webb et al., 2000] Webb, N., Roeck, A. D., Kruschwitz, U., Scott, P., Steel, S., and Turner, R. (2000). Evaluating a natural language dialogue system : Results and experiences. In *Proceedings of the LREC 2000 Workshop on From Spoken Dialogue to Full Natural Interactive Dialogue : Theory, Empirical Analysis and Evaluation*.
- [Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9 :26–45.
- [Winograd, 1972] Winograd, T. (1972). *Understanding natural language*.
- [Witt and Williams, 2003] Witt, S. M. and Williams, J. D. (2003). Two studies of open vs. directed dialog strategies in spoken dialog systems. In *Proceedings of the 8th European Conference on Speech Communication and Technology*.
- [Young, 2002a] Young, S. (2002a). The statistical approach to the design of spoken dialogue systems. Technical report, Cambridge University Engineering Department.
- [Young, 2002b] Young, S. (2002b). Talking to machines (statistically speaking). In *Proceedings of the 7th International Conference on Spoken Language Processing*.
- [Zeiliger et al., 1997] Zeiliger, J., Caelen, J., and Antoine, J.-Y. (1997). Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme-machine. In *Actes des Journées Scientifiques et Techniques*.

- [Zue, 1997] Zue, V. (1997). Conversational interfaces : Advances and challenges. In *Proceedings of the 5th European Conference on Speech Communication and Technology*.