Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat)

Jean-Yves Antoine

VALORIA - Université de Bretagne Sud

IUP Vannes, r. Yves Mainguy, 56000 Vannes

Mél : Jean-Yves.Antoine@univ-ubs.fr

Toile: web.iu-vannes.fr/public/IUP/recherche/JYA

Jean Caelen

CLIPS - IMAG

BP 53, 38041 Grenoble Cedex 9

Mél : Jean.Caelen@imag.fr

Toile : http://herakles.imag.fr/Geod/

Rubrique — Ingénierie de la langue

Titre court — Evaluation qualitative de la compréhension de parole : le paradigme DCR

Title — Towards an objective, predictive and generic assessment of spoken language understanding: the DCR paradigm

Mots clés — Evaluation, compréhension de parole, predictivité, objectivité

Correspondance — Jean-Yves Antoine, VALORIA, IUP Vannes, rue Yves Mainguy, 56000 Vannes. Tél : 02 97 68 32 10, Mél : Jean-Yves.Antoine@univ-ubs.fr

Tables

Figures

Résumé — Lorsqu'elle est envisagée de manière quantitative, l'évaluation de la compréhension de la parole est généralement basée sur des mesures globales de performance. Cette approche, largement utilisée dans le programme DARPA-ATIS, pèche cependant par son caractère prédictif limité et son manque de généricité. Dans cet article, nous présentons un nouveau paradigme d'évaluation cherchant à répondre à ces limites.

Abstract — Generally speaking, the objective evaluation of spoken language understanding is based on the measurement of the overall performances of the system. This approach, which was chiefly used in the DARPA-ATIS program, lacks however some predictive power to enable a really informative diagnosis. We present in this paper a novel methodology that intends to achieve a really predictive and generic evaluation.

1. Problématique

La communication orale homme-machine (CHM orale par la suite) a atteint une maturité que traduit l'apparition récente d'applications réelles telles que, par exemple, le système automatique de routage téléphonique grand public mis en place par AT&T (Lokbani & White, 1998), ou encore le système de réservation par téléphone des chemins de fer néérlandais faisant suite au projet européen ARISE (Baggia *et al*, 1999). D'une manière générale, l'ensemble des traitements automatiques impliqués dans la CHM orale a connu des progrès significatifs au cours des dernières années. Pour capitaliser ces avancées et orienter les recherches futures, la mise en place de procédures d'évaluation adaptées au dialogue oral constitue un enjeu central pour la communauté parole. Force est néanmoins de reconnaître qu'en dépit de plusieurs campagnes d'évaluation lourdes menées dans le domaine (Pallett *et al*., 1994; Hirschman 1998), le développement des systèmes de dialogue oral repose toujours sur un certain empirisme (Bernsen & Dybkjaer, 1997). En particulier, on ne dispose pas encore d'indices suffisamment prédictifs pour orienter les recherches, ou simplement certains choix de mise en œuvre, à partir d'une analyse préalable des besoins[1] (critère de **prédictivité** non

[1] Il peut s'agir par exemple, d'une analyse linguistique ou ergonomique des usages effectuée aussi bien sur des corpus

satisfait).

Une des causes de ces difficultés réside dans la complexité des systèmes de dialogue oral. La structure de ces systèmes, composés de plusieurs modules interdépendants, repose généralement su une architecture semblable à celle donnée ci-dessous[2] (figure 1).

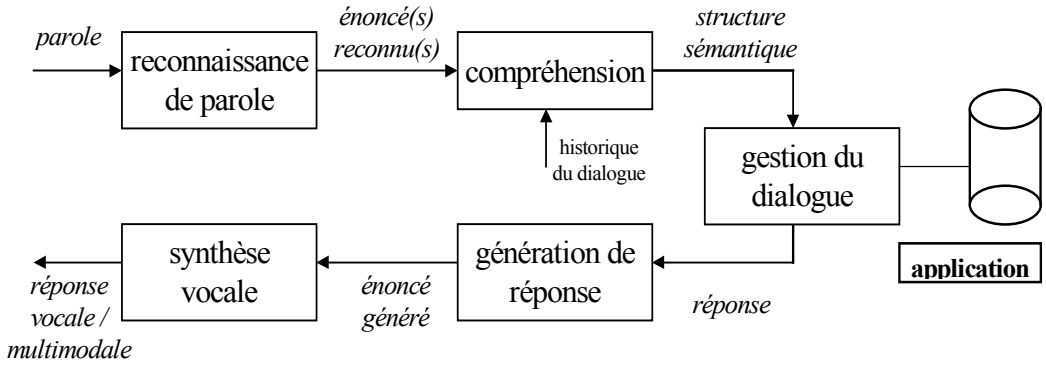


Figure 1 — Architecture générale d'un système de dialogue oral homme-machine.

La requête orale de l'utilisateur est tout d'abord traitée par un module de reconnaissance de parole qui fournit en sortie un (ou plusieurs) énoncé(s) reconnu(s). Ce processus de reconnaissance intègre un modèle de langage qui rend compte plus ou moins de la syntaxe de la langue[3]. L'énoncé reconnu est alors adressé au module de compréhension qui en construit une représentation sémantique. Nous reviendrons au paragraphe suivant sur une présentation plus précise de cette étape de traitement.

---

d'interaction humaine que sur des corpus simulés (technique du magicien d'Oz) de communication homme-machine.

[2] Le caractère sériel de cette architecture n'est en fait acceptable qu'en première approximation. De nombreux systèmes intègrent ainsi fortement les processus de compréhension et de reconnaissance de parole (Antoine & Genthial, 1999). De même, certains aspects de la compréhension (résolution des anaphores par exemple) peuvent être gérés par le module de dialogue.

[3] Ces modèles reposent fréquemment sur une approche statistique limitant la syntaxe à la probabilité d'occurence de séquences de N mots (modèles N-grams). Il s'agit d'une modélisation appauvrie nécessitée par le caractère agrammatical du langage parlé (cf § 2.1.). Elle ne concerne que la dimension linéaire du langage et s'avère incapable de rendre compte de la structure profonde des énoncés. D'où les nombreuses recherches qui sont menées à l'heure actuelle sur l'intégration de connaissances syntaxiques plus fines au processus de reconnaissance (Antoine & Genthial, 1999).

La structure sémantique de l'énoncé est ensuite adressée au module de dialogue (Siroux *et al*., 1995; Pieraccini *et al*., 1997) qui a à gérer :

- l'*interface avec l'application* — Pour un système d'information, la représentation sémantique est transformée en une requête d'interrogation SQL[4], qui donne en retour un ensemble de réponses extraites de la base de données de l'application.
- la *gestion du dialogue* — Le module de dialogue doit gérer la construction interactive de la solution au problème posé par l'utilisateur en lui présentant les réponses obtenues, en lui demandant des éclaircissements ou en devançant ses requêtes.

Enfin, les réponses élaborées par le module de dialogue sont transmises à l'utilisateur par l'intermédiaire d'une chaîne de génération variant en fonction de la modalité de sortie du système.

Plusieurs approches peuvent être envisagées pour l'évaluation de tels systèmes complexes. On peut les caractériser suivant deux dimensions transversales :

- **Evaluation objective / subjective** — L'évaluation est subjective (Bernsen *et al*., 1995; Lamel *et al*., 1995) lorsqu'elle repose sur une analyse de l'opinion des utilisateurs sur le système : facilité d'utilisation, convivialité, etc. A l'opposé, l'évaluation est dite objective lorsqu'elle appréhende les performances du système par des mesures quantitatives du type taux d'erreurs. Jusqu'à présent, cette approche centrée technologie a été largement privilégiée (Hirschman, 1998). Certains auteurs ont par ailleurs montré qu'une évaluation subjective est peu fiable car trop dépendante de chaque utilisateur. Il n'en reste pas moins que la finalité des systèmes de CHM orale reste la satisfaction de l'utilisateur. Certaines propositions telles que le paradigme PARADISE (Walker *et al*., 1997) visent ainsi à combiner évaluations objective et subjective.
- **Evaluation globale / détaillée** — D'autre part, l'évaluation peut également considérer le système globalement, ou s'attacher au contraire à l'étude de chacun de ses composants. On parle dans le

[4] SQL (Structured Query Language) est un langage informatique spécialisé dans le domaine des bases de données. Il est

premier cas d'une évaluation de type "boîte noire" (*black box*) et d'une évaluation de type "boîte grise" (*glass box*) dans l'autre.

Compte tenu des multiples interactions entre les différents modules de traitements d'un système de dialogue oral, une évaluation de type *glass box* ne peut suffire à estimer les performances globales du système (Polifroni *et al.*, 1998). A l'opposé, une évaluation de type boîte noire ne peut conduire à aucun diagnostic sur le comportement des différents niveaux de traitement du système. Aussi est-il généralement recommandé de conduire de front ces deux approches (Polifroni *et al.*, 1998; Minker 1998).

Ainsi, un système peut être évalué suivant diverses approches. Dans cet article, nous nous intéresserons spécifiquement à l'évaluation objective du module de compréhension des systèmes de dialogue. Après une présentation rapide de la compréhension automatique de la parole en CHM orale, nous proposerons en premier lieu une revue critique des méthodes actuelles d'évaluation objective de la compréhension. Nous montrerons en particulier qu'elles ne satisfont pas aux critères essentiels de **prédictivité** (cf *supra*) et de **généricité** (généralisation à d'autres domaines des résultats obtenus sur une application ; indépendance de l'évaluation vis à vis des méthodes utilisées par les systèmes). Nous proposerons alors une approche alternative adaptée à la CHM orale : le paradigme DCR. Après avoir présenté les principes de cette méthodologie, nous en détaillerons la mise en œuvre dans le cadre de l'Action de Recherche Concertée (ARC par la suite) "Dialogue Oral" de l'AUPELF-UREF. Nous montrerons d'une part que le paradigme DCR remplit les objectifs des approches traditionnelles, et que d'autre part il apporte des réponses en matière de prédictivité et de généricité.

2. Evaluation objective de la compréhension de la parole : l'approche ATIS

La compréhension de la parole a pour tâche de fournir une représentation rendant compte du sens de l'énoncé. Compte tenu des contraintes fortes de robustesse qui sont imposés aux systèmes de

---

utilisé en particulier pour formuler les requêtes d'interrogation de la base de données.

dialogue oral, cette analyse sémantique est le plus souvent envisagée suivant des approches qui diffèrent sensiblement de celles rencontrées en linguistique ou en traitement du langage écrit (TALN). Il semble donc utile de présenter rapidement cette étape de traitement avant d'aborder la question de son évaluation.

2.1. Compréhension automatique de la parole en CHM orale

Tout en travaillant sur un objet d'étude qu'elle partage avec les sciences du langage et le TALN, la CHM orale, et plus précisément la compréhension automatique de la parole, doivent faire face à des problèmes spécifiques. Ceux-ci sont de deux natures différentes :

1) Tout d'abord, la compréhension de la parole intervient après (ou en parallèle avec) une étape de reconnaissance automatique de la parole qui travaille sur un signal acoustique non décodé, et non pas sur un énoncé linguistiquement connu. Il en résulte que le module de compréhension peut être amené à travailler non pas sur l'énoncé effectivement prononcé, mais sur une "traduction" comportant des erreurs de reconnaissance. Par exemple :

   Bonjour le musée Grévin est-il ouvert le matin ?

   * *Bonjour le musée Rodin est-il ouvert ce matin ?*

2) D'autre part, le caractère spontané du langage parlé se traduit au niveau linguistique par l'apparition de nombreux inattendus structurels qui cassent la régularité syntaxique de l'énoncé (Blanche-Benveniste, 1990). On citera en particulier les procédés suivants :

   - hésitations *Je voudrais un aller simple pour euh attendez Rosporden c'est cela*

   - répétitions *Pouvez vous me dire s'il y a s'il y a un restaurant dans la gare*

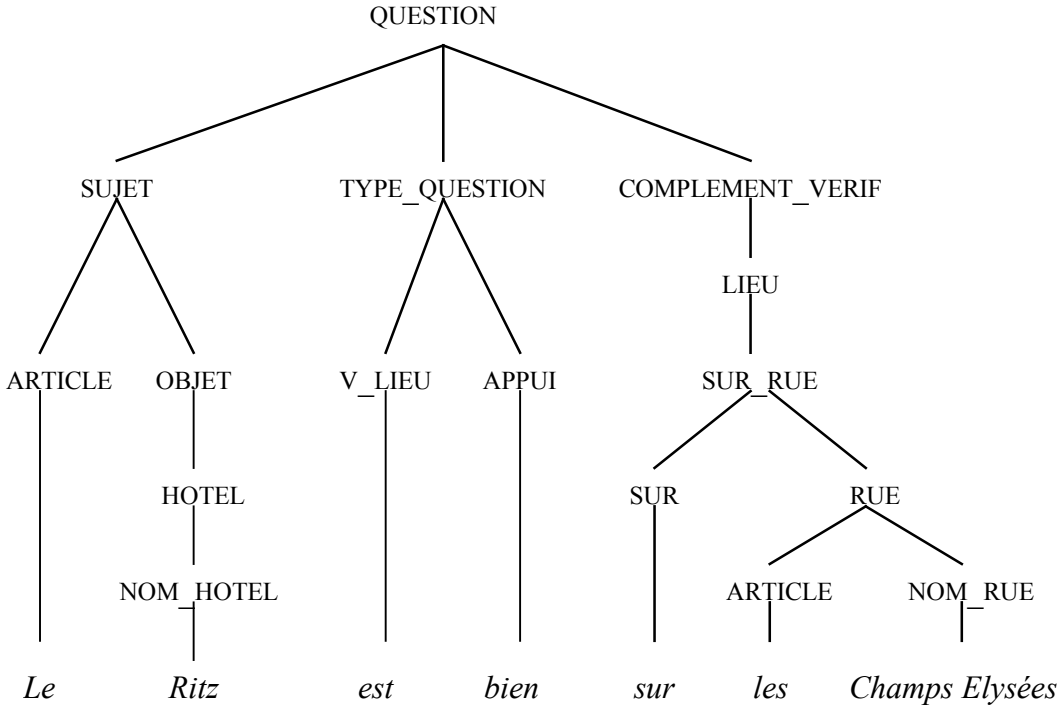
   - reprises *Je voudrais un biller pour demain enfin pour demain matin si possible*

   - corrections *Ce serait donc pour le retour non excusez-moi pour l'aller*

   - incises *Je voudrais le premier départ c'est à dire que je suis très pressé hein vous comprenez et donc c'est pour Paris*

- ellipses *Pour le retour c'est pareil qu'à aller sauf pour la réduction bien sûr*

La fréquence élevée d'apparition de ces procédés irréguliers (Lettelier-Zarshenas *et al.*, 1999), associée aux erreurs de reconnaissance, expliquent la nécessité de méthodes spécifiques au traitement de l'oral. A l'heure actuelle, il n'existe pas de réel consensus sur l'approche à adopter pour la compréhension de la parole. Différentes méthodes sont ainsi utilisées (Antoine & Genthial, 1999), qui vont d'une modélisation conceptuelle entièrement stochastique (Pieraccini & Levin, 1992) à l'utilisation d'approches à base de règles, éventuellement probabilistes (Seneff, 1992). D'un point de vue linguistique, certaines approches sont basées sur une analyse fine conduisant à la construction d'une structure sémantique complète représentant l'ensemble des relations de dépendances au sein de l'énoncé (Seneff, 1992; Antoine, 1996). La figure 2 représente par exemple la représentation[5] — vue comme un arbre de dérivation — construite par le système de compréhension TINA du MIT (Seneff, 1992).



[5] On remarquera que de nombreuses étiquettes sémantiques reposent en réalité sur une connaissance du monde de la tâche (A-HOTEL par exemple). Nous sommes donc plutôt en présence, comme souvent en CHM orale, d'une analyse sémantico-pragmatique profitant du caractère finalisé du domaine d'application visé.

Figure 2 — Structure pragmatico-sémantique élaborée par le système TINA pour l'énoncé *Le Ritz est bien sur les Champs-Elysées ?* Exemple adapté de (Seneff, 1992).

A l'opposé, les méthodes les plus répandues à l'heure actuelle peuvent être qualifiées de sélectives (De Mori, 1994) : seuls certains îlots clefs de l'énoncé sont jugés pertinents pour représenter le sens "utile" de l'énoncé, c'est à dire permettant de construire la requête SQL adéquate. Reprenons l'exemple de l'énoncé précédent (*Le Ritz est bien sur les Champs-Elysées ?*). Dans cette phrase, trois mots clés seulement vont être considérés par une approche sélective :

type_requete	*est*
nom_objet	*Ritz*
lieu	*Champs-Elysées*

On obtient alors une structure sémantico-pragmatique partielle, correspondant à l'information minimale de l'énoncé nécessaire à son interprétation sous la forme de requête SQL. Cette structure se représente généralement sous la forme d'un cadre (*frame*) sémantique (Oerder & Aust, 1994) :

```
<type_requete = confirmer_lieu>
<objet =    <type = hotel>
            <nom  = Ritz> >
<lieu> =    Champs-Elysees>
```

Le caractère partiel de cette analyse sélective garantit une certaine robustesse de traitement face aux inattendus structurels de l'oral. Les approches sélectives présentent ainsi des performances encourageantes dans le cadre d'une CHM orale très finalisée. C'est ce caractère finalisé qui permet de limiter la compréhension à de telles approches linguistiquement pauvres. La généralisation des approches sélectives à des contextes applicatifs moins restreints, ou à la considération de phénomènes de langues plus fins dans l'optique d'une meilleure gestion du dialogue, n'est cependant pas garantie. Cette question, essentielle à nos yeux, ne constitue pas l'objet central de cet article et ne sera abordée qu'en filigrane du point de vue de la généricité de l'évaluation. Pour une argumentation plus complète sur le sujet, le lecteur pourra se reporter par exemple à (Antoine,

1995).

2.2. Le programme ATIS d'évaluation de la compréhension de parole

L'investissement le plus conséquent mené à ce jour en matière d'évaluation des systèmes de dialogue oral est sans aucun doute celui du programme ATIS de l'ARPA[6] (Pallett *et al*., 1994). Conduit sur une période de 5 ans, ce programme s'est intéressé à l'évaluation objective des systèmes de dialogue dans le cadre bien précis du renseignement aérien (ATIS : *Air Transport Information Systems*). Cette spécificité du domaine d'application ne va pas sans poser des problèmes de généricité sur lesquels nous reviendrons ultérieurement. L'évaluation a porté à la fois sur les performances globales des systèmes et sur celles des modules de reconnaissance et de compréhension (Hirschman, 1998). D'autres programmes d'évaluation ont également été menés en dehors du cadre ATIS. Il concernaient essentiellement les systèmes d'information : météorologie (Polifroni et al., 1998), renseignement ferroviaire (Baggia *et al*., 1999; Gauvain *et al*., 1997; Aust *et al*., 1995). Les évaluations conduites étaient subjectives ou suivaient une démarche proche de celle du programme ATIS, spécifiquement étudié dans cet article.

Comme nous l'avons noté, la compréhension peut reposer sur des approches sensiblement différentes. Cette diversité pose problème pour les campagnes d'évaluation, où l'on doit définir une plate-forme commune comportant les mêmes représentations sémantiques. Afin de surmonter cette difficulté, l'évaluation ATIS a porté non pas sur les représentations sémantiques produites par la compréhension, mais sur les réponses obtenues après interrogation de la base de données (Hirschman, 1998; Minker, 1998). L'idée étant qu'un énoncé compris correctement et traduit en une requête SQL donnera une réponse cohérente de la part de la base de données.

---

[6] ARPA (ou DARPA) : agence responsable des activités de recherche dépendant du ministère américain de la Défense., L'ARPA a financé plusieurs projets phares pour le développement et l'évaluation de systèmes en ingénierie linguistique écrite ou orale.

Référence Minimale		Référence Maximale		
code-tarif	no-vol	code-tarif	tarif	no-vol
Plein_Ciel	AF_2137	Plein_Ciel	615 F	AF_2137
Azur	AF_2137	Azur	875 F	AF_2137

Tableau 1 — Evaluation ATIS de la compréhension à travers les réponses globales du système : références correspondant à la requête : *Quels sont les tarifs sur les vols Paris-Toulouse arrivant avant 15h ?*. Exemple adapté de (Minker, 1998) .

D'une manière plus précise, chaque requête est associée à deux réponses canoniques appelées référence minimale et référence maximale (tableau 1). La référence minimale contient l'ensemble des informations attendues explicitement par l'utilisateur, tandis que la référence maximale contient des informations supplémentaires considérées comme cohérentes avec la requête[7]. La réponse du système est alors jugée correcte si elle contient la référence minimale et n'excède par la référence maximale.

Le tableau 1 illustre la différence qui existe entre le sens de la requête et la réponse de la base de données. Il s'agit clairement d'une méthodologie adaptée à une compréhension sélective, puisqu'on ne vérifie à travers la réponse de la base de données que la bonne compréhension des îlots clés "utiles" à la formulation de la requête SQL.

L'évaluation ATIS est conduite suivant deux types de tests :

- **Enoncés de type A** — Il s'agit d'énoncés dont la compréhension est indépendante du contexte dialogique. Ils évaluent la *compréhension littérale* des énoncés, considérés hors de tout contexte.
- **Enoncés de type D** — Il s'agit d'énoncés dont l'interprétation dépend du contexte dialogique. Ils

[7] Dans l'exemple du tableau 1, le terme *tarif* est ambigu. L'intention de l'usager peut être en effet d'obtenir les différents prix ou bien les différentes classes de tarifs disponibles sur ce vol. Il est à noter que le paradigme d'évaluation ATIS ne permet pas l'emploi de jeux multiples de références pour rendre compte de ces ambiguïtés.

évaluent la *compréhension réelle* (ou *complète*), qui correspond à la contextualisation de la représentation sémantique (résolution des références anaphoriques et des ellipses, par exemple).

Enfin, l'évaluation de la compréhension est conduite de deux manières différentes afin d'étudier les interactions entre reconnaissance de parole et compréhension (Hirschman, 1998) :

- **compréhension langagière** (*natural language understanding error rate*) : on considère en entrée la transcription de l'énoncé prononcé et en sortie la réponse de la base de données.
- **compréhension de la parole** (*spoken language understanding error rate*) : on considère en entrée le signal de parole et en sortie la réponse de la base.

On notera que cette distinction nécessite de séparer — du moins pour l'évaluation — les processus de reconnaissance et compréhensoin. Or, cette contrainte n'est pas obligatoirement respectée par tous les systèmes. On retrouve, cette fois sous un angle architectural, le problème de la généricité de l'évaluation ATIS. Nous allons précisément étudier les limites de ce paradigme.

3. Limites de l'évaluation ATIS

Par son étude ambitieuse des performances des systèmes de dialogue, le programme ATIS a beaucoup apporté à la CHM orale. Outre la généralisation du recours à l'évaluation qu'il a initié, ce programme a fourni une vision globale des avancées de la CHM orale, au moment où celle-ci arrivait à maturité (cf. tableau 2). Il a ainsi contribué à la consolidation d'une communauté qui s'est ainsi vue renforcée dans ses choix. L'apport de ce type d'évaluation apparaît cependant limité quant à la conduite de recherches futures sur le sujet. Deux caractéristiques essentielles manquent en effet à ce paradigme : une généricité appréciable et un caractère prédictif marqué.

3.1. Prédictivité

Par delà l'observation globale de performances, le principal résultat que l'on attend d'une série de tests réside dans la définition de critères prédictifs pour l'évaluation **prospective** des systèmes. Ces critères prédictifs sont essentiels dans la phase de conception d'un système ainsi que dans la

conduite de recherches à long terme, dans la mesure où ils évitent de s'engager dans certaines voies vouées à l'échec.

Or, une analyse détaillée du programme ATIS montre que celui-ci reste très perfectible en terme de prédictivité. D'une part, l'évaluation de la compréhension au niveau des réponses de la base de données prête à caution. En effet, il n'est pas impossible qu'une représentation sémantique erronée conduise à une réponse correcte de la base de données (Minker, 1998). Certains auteurs ont alors envisagé de réaliser l'évaluation au niveau des représentations sémantiques (Minker & Bennacef, 1996). Cette solution nécessite la définition de représentations de références communes. Nous verrons plus loin que cette approche pose problème en matière de généricité.

Par ailleurs, l'évaluation ATIS repose sur le calcul d'un taux d'erreur portant sur l'ensemble des énoncés du corpus de test, sans classification de ces derniers suivant des critères linguistiques, pragmatiques ou autres. Or, quels enseignements tirer d'un taux aussi global ? Le système est-il déficient en matière de résolution des co-références ? Le modèle conceptuel est-il adapté à la complexité de la tâche ? A dire vrai, les résultats de l'évaluation ATIS ne permettent pas de trancher. Ainsi, les systèmes les plus performants au vu de l'évaluation (tableau 2) reposent sur des approches sensiblement différentes — modèle conceptuel stochastique pour AT&T (Pieraccini & Levin, 1992) et règles de grammaire probabilistes pour le MIT (Seneff, 1992) — sans qu'aucun résultat ne permette de distinguer leurs faiblesses et points forts respectifs.

Taux d'erreur (%)	AT&T	CMU	MIT	SRI	BBN	Unisys	MITRE
Compréhension langagière	3,8	3,8	4,5	7,0	9,4	23,6	30,6
Compréhension de parole	7,0	7,4	10,3	10,6	11,9	27,4	44,9

Tableau 2 — Campagne ATIS 1994 : évaluation de la compréhension de la parole. Taux d'erreurs globaux sur des énoncés de type A (indépendants du contexte dialogique). D'après (Pallett et al., 1994).

Ce caractère global interdit ainsi à l'évaluation ATIS toute conclusion prédictive. Ce fait est d'ailleurs bien connu des chercheurs, pour qui l'étude des sessions d'utilisation du système (analyse

de *logfiles*) est autrement plus riche d'enseignements (Polifroni, *et al*. 1998). Il est alors possible de détecter certains dysfonctionnements du système sur des phénomènes précis. Néanmoins, cette évaluation par étude de cas n'est ni systématique, ni objective. Elle est donc inadaptée à des campagnes d'évaluation regroupant de multiples participants. Elle montre en revanche que le caractère prédictif de l'évaluation va de pair avec son caractère discriminant.

3.2. Généricité

La généricité est une propriété essentielle en ce sens qu'elle permet d'étendre la portée des résultats obtenus par une campagne d'évaluation spécifique. Elle peut également être une garantie de réutilisabilité de la procédure de test. Cette généricité revêt plusieurs aspects :

- **généricité vis à vis de l'application** — Cette propriété est une garantie contre tout enfermement dans l'amélioration de méthodes ad hoc. Les participants du programme ATIS regrettent la trop grande spécificité de cette évaluation, qui ne fournit aucune indication sur la portabilité des méthodes employées à d'autres cadres applicatifs (Hirschman, 1998; Minker, 1998). Il nous semble que ce problème est cependant sous-estimé et ne se limite pas à la réutilisabilité du système d'une tâche de renseignement à une autre. Or, l'avenir de la CHM orale dépend en premier lieu de la capacité des systèmes de dialogue oral à aborder des contextes applicatifs sensiblement plus riches que ceux sur lesquels ils ont fait leurs preuves jusqu'ici. Par son manque de généricité, l'évaluation ATIS n'est que peu éclairante sur ce futur proche de la CHM.

- **généricité vis à vis des méthodes** — Sous peine d'introduire un biais dans la mesure, un paradigme d'évaluation ne doit reposer sur aucun a priori quant aux représentations et aux méthodes utilisées par le système. Effectuée au niveau des réponses de la base de données, l'évaluation ATIS satisfait à première vue ce critère. On peut néanmoins douter de la généricité d'un postulat qui identifie représentation sémantique et réponse du système. Si, comme nous l'avons vu au paragraphe 2.2, celui-ci est adapté à une compréhension sélective, on sera plus circonspect quand à sa généralisation à des approches plus fouillées. L'évaluation ATIS s'interdit ainsi de rendre compte d'approches plus fines dont l'intérêt risque de s'afirmer dans un avenir

proche (Antoine & Genthial, 1999). Une évaluation directe au niveau des représentations sémantiques semble donc préférable (Minker & Bennacef, 1996). Cette approche nécessite cependant une définition de représentations canoniques communes qui impose une dépendance de l'évaluation vis à vis des méthodes utilisées. Ainsi, les propositions d'évaluation à ce niveau (Minker & Bennacef, 1996) reposent sur des références à base de représentations sélectives.

La CHM orale a besoin d'une méthodologie générique et prédictive qui lui permette d'étendre réellement la portée de ses premiers succès. Comme le montre cette étude, l'approche ATIS ne peut jouer ce rôle en l'état. La méthodologie DCR proposée dans cet article tente précisément d'apporter une réponse à ce problème.

4. Critères pour une évaluation générique et prédictive

En conclusion de cette analyse critique du paradigme ATIS, deux conditions semblent nécessaires à la mise en oeuvre d'une évaluation générique et prédictive.

D'une part, l'évaluation doit être discriminante tout en restant quantitative, donc objective. C'est à dire qu'elle doit se composer de sous-sessions d'évaluation portant sur des phénomènes linguistiques bien identifiés (traitement des répétitions, par exemple). Ces phénomènes étant indépendants du contexte applicatif choisi, on disposera ainsi d'un diagnostic détaillé qui sera généralisable dans une certaine mesure à d'autres domaines. Cette généralisation sera conditionnée par une caractérisation préalable des usages à l'aide d'analyses de corpus pilote ou de magicien d'Oz (Caelen *et al*., 1997). L'évaluation ne peut cependant pas se résumer à une étude purement linguistique, sous peine d'ignorer le caractère finalisé de la CHM orale. Elle n'a donc de sens qu'en complément d'approches globales de type ATIS.

D'autre part, l'évaluation doit être envisagée directement au niveau de la compréhension, sans nécessiter pour autant l'adoption d'un système de représentations communes. Une solution est de fonder l'évaluation non pas sur une référence en sortie, mais au contraire de définir un énoncé de contrôle qui tiend le rôle de référence en entrée. L'évaluation porte alors sur la comparaison des

structures sémantiques construites pour ces deux énoncés et ne fait intervenir que le système de représentation interne du système.

Cette approche a été adoptée (sous une forme relativement différente) dans le cadre du projet FRACAS (FRACAS, 1996) qui concernait la compréhension d'énoncés écrits. Cette méthodologie, qui est reprise dans le cadre de l'ARC "Compréhension de texte" de l'AUPELF-UREF (Sabatier, 1997), repose sur la définition de séries de tests spécifiques à un phénomène linguistique donné. Chaque test est constitué d'une déclaration D (énoncé à comprendre), d'une question fermée Q et d'une réponse attendue R). Voici un exemple d'évaluation de résolution d'anaphore (FRACAS, 1996) :

(D) *Peter is attending a meeting. He is to chair it.*

(Q) *Is Peter to chair a meeting ?*

(R) [Yes]

L'évaluation consiste alors à comparer la réponse fournie par le système avec la référence (R). L'originalité de cette méthodologie réside dans l'introduction de la question (Q), qui déplace l'objet de l'évaluation. En effet, il n'est plus nécessaire de comparer la structure sémantique construite à partir de (D) avec une référence prédéfinie. C'est au contraire la question (Q) qui impose au système une évaluation interne. En conséquence, la comparaison avec la réponse (R) est neutre vis à vis de la méthode utilisée par la compréhension.

Il n'en reste pas moins que la méthodologie DQR n'est pas transposable à la CHM orale. Les systèmes de dialogue sont en effet conçus pour comprendre les requêtes de l'usager, et non pas pour s'interroger sur leur propre comportement ! Les idées directrices de ce paradigme ont néanmoins inspiré la méthodologie DCR que nous allons présenter.

5. La méthodologie DCR

5.1. Principes

La méthodologie DCR repose sur la définition de séries de tests spécifiques pour chaque

phénomène propre à la CHM orale. Elle se focalise en particulier sur :

- **les inattendus propres à la parole spontanée** : hésitations, reprises, répétitions, etc.
- **la résolution des co-références** (anaphores, déictiques, ellipses) qui revêtent une grande importance en dialogue oral du fait de leur grande fréquence d'apparition.

Nous verrons ultérieurement que cette méthodologie peut également se substituer à une évaluation globale de type ATIS. Chaque test DCR se compose de trois éléments :

- la **demande (D)**, qui correspond à la requête de l'utilisateur,
- le **contrôle (C)**, qui correspond à un énoncé susceptible de comporter une information présente dans la requête (D).
- la **référence (R)**, issue de la comparaison de (D) et (C). Celle-ci est positive si les deux énoncés sont compatibles.

Il est essentiel de noter que **l'énoncé de contrôle (C) n'est pas une question posée au système sur son propre comportement**. Il s'agit au contraire d'une simple requête qui pourrait être prononcée par un utilisateur lambda et qui vise à contrôler la compréhension d'une information précise au sein de la demande (D). En d'autres termes, il s'agit d'une requête ordinaire (par exemple : *quel est le prix du trajet en seconde classe ?*) et non pas une question réflexive (par exemple : *Le tarif demandé par l'utilisateur concerne-t-il la seconde classe ?*).

Ainsi,

- La compréhension de (C) ne nécessite aucune modification du système.
- L'énoncé (C) est aussi simple que possible, de manière à être compris sans faute par le système.

A titre illustratif, reprenons l'exemple du tableau 1. Plusieurs tests peuvent être définis, qui vérifient la compréhension littérale des différents éléments de l'énoncé. On remarquera que ces tests peuvent être positifs ou négatifs :

(1) D *Quels sont les tarifs sur les vols Paris-Toulouse arrivant avant 15h ?*

C *Quelles sont les prestations ?*

R [Non]

(2) D *Quels sont les tarifs sur les vols Paris-Toulouse arrivant avant 15h ?*

C *Quels sont les tarifs* ?

R [Oui]

(3) D *Quels sont les tarifs sur les vols Paris-Toulouse arrivant avant 15h ?*

C *Quels sont les tarifs sur les Toulouse-Paris* ?

R [Non]

La procédure d'évaluation consiste alors en la succession de trois étapes :

Ì *Traitement séparé des énoncés (D) et (C)* — C'est à dire que pour le système, l'évaluation DCR se traduit par la compréhension d'énoncés **parfaitement indépendants**. De ce point de vue, les approches ATIS ou DCR sont équivalentes, le paradigme DCR nécessitant simplement un doublement des traitements pour chaque test.

Ò *Comparaison des représentations sémantiques (D) et (C)* — Cette comparaison revient à juger **en interne** si les représentations produites par le système sont compatibles ou non. Notons qu'un jugement négatif n'est pas synonyme de compréhension incorrecte, puisque que les énoncés (D) et (C) peuvent être initialement incompatibles (test négatif). Une comparaison avec la référence (R) est donc nécessaire.

Ó *Comparaison avec la référence (R)* — L'évaluation est positive si le jugement de compatibilité précédent est identique au résultat attendu (R).

Les résultats sont regroupés par série de tests. Ils fournissent alors un taux d'erreur pour chaque type de phénomène. On dispose ainsi d'un ensemble d'indices objectifs décrivant de manière discriminante le comportement du système.

(D) : *Quels sont les tarifs sur les vols Paris-Toulouse arrivant au plus tard à 15h ?*	(C) *Quel sont les tarifs* ?	Unification (D),(C)

type_requete : tarif iti.depart : Paris iti.arrivee : Toulouse h.depart : _ h.arrivee : < 15.00	type_requete : tarif iti.depart : _ iti.arrivee : _ h.depart : _ h.arrivee : _	oui

Figure 3 — Evaluation DCR : comparaison par unification des représentations sémantiques des énoncés (D) et (C). Représentation sous forme de cadre (*frame*) sémantique (Oerder & Aust, 1994).

La comparaison des structures sémantiques de (D) et (C) est propre au formalisme de représentation de chaque système. Cette comparaison en interne garantit la généricité de l'évaluation vis à vis des méthodes de compréhension. Il s'agit d'une opération immédiate reposant sur une simple unification de structure (figure 3). L'évaluation DCR ne nécessite donc le développement que d'un minimum d'outils spécifiques.

5.2. Evaluation multi-niveaux

Nous avons distingué deux niveaux d'évaluation correspondant respectivement aux deux étapes vues en introduction :

- **Compréhension littérale** — On considère ici des énoncés simples ne comportant aucune co-référence. Il s'agit donc d'énoncés indépendants du contexte — type A dans la terminologie ATIS (Pallett *et al.*, 1994). Les principales difficultés rencontrées à ce niveau sont la prise en compte des inattendus inhérents au caractère spontané du langage parlé. On traitera également les phénomènes de coordination complexe qui posent généralement problème aux systèmes de compréhension.

  Les tests (1) à (3) précédents correspondent à l'évaluation de la compréhension littérale.

- **Compréhension réelle** — A ce niveau, on considère avant tout la résolution des co-références. Cette résolution nécessite généralement une prise en compte du contexte dialogique (énoncés de type D dans la terminologie ATIS) qui sera fourni au système lors de l'évaluation. Là encore, la

demande (D) et l'énoncé de contrôle (C) sont traitées de manière indépendantes : on fait "rejouer" deux fois le même dialogue au système. Par exemple :

(4) Dialogue antérieur : *Pouvez-vous me donner les prochains Lyon - Lorient ?*

D Dialogue antérieur + *Ces vols sont-ils directs ?*

C Dialogue antérieur + *Les vols Lyon-Lorient sont-ils directs ?*

R [Oui]

6. Mise en oeuvre de la Méthodologie DCR

La définition de jeux de tests systématiques revêt une importance centrale dans la méthodologie DCR. Afin d'en illustrer la mise en oeuvre pratique, nous donnons dans ce paragraphe plusieurs exemples de tests DCR. Ces exemples sont extraits du corpus PARISCORP (Bonneau-Maynard & Devillers, 1998) réalisé dans le cadre de l'ARC "Dialogue Oral" de l'AUPELF-UREF.

6.1. Compréhension littérale

Les tests (1) à (3) concernaient la compréhension littérale dans le domaine ATIS. Les exemples suivants jouent le même rôle pour le renseignement touristique, cadre applicatif du corpus PARISCORP.

(5) D *Je voudrais connaître les restaurants les moins chers s'il vous plaît.*

C *Quels sont les restaurants les moins chers ?*

R [Oui]

(6) D *Je voudrais connaître les restaurants les moins chers s'il vous plaît.*

C *Quels sont les restaurants pas chers ?*

R [Non]

(7) D *Je voudrais connaître les restaurants les moins chers s'il vous plaît.*

C *Quels sont les tarifs des restaurants ?*

R [Non]

Tous ces exemples testent la compréhension du but central de la requête (D). Ils remplissent ainsi un objectif analogue à ceux d'une évaluation ATIS. On remarquera d'ailleurs que ces tests DCR s'interprètent dans l'approche ATIS. L'énoncé de contrôle (C) peut en effet être vue comme l'énoncé réellement compris par un système virtuel parfait (figure 4) : de même que dans l'approche ATIS, les références minimales et maximales correspondent aux réponses susceptibles d'être produites par un système idéal, dans la méthodologie DCR, l'énoncé de contrôle (C) correspond à l'information que pourrait comprendre un tel système travaillant sur la requête (D). Un test positif (R= [oui]) correspond alors à un comportement du système compatible avec celui de son compère virtuel. Dans cette interprétation, l'objectif du système évalué est donc d'être en accord uniquement avec les comportements virtuels pertinents.

Test	Référence (minimale et maximale) correspondant à (D)	Réponse du système virtuel : (R)
(6)	restaurant Café de la Gare	restaurant Brasserie Centrale Café de la Gare
(7)	restaurant Café de la Gare	restaurant prix Brasserie Centrale 75 F Café de la Gare 45 F Le Train Bleu 125 F

Figure 4 — Evaluation DCR et évaluation ATIS : exemple de références minimales dans la méthodologie ATIS correspondant aux tests DCR (6) et (7).

L'évaluation DCR permet néanmoins une analyse beaucoup plus instructive que dans l'approche ATIS. En particulier, les tests négatifs permettent un diagnostic très fin des erreurs du système. Le test (6) correspond ainsi à un non repérage du marqueur "*moins*", essentiel à la compréhension dans ce cas. Le test (7) correspond à une erreur plus grossière, l'adjectif "cher" n'ayant été interprété que comme marqueur de recherche de tarifs. S'il est utile de relever les énoncés mal compris par le système (tests positifs), il est encore plus intéressant de connaître les causes de ces erreurs (tests

négatifs). Le paradigme DCR permet ainsi un diagnostic proche comparable à celui d'une analyse manuelle de *logfiles*, avec l'avantage supplémentaire de systématiser et objectiver cette étude.

Les exemples suivants montrent enfin que le paradigme DCR s'applique aisément au traitement des inattendus de la parole spontanée tels que les corrections (8,9) et les répétitions (10) :

(8) D *Je voudrais connaître les restaurants les plus proches euh à la gare Saint Lazare.*

C *Quels sont les restaurants de la gare Saint Lazare?*

R [Oui]

(9) D *Je voudrais connaître les restaurants les plus proches euh à la gare Saint Lazare.*

C *Quels sont les restaurants proches de la gare Saint Lazare?*

R [Non]

(10) D *Je voudrais une chambre [...] pour trois nuits donc une chambre double avec douche*

C *Je voudrais une chambre double*

R [Oui]

6.2. Compréhension réelle

En termes de mise en oeuvre, ce niveau se distingue du précédent par l'introduction préalable d'un contexte discursif. Pour le reste, la constitution et l'analyse des tests DCR reste inchangée, comme le montrent les exemples suivants concernant des résolutions de références elliptiques (11), anaphorique (12) ou déictiques (13) :

(11) Dialogue antérieur : *En fait je cherche un hôtel*

D *Quels sont les moins chers ?*

C *Quels sont les restaurants les moins chers?*

R [Non]

(12) Dialogue antérieur : *Quel est le prix d'une chambre simple à l'hôtel Bellevue ?*

D *Comment je peux y aller ?.*

C *Comment aller à l'hôtel Bellevue ?*

R [Oui]

(13) Dialogue antérieur : *Oui, je voudrais une chambre dans l'hôtel d'Artois pour trois nuits donc une chambre double avec douche*

D *Comment je peux aller là-bas ?*

C *Comment aller à l'hôtel d'Artois ?*

R [Oui]

Ces exemples concernaient des co-références externes, c'est à dire pour lesquelles le référent se trouve dans un tour de parole antérieur. Dans le cas de co-références interne au tour de parole, aucun dialogue préalable n'est alors à rejouer.

Pour terminer, nous donnerons l'exemple d'une ellipse dont la résolution requiert une inférence de nature pragmatique, capacité essentielle dans le cadre d'une CHM finalisée.

(14) Dialogue antérieur : *Dans quelles salles passe la palme d'or du festival de Cannes ?*

D *Bon je prends le Gaumont.*

C *Je choisis le cinéma Gaumont*

R [Oui]

Ici, le système doit savoir associer *Gaumont* à un cinéma. Cette connaissance propre à l'application ne permet aucune généralisation à d'autres domaines. Cet exemple confirme néanmoins que la paradigme DCR est utilisable pour la mise en oeuvre de campagnes d'évaluation non génériques de type ATIS.

7. Conclusion

Dans cet article, nous insistons sur l'importance de la prédictivité et de la généricité en matière d'évaluation. Nous pensons avoir montré que l'objectif de prédictivité pouvait être raisonnablement atteint par l'approche DCR. Sur le second point, nous avons opéré une distinction entre généricité vis à vis de l'application et généricité vis à vis des traitements. Le paradigme DCR remplit

complètement ce dernier critère : il ne repose en effet sur aucun a priori quand à la nature des méthodes et représentations employées par le système.

Par contre, la question de la généricité vis à vis de l'application reste plus ouverte. En effet, une série de tests DCR définie pour une application donnée ne peut être utilisée directement sur un autre domaine. Cependant, si les tests ne sont pas portables, les résultats d'une campagne d'évaluation DCR restent généralisables dans une certaine mesure. Supposons par exemple qu'un système présente une robustesse appréciable en matière de résolution d'anaphores. Il sera alors un bon candidat pour un portage dans un autre domaine caractérisé par un usage très marqué de co-références anaphoriques.

En outre, on remarquera que le paradigme DCR ne repose sur aucun a priori quant à la nature de la tâche, ce qui n'est pas le cas de l'approche ATIS, comme nous l'avons noté au paragraphe 3.2.

Pour autant, le paradigme DCR comporte aussi des limites. Tout d'abord, nous avons évoqué le risque de biais méthodologique introduit par la compréhension de l'énoncé de contrôle (C). Comme nous l'avons montré, une définition rigoureuse de cette référence permet de surmonter ce problème. Pour être menée à bien, cette approche requiert un travail d'analyse sur corpus mené à la fois par des linguistes (pour caractériser les phénomènes de langues pertinents) et des chercheurs en CHM (pour identifier les cas intéressants du point de vue du système, mais aussi caractériser leurs erreurs éventuelles dans le cas des tests négatifs).

Afin d'autoriser une analyse fortement discriminante, l'évaluation DCR requiert d'autre part la définition d'un nombre important de séries de tests. Cette approche nécessite donc un investissement important, certainement comparable aux efforts fournis dans le cadre des campagnes DARPA-ATIS. Reste à savoir si celui-ci doit être interprété comme une limite, ou comme le prix à payer pour la mise en place d'une méthodologie réellement prédictive. L'ARC "Dialogue Oral" de l'AUPELF-UREF, où il a été décidé de mener de front une évaluation globale de type ATIS et une évaluation suivant le paradigme DCR donnera certainement des éléments de réponse à cette interrogation.

Remerciements

Les auteurs tiennent à remercier Jérôme Zeiliger (ICP, Grenoble, France), pour ses travaux antérieurs sur la méthodologie DCR, ainsi que Jacques Siroux (LLI-IRISA, Lannion, France), participant à l'ARC "Dialogue Oral" de l'AUPELF-UREF, pour ses commentaires sur la méthodologie DCR.

Références

Antoine J-Y. 1995. Conception de dessins et CHM : améliorer l'interaction orale au niveau linguistique, In : Caelen J. & Zreik K. (éds.) *Le Communicationnel pour concevoir*, Europia, Paris, France, 161-184.

Antoine J-Y. 1996. Parsing spontaneous speech without syntax. International Conference on Computational Linguistic, COLING'96, Copenhague, Danemark, 47-52.

Antoine J-Y. & Genthial D. 1999. Méthodes hybrides issues du TALN et du TAL Parlé : état des lieux et perspectives. In Antoine J-Y. & Genthial D. 1999. (éds.) *Atelier thématique Méthodes hybrides TALN / TALP*. TALN'1999, Cargèse, France, 1-17.

Aust H., Oerder M., Seide F., Steinbiss V. 1995. The Philips automatic train timetable information system, *Speech Communication*, 17, 249-262.

Baggia P., Kellner A., Perennou G., Popovici C., Sturm J., Wessel F. 1999 (à paraître). Language modelling and spoken dialogue systems: the ARISE experience. Eurospeech'99, Budapest, Hongrie, Septembre 1999.

Bernsen N.O., Dybkjaer H., Dybkjaer L. 1995. Exploring the limits of system-directed dialogue. Dialogue evaluation of the Danish Dialogue System. Eurospeech'95, Madrid, 1995.

Bernsen N.O., Dybkjaer L. 1997. The DISC Concerted Action. SALT workshop on Evaluation in Speech and Language Technology, Sheffield, Royaume-Uni, 35-42.

Blanche-Benveniste C., Bilger M., Rouget C., Van den Eynde K. 1990. Le français parlé : études grammaticales. CNRS, Paris, France.

Bonneau-Maynard H. & Devillers L. 1998. Acquisition, Transcription et annotation du corpus

PARISCORP, Rapp. Tech. AUPELF-UREF, ARC B2 Dialogue Oral.

Caelen J., Zeiliger J., Bessac M., Siroux J., Pérennou G. 1997. Les corpus pour l'évaluation du dialogue homme-machine, 1ères Journées Scientifiques et Techniques du réseau Francil, JST-FRANCIL'97, Avignon, France, 215-223.

De Mori R. 1994. Apprentissage automatique pour l'interprétation sémantique. Journées d'Etudes de la Parole, JEP'94, Trégastel, France, 11-19.

FRACAS consortium. 1996. Using the framework. Fracas Project LRE 62-051, Deliverable D16, chapitre 3.

Gauvain J-L., Bennacef S., Devillers L., Lamel L., Rosset S. 1997. Spoken Language component of the MASK kiosk. In :Varghese K. & Pfleger S (éds*.) Human comfort and security in information systems*, Springer Verlag, Berlin, RFA, 93-103.

Siroux J., Guyomard M., Jolly Y., Multon F. Remondeau C. 1995. Speech and tactile-based GEORAL system. Eurospeech'97, Madrird, Espagne, 1943-1946.

Hirschman L. 1998. Language understanding evaluations: lessons learned from MUC and ATIS. 1st International Conference on Language Ressource and Evaluation, LREC'98, Granada, Espagne, 117-122.

Lamel L., Rosset S., Bennacef S., Bonneau-Maynard H., Devillers L., Gauvain J-L. 1995. Development of spoken language corpora for travel information. Eurospeech'95, Madrid, Espagne, 1961-1964.

Letellier-Zarshenas S., Nicolas P., Goulian J., Antoine J-Y. 1999 (à paraître). Inattendus structurels et communication orale finalisée : influence de la tâche et du contexte interactif, Journées Internationales de Linguistique Appliquée, JILA'99, Nice, France.

Lokbani M. N. & White S. 1998. La reconnaissance de la parole. *La Recherche*, 319, 82.

Minker W. 1998. Evaluation methodologies for interactive speech systems. 1st International Conference on Language Ressource and Evaluation, LREC'98, Granada, Espagne, 199-206.

Minker W. & Bennacef S. 1996. Compréhension et évaluation dans le domaine ATIS. Journées d'Etudes de la Parole, JEP'96, Avignon, France, 417-421.

Oerder M. & Aust H. 1994. A realtime prototype of an automatic inquiry system, International Conference on Spoken Language Processing, ICSLP'94, Yokohama, Japon, 703-706.

Pallett D.S, *et al*. 1994. Benchmark tests for the ARPA Spoken Language Program. ARPA Workshop on Spoken Language Technology, 5-36.

Pieraccini R. & Levin E. 1992. Stochastic representation of semantic structures for speech understanding. *Speech Communication*, 11, 283-288.

Pieraccini R, Levin E., Eckert W. 1997. AMICA: the AT&T mixed initiative conversational architecture. Eurospeech'97, Rhodes, Grèce, 1875-1878.

Polifroni J., Seneff S., Glass J., Hazen T.J. 1998. Evaluation methodology for a telephone-based conversational system. 1st International Conference on Language Ressource and Evaluation, LREC'98, Granada, Espagne, 43-49.

Sabatier P. 1997 Evaluer les systèmes de compréhension de textes, actes 1ères Journées Scientifiques et Techniques du réseau Francil, JST-FRANCIL'97, Avignon, France, 223-226.

Seneff S. 1992. TINA, a natural language system for spoken language applications. *Computational Linguistics*, 18(1), 61-86.

Walker M.A, Litman D.J, Kamm C.A, Abella A. 1997. PARADISE: a framework for evaluating spoken dialogue agents. 35th meeting of the ACL/EACL, Madrid, Espagne, 271-280.