# Vers une Méthodologie d'Evaluation Qualitative des Systèmes de Compréhension et de Dialogue Oral Homme-machine

**Jérôme Zeiliger**
Institut de la Communication Parlée
INPG
46 Av. Félix Viallet
38031 Grenoble Cedex
Email : zeiliger@icp.grenet.fr

**Jean Caelen**
CLIPS - IMAG
Domaine Universitaire
BP 53
38041 Grenoble Cedex 9
Email : Jean.Caelen@imag.fr

**Jean-Yves Antoine**
VALORIA
Université de Bretagne Sud
CUS de Vannes
1 r. de la Loi, 56000 Vannes
Email : antoine@univ-ubs.fr

**Résumé**

Cet article présente une méthodologie d'évaluation pour les systèmes de compréhension et de dialogue oral homme-machine. Placée dans le prolongement de travaux récents en TALN, cette méthodologie est basée sur la définition et l'emploi de batteries de tests DQR (Donnée, Question, Réponse). Ces tests sont adaptés aux particularités du langage oral et aux situations de dialogue, et visent à vérifier la prise en compte individuelle par n'importe quel système de chacun des phénomènes linguistiques répertoriés. L'évaluation ainsi obtenue se définit comme qualitative, générique et portable.

## 1. Problématique

La communication orale homme-machine a atteint une maturité qui laisse espérer le développement futur de systèmes opérationnels en conditions réelles. En particulier, les problèmes centraux du domaine sont désormais bien identifiés, et certaines réalisations dépassant le simple cadre du prototype de laboratoire présentent des performances encourageantes. L'ensemble des traitements automatiques impliqués dans la communication orale (reconnaissance et synthèse de parole, compréhension du langage parlé, modélisation du dialogue) ont ainsi connu des progrès significatifs au cours de ces dernières années. Pour capitaliser ces avancées et orienter les recherches futures (Cole *et al*, 1995), la mise en place de procédures d'évaluation adaptées au dialogue oral constitue un enjeu central pour la communication parole.

Le recours à l'évaluation est une pratique déjà bien établie dans le domaine du traitement automatique du langage écrit (TALN). Des programmes tels que TSNLP[1] ont ainsi permis l'élaboration et la validation de méthodologies fouillées d'évaluation (Lehmann *et al*, 1996). Reposant sur la définition de jeux de test très détaillés, ces procédures permettent d'étudier le comportement des systèmes sur chaque phénomène linguistique bien précis.

A l'opposé, les systèmes de reconnaissance de la parole ou de dialogue oral ont jusqu'ici été principalement évalués en termes de performances globales (programme ATIS[2] de la DARPA). Ce type d'évaluation permet avant tout de mesurer le chemin qui nous sépare d'un dialogue oral en conditions réelles. Il reste cependant à étendre cette méthodologie afin d'atteindre un diagnostic plus précis et donc plus riche en enseignements. On ne saurait ignorer les recherches menées sur ce sujet en TALN (Estival *et al*, 1994). C'est pourquoi nous proposons dans cet article une méthodologie d'évaluation inspirée des réflexions de l'ARC ILEC A4[3] (*Compréhension de textes*) et adaptée à la spécificité de la langue et de la communication orale.

Dans un premier temps, nous allons présenter les objectifs que nous assignons à la méthodologie proposée. Nous décrirons ensuite les procédures d'évaluation employées dans le cadre du TALN, pour ensuite discuter de leur adaptation aux spécificités de la modalité orale. A l'aide de multiples exemples, nous détaillerons enfin la mise en oeuvre pratique de notre méthodologie, tant pour l'évaluation de la compréhension de la parole que pour celle du dialogue oral.

## 2. Pour une Evaluation Générique et Qualitative

L'évaluation de systèmes requiert des critères objectifs car les critères subjectifs sont trop dépendants de l'état mental de l'observateur même expert. Mais "objectivité" en matière d'évaluation de dialogue a le plus souvent rimé jusqu'ici avec "quantitatif" donc "métriques" (Cole *et al*, 1996 ; EAGLES, 1997). En même temps des notions propres au dialogue, comme la redondance, la coopérativité, l'à-propos... sont des concepts non pas absolus mais "dégradables" (Jokinen, 1996), donc difficiles à mesurer par des moyens quantitatifs. D'où le recours fréquent à une évaluation par des usagers - sous forme d'enquêtes (Bernsen *et al*, 1995 ; Lamel *et al*, 1995) - qui fait retomber l'évaluation dans les mêmes travers.

La méthodologie proposée ici revêt deux caractéristiques essentielles à nos yeux que l'on ne retrouve pas dans les paradigmes d'évaluation usuellement employés en dialogue oral :

- **généricité** — Les systèmes de dialogue oral reposent généralement sur des théories et des domaines d'application différents. Ainsi, les systèmes vocaux centrés sur l'interrogation de bases de données

[1] TSNLP (Test Suites for Natural Language Processing) est un projet du programme LRE (Language Research and Engeenering) de la Communauté Européenne.

[2] ATIS : Air Transport Information Systems

[3] Les Actions de Recherche Concertée sont une initiative de l'AUPELF-UREF, notamment les thèmes ILEC (Informatique, Linguistique et Corpus Ecrits) et ILOR (Informatique, Linguistique et Corpus Oraux).

(renseignement de type ATIS) mettent généralement en jeu une compréhension reposant sur la recherche d'îlots-clefs (De Mori, 1994 ; Minker & Bennacef, 1996) ou de segments conceptuels (Pérennou, 1996). A l'opposé, les contextes applicatifs plus complexes tels que le dessin ou la conception assistée par ordinateur nécessitent une compréhension sensiblement plus fouillée, se rapprochant du niveau de détail exigé à l'écrit (Antoine, 1996a, 1996b). Cette diversité explique les problèmes généralement rencontrés lors des campagnes d'évaluation pour définir une plate-forme commune (portage des systèmes vers une même application, adoption de modes de représentation communs, etc.). Il est donc nécessaire de définir un paradigme générique d'évaluation, qui ne pose aucun a priori sur les représentations utilisées par les systèmes, ni sur les modèles de langage et de dialogue mis en jeu dans chaque contexte applicatif.

On notera que cette généricité est une garantie de réutilisabilité et d'évolutivité des procédures d'évaluation. A la différence des paradigmes utilisés à l'heure actuelle très dépendants des représentations sémantiques manipulées par les systèmes.

- **qualitativité** — Le langage parlé est un objet complexe qui fait intervenir de nombreux niveaux de traitements généralement interdépendants, aux frontières incertaines, et liés au domaine d'application considéré. Comme nous l'avons déjà laissé entendre, une évaluation purement quantitative des performances n'a ainsi qu'un intérêt limité, puisqu'elle ne conduit à aucun diagnostic prédictif sur le comportement de chaque niveau de traitement. Quels composants doivent-ils être crédités des bonnes performances globales d'un système: la reconnaissance de la parole, les niveaux de compréhension ou encore le module de dialogue ? Ce raisonnement peut s'étendre à toute évaluation quantitative portant sur un niveau précis de traitement. Par exemple, quels enseignements tirer d'un taux global de 90% de bonne compréhension quant à l'amélioration future du système ? Et surtout, quelle assurance avons-nous que les méthodes employées par le système, qui sont robustes dans le contexte de l'application, le seront pour d'autres formes de CHM orale ? Il n'est qu'à observer les différences structurelles de langage entre plusieurs contextes applicatifs pour en douter (Antoine, 1995).

On voit maintenant l'intérêt d'une approche qualitative, qui consiste à évaluer les systèmes sur des phénomènes linguistiques ou dialogiques bien identifiés: on disposera alors d'un diagnostic détaillé du comportement du système dans des situations bien définies.
Enfin, on notera que l'évaluation des systèmes de dialogue n'a de sens qu'inscrite dans la perspective d'une communication naturelle. D'où l'intérêt de corpus pilotes tel celui réalisé dans le cadre de l'ARC ILOR B2 (*Dialogue oral*) pour recenser les phénomènes linguistiques observés en situation, à travers une analyse d'usage (Caelen *et al*, 1997).

## 3. Harmoniser les Approches : l'Evaluation en TALN

L'évaluation a déjà une longue histoire dans le domaine du traitement du langage écrit (TALN). Ainsi en Europe le consortium du projet FRACAS (Framework for Computational Semantics) a cherché à "harmoniser les approches" et jeté les bases d'un cadre général à toutes les questions de sémantique computationnelle. Quels que soient les systèmes de traitement du langage considérés (traduction automatique, compréhension de message, "information retrieval", systèmes de dialogue H-M., etc.), les théories sémantiques qui les sous-tendent doivent, selon ces auteurs, répondre d'un certain nombre de phénomènes linguistiques de base qu'il doit être possible de recenser. Ils ont commencé d'en dresser la liste (FRACAS, 1994), et ont proposé une méthodologie de test pour vérifier la prise en compte de chaque phénomène par un système (FRACAS, 1996). Ces travaux ont inspiré par exemple les évaluateurs francophones dans le domaine de la "compréhension de textes" (Rolbert & Sabatier, 1996) qui, ayant passé en revue les différentes formes de test, retiennent une méthodologie d'évaluation de type "boîte noire" et reposant sur des tests génériques DQR. Ils distinguent en effet:
• le type QR (Q pour *question* + R pour *réponse*) : qui est adapté à l'évaluation de systèmes d'interrogation en langage naturel de bases de connaissances structurées. "*Il implique de fournir aux compétiteurs la même base de données à interroger, le modèle conceptuel explicite associé (individus, objets, relations, etc.), ainsi que le lexique sur la base duquel les questions seront formulées. Dans un second temps, il faut fournir aux testeurs un ensemble de questions avec les réponses attendues*".
• le type DF (D pour *déclaration* + F pour *représentation Formelle*): o˘ "*la représentation formelle (F) est une représentation prédéfinie (liste annotée, tableau, etc.) où sont exprimés les résultats de tâches particulières considérées comme liées à la compréhension du langage naturel*". Mais, notent-ils, faisant l'impasse sur la Question et la Réponse évalue-t-on vraiment des systèmes de compréhension ?
• le type DQR (D pour *déclarations*, Q pour *question* et R pour *réponse*).qui est donc considéré comme le meilleur. Car "*Tout système se prêtant à des tests DQR peut être considéré comme un système complet de compréhension du langage naturel, c.-à-d. un système qui analyse (D et Q) et qui synthétise (R) du langage naturel*". Dans chaque test, D est l'ensemble de données (énoncé(s) à comprendre) contenant les informations suffisantes pour répondre à une question Q testant un phénomène précis. La réponse attendue R, sur laquelle se base l'évaluation, est dans l'ensemble : *oui* / *non* / *ne sais pas*.
Tel l'exemple suivant: (Extrait de: Ellipsis, Gapping, cross-sentential gapping (FRACAS D16 ).

*John went to Paris by car.*
*Bill by train.*
---
*Did Bill go to Paris by train ?*
*[Yes]*

Le projet TSNLP également, a bâti des jeux de phrases-test pour l'évaluation des applications en TALN, avec un souci de portabilité et de réutilisabilité (Lehman *et al*,

1996 ; Fouvry & Balkan, 1996). Chaque test est soigneusement annoté et réalisé en plusieurs langues, le tout venant nourrir une base de données qui s'enrichit progressivement.
L'intérêt principal de ces approches est qu'elles obligent les concepteurs de test à avoir une très bonne connaissance des phénomènes à tester, et que les tests ainsi produits sont relativement simples à mettre en oeuvre.
Ce paradigme permet une évaluation qualitative car indifférente aux méthodes et théories utilisées, globale (pas de désaccord sur les résultats des sous-systèmes) et centrée sur les problèmes proprement linguistiques communs aux divers types d'application. Nous proposons de l'appliquer à l'évaluation de la compréhension de l'oral comme du dialogue en procédant par complexification progressive des ensembles D dans DQ, passant ainsi de la simple requête à l'enchaînement de un ou plusieurs tours de dialogue.
Cette évaluation plus générique sera aussi moins "coûteuse" et gagnera en portabilité: pas de bases de connaissances nouvelles à intégrer dans le système (D est autosuffisant), pas de modèle d'application ou de domaine à modifier, pas de représentations ou formalismes communs à adopter. Pas non plus de représentations de référence forgées par des experts, et le dépouillement des résultats pour produire des scores en sera simplifié. Le travail se concentre sur la mise en évidence des phénomènes qu'on veut tester et la construction d'exemples génériques qui permettent de le faire, d'où une meilleure compréhension des phénomènes dialogiques. L'évaluation elle-même se "limitant" à la constitution d'un corpus de test particulier, sorte d'implémentation locale du test générique pour un domaine d'application plutôt qu'un autre.

## 4. Application de la Méthodologie DQR à l'Evaluation du Dialogue Oral

La démarche consistera donc ici à:
⇒ définir l'ensemble des phénomènes linguistiques qu'un système de dialogue oral devrait savoir traiter
⇒ se focaliser sur les phénomènes particuliers à l'oral et au dialogue
⇒ en choisir un sous-ensemble représentatif et proposer des exemples de test pour chacun des phénomènes.

Pour le premier point on se référera au travail des projets cités ci-dessus qui sans prétendre à l'exhaustivité ont déjà jeté de bonnes bases.
Les spécificités du langage oral, au sens de la parole spontanée et non lue, ont trait (Blanche-Benveniste *et al*, 1990):

- à l'élocution: hésitation, bruits d'élocution, interruptions d'interlocuteurs potentiels, recouvrement de voix, spécificités acoustiques de l'environnement... Ces caractéristiques jouent un rôle surtout dans la phase de reconnaissance de la parole, mais, même s'il faut le remettre à plus tard, leur rôle sémantique serait également à prendre en compte. Les phénomènes de reprise (et de réparation induite) tiennent également une place importante: reprise avec effacement, avec ajout, avec répétition. Les épellations sont également propres à l'oral.
- au linguistique, dès le niveau morpho-phonologique (" il y a " est prononcé " y'a "), puis le niveau lexical (mots d'argot, emprunts, mots étrangers, etc.) jusqu'au niveau syntaxico-sémantique avec, par exemple, la déstructuration de la syntaxe dans les questions: suppression des mots interrogatifs, renversement de l'ordre, mises en opposition, en apposition, répétitions, etc.
- au niveau pragma-linguistique pour ce qui concerne les référents au monde (marqués à l'oral surtout par les déictiques comme " prends ce cercle "), les référents situationnels (objets visibles et non-visibles, comme " prends le rouge ", spatialité " je suis ici ", sujet parlant " c'est moi qui te parle "), et les référents d'arrière-plan (le quai A est souvent à côté du quai B, l'heure d'arrivée ne peut précéder l'heure de départ encore qu'il peut s'agir de deux jours différents ou d'un changement de fuseau horaire...)
- au dialogique: anaphores, ellipses, reprise d'une partie de la question précédente (" puisque tu veux savoir où j'étais hier, eh, bien oui, j'étais à Paris ") marques d'articulation du dialogue (" donc maintenant, je voudrais aller au cinéma ") et signes de compréhension ou d'acquiescement, voire de relance (" oui, oui, j'écoute "), marqueurs rhétoriques et argumentatifs (connecteurs, ligateurs, présentatifs, phatiques, etc., " comme le dit si bien Boileau dans son discours sur la persuasion, je pense qu'il faudrait que l'on considère ce problème dans toute son étendue rhétorique ") dans la mesure où ces phénomènes sont plus fréquents à l'oral qu'à l'écrit, et jouent sur plusieurs degrés du fait des tours de parole.

Il conviendra donc de s'inspirer de dialogues réels spontanés pour y rechercher les phénomènes linguistiques importants, et pour les illustrer à partir d'exemples véridiques dans les tests DQR.
Enfin la notion de dialogue implique qu'il y ait au moins deux interlocuteurs, et donc au moins deux répliques de préférence.

## 5. Une Méthodologie Multiniveaux pour une Evaluation Qualitative

Afin d'étudier les systèmes de dialogue sur des phénomènes linguistiques et dialogiques bien précis, nous avons défini sept niveaux d'évaluation relativement indépendants. Nous allons présenter brièvement ces différents niveaux d'évaluation. Le paragraphe suivant, comportant plusieurs exemples de tests DQR pour chaque niveau, sera l'occasion d'une présentation plus détaillée.
Les trois premiers niveaux concernent l'étape de compréhension de la parole et ne mettent en jeu aucune information dialogique. La phrase de donnée D correspond ainsi à une[4] requête utilisateur. L'objectif est de vérifier si le système a extrait une représentation sémantique correcte de l'énoncé, indépendamment du formalisme adopté. Chaque question Q fait appel à la caractérisation d'une relation sémantique[5] précise au sein de l'énoncé D. Les trois niveaux d'évaluation mettent en jeu des processus de complexité croissante pour la récupération de cette information sémantique :

[4] Ou éventuellement plusieurs, comme nous le verrons dans les exemples du paragraphe 6.
[5] Suivant le formalisme adopté, on parlera également de rôle thématique ou de cas sémantique (Minker, 1996)

- *Information explicite* (niveau 1) — Repérage d'une information explicitée dans l'énoncé.
  Ce niveau concerne la compréhension d'énoncés simples ne comportant aucune ellipse ou anaphore. A ce niveau, les principales difficultés de traitement concernent la prise en compte robuste de la variabilité structurelle du langage spontané (hésitations, répétitions, corrections, dislocations, etc.) ainsi que sa richesse informationnelle, caractéristique largement ignorée dans les procédures d'évaluation actuelles[6].
- *Information implicite* (niveau 2) — Résolution des références implicites à une information.
  L'objet de l'évaluation est ici la résolution des anaphores, des ellipses, et des incomplétudes, phénomènes omniprésents en dialogue oral spontané. A ce niveau, on reste dans le cadre d'ellipses et d'anaphores littérales, c'est à dire récupérables à un niveau structurel (syntaxe ou sémantique). Les constructions implicites faisant appel à des traitements pragmatiques ou dialogiques font l'objet d'une évaluation au niveau suivant.
- *Inférence* (niveau 3) — Construction du sens complet de l'énoncé.
  On s'intéresse à ce niveau à la construction du sens complet (ou sens réel (Pérennou, 1996) de l'énoncé, c'est à dire replacé dans son contexte pragmatique et dialogique. Le passage du sens littéral au sens complet nécessite l'intervention de processus inférentiels afin de récupérer la part de sous-entendu qui est généralement présente dans les requêtes de l'utilisateur. Ces sous-entendus dépassent largement le cadre des références anaphoriques et elliptiques. On distinguera les résolutions faisant appel à un raisonnement de sens commun de celles faisant intervenir des inférences pragmatiques.

Les deux niveaux suivants concernent le dialogue. On peut se placer à divers degrés: au niveau de l'échange (2 tours de parole) ou au niveau de la transaction complète (de but posé à but atteint ou abandonné ou atteint et satisfait). A chacun de ces niveaux on peut se placer soit du point de vue de la machine (ou allocutaire), c'est-à-dire en entrée du système, soit du point de vue de l'utilisateur (ou locuteur), c'est-à-dire en sortie du système. Dans le premier on adresse les niveaux 4 et 5, dans le second cas on adresse les niveaux 6 et 7.

- *Interprétation du type d'acte illocutoire* (niveau 4) — On évalue ici si une demande, même indirecte, a été bien interprétée comme telle, si une confirmation, une contestation, une assertion, etc. ont été bien reconnues. Cette interprétation est en fait une identification du *but illocutoire* de l'acte en cours. La portée de cette évaluation est donc de l'ordre de la réplique (intention en action chez Searle).
- *Reconnaissance des intentions (ou du but)* (niveau 5) — On s'intéresse à ce niveau à l'échange ou à la transaction pour évaluer la reconnaissance de buts plus profonds (intention préalable chez Searle). Il s'agira par exemple d'identifier rapidement que tel client ne souhaite pas prendre son petit-déjeuner à l'hôtel, cela évitera plus tard de lui parler des prix dudit petit-déjeuner.
- *Pertinence de la réponse* (niveau 6) — On évalue ici la réponse du système suite à chaque requête utilisateur, en progressant au cours du dialogue. Cette pertinence est à pondérer en fonction des connaissances de la machine, de ses capacités linguistiques et du type d'utilisateur.
- *Pertinence de la stratégie* (niveau 7) — On évalue ici la réponse du système à la fin d'un échange ou à la fin du dialogue : la transaction est-elle réussie ? A-t-elle été efficacement menée ?

## 6. Mise en Oeuvre de la Méthodologie DQR pour l'Evaluation du Dialogue Oral

Afin d'illustrer la mise en oeuvre pratique de la méthodologie proposée, nous donnons dans ce paragraphe plusieurs exemples de tests DQR par niveau d'évaluation. Nous allons tout d'abord relever les propriétés caractéristiques des jeux de tests que nous nous proposons de réaliser, afin de bien marquer les spécificités de cette méthodologie par rapport aux procédures d'évaluation employées tant à l'oral qu'à l'écrit.

### 6.1. Spécificités de la méthodologie proposée

Avant toute chose, il convient de rappeler que cette méthodologie concerne l'évaluation de systèmes de dialogue oral finalisé. Dans ce cadre, l'accent est mis sur la robustesse des processus de compréhension et sur la pertinence de la stratégie de dialogue. Il ne s'agit pas de vérifier in extenso que le système a tiré tout ce qu'il y avait à conclure de l'échange, mais seulement ce qui est pertinent pour établir sa compréhension des phénomènes dialogiques et de la marche du dialogue en cours. Ainsi, la recherche d'une grande couverture linguistique, tout en étant un prérequis indispensable à une bonne compréhension pour des contextes applicatifs évolués (CAO par exemple), ne constitue pas un objectif d'évaluation en soi. En particulier, on ne retrouvera pas ici l'intérêt marqué par la communauté TALN pour une évaluation centrée sur la syntaxe (phénomènes d'accord, de complémentation etc.).

Parallèlement, la méthodologie d'évaluation doit prendre en compte, d'une part le coût de la reconnaissance de parole, d'autre part la spécificité des modèles de langages utilisés par la plupart des systèmes de dialogue : alors que les systèmes de compréhension de l'écrit visent une grande couverture linguistique, ceux utilisés à l'oral se concentrent sur la modélisation des requêtes de l'utilisateur. En pratique, il est donc nécessaire de simplifier au maximum la question Q pour la centrer uniquement sur le phénomène que l'on désire tester. Cette contrainte garantira la généricité de l'évaluation et évitera que ce ne soient les difficultés du système à comprendre la question Q — et non la requête D — qui soient évaluées ! Par ailleurs, la prise en compte dans un jeu unique de test DQR de plusieurs phénomènes est susceptible d'introduire des biais de même nature. On n'y aura donc pas recours dans un premier temps.

Enfin, on notera l'emploi de données D positives (R = NON) et négatives (R = OUI) dans les jeux de tests proposés. Cette approche, absente des procédures utilisées à l'oral, doit aider à l'établissement d'un diagnostic utile pour chaque phénomène. Plutôt que de s'en remettre à une évaluation de type tout ou rien, un choix pertinent de données négatives peut en effet conduire à une

[6] Citons par exemple les phénomènes de coordinations non triviales et, pour le niveau suivant, d'anaphore plurielles.

caractérisation des causes des échecs subis par le système. Plusieurs illustrations seront données, de ce point de vue, dans les exemples ci-dessous.

## 6.2. Niveau information explicite

On rappelle que l'objectif est ici le repérage d'une information explicitée dans l'énoncé. Celle-ci est traduite sous la forme d'une relation sémantique sur laquelle se concentre la question Q :

(1) D *Ce serait pour partir demain pour Vannes*
Q *Aller à Vannes ?*
R *Oui*

Dans cet exemple, on cherche à retrouver le cas Destination *Vannes*. On remarquera l'extrême simplicité de la question, qui ne doit poser aucun problème de compréhension à un système de dialogue. L'information recherchée correspondait ici à un élément clef de la requête. Ce niveau de détail est largement suffisant pour des applications de type ATIS. D'autres cadre applicatifs plus complexes, peuvent cependant nécessiter l'extraction d'informations de plus bas niveaux (arguments d'arguments) :

(2) D *Vous prenez à droite après les bâtiments blancs aux volets bleus*
Q *Volets bleus ?*
R *Oui*
(3) D *Vous prenez à droite après les bâtiments blancs aux volets bleus*
Q *Bâtiments aux volets bleus ?*
R *Oui*

La relation d'attribution entre le substantif *volets* et la couleur *bleu* se situe à un niveau de profondeur 3 dans la structure sémantique de l'énoncé. Il s'agit donc d'une information de second plan dont la compréhension peut cependant être cruciale dans certains contextes. On regrettera ainsi que les procédures d'évaluation traditionnelles ignorent largement cette richesse informationnelle de la communication orale.
On remarquera par ailleurs que les tests (2) et (3) permettent un diagnostic très progressif : rattachement de l'adjectif attribut *bleus* dans le cas de la question (2), rattachement du groupe prépositionnel *aux volets bleus* dans le cas de la question (3). Enfin, l'utilisation de données négatives telles que (4) permet de détecter la cause d'éventuelles erreurs :

(4) D *Vous prenez à droite après les bâtiments blancs aux volets bleus*
Q *Bâtiments bleus ?*
R *Non*

Ici, en cas d'erreur (réponse : OUI), on a affaire à un mauvais rattachement de l'adjectif attribut.
Nous avons noté qu'une des difficultés principales des systèmes à ce niveau consiste en la prise en compte des inattendus structurels de la parole spontanée. Hésitations, répétitions, reprises, corrections, dislocations peuvent être évaluées sans difficultés à l'aide de tests DQR. Par exemple :

(5) D *Je voudrais partir demain non après-demain*
Q *Partir demain ?*
R *Non*
(6) D *Je voudrais partir demain non après-demain*
Q *Partir après-demain ?*
R *Oui*

On retrouve sur cet exemple de correction la progressivité de la méthodologie : détection de la correction dans le cas de la question (5) et test de la résolution de cette dernière dans le cas de la question (6).

## 6.3. Niveau information implicite

A ce niveau, on s'intéresse essentiellement à la résolution des références anaphoriques et elliptiques qui peuvent être traitées sans intervention du contexte pragmatico-dialogique :
On rappelle que l'objectif est ici le repérage d'une information explicitée dans l'énoncé. Celle-ci est traduite sous la forme d'une relation sémantique sur laquelle se concentre la question Q :

(7) D *Vous prenez la rue à droite et vous la suivez sur 300 m*
Q *Suivre la ?*
R *Oui*
(8) D *Vous prenez la rue à droite et vous la suivez sur 300 m*
Q *Suivre rue à droite ?*
R *Oui*
(9) D *Vous prenez la rue à droite et vous la suivez sur 300 m*
Q *Suivre à droite ?*
R *Non*

Les exemples précédents correspondent à un cas de résolution d'anaphore pronominale. La question (7) permet de vérifier le rattachement du pronom la au bon rôle sémantique, tandis que la question (8) s'assure de la bonne résolution de la référence anaphorique. En cas d'erreur, la donnée négative (9) permet enfin d'identifier un mauvais rattachement de référent.

(10) D *Donnez moi un billet pour Paris et aussi pour Lyon*
Q *Billet pour Lyon ?*
R *Oui*
(11) D *Donnez moi un billet pour Paris et aussi pour Lyon*
Q *Billet de Lyon vers Paris ?*
R *Non*

Le test (10) correspond à la résolution de l'ellipse du verbe *donner* par l'adverbe *aussi*. La donnée négative permet de tester un cas, parmi d'autres, de mauvaise résolution. Les ellipses non marquées peuvent faire de même l'objet d'une évaluation.
Les exemples précédents correspondaient à des anaphores et des ellipses internes à la structure de phrase. Il arrive cependant fréquemment que le référent recherché se trouve dans un énoncé prononcé auparavant. Il est donc impératif de construire des jeux de tests portant sur de multiples énoncés pouvant couvrir plusieurs tours de parole :

(12) D *Vous prenez la première rue qui se présente*
D *A droite ou à gauche ?*
D *Celle de droite*
Q *Rue de droite ?*

R *Oui*

Les tests sont alors de la forme $D^n$ QR, où n correspond à la portée de l'anaphore (resp. de l'ellipse) en terme de nombre de tours de paroles. On prendra cependant garde de ne pas interpréter ces tests comme des tours de parole réels entre le système et l'utilisateur. Il n'est en effet pas souhaitable d'introduire la réponse du système à la première requête D comme donnée suivante. Ce serait en effet introduire le comportement dialogique du système (évalué aux niveaux 4 et 5) dès cette étape. La succession des données D correspondra donc à un dialogue fictif, la tâche du système évalué étant d'intégrer convenablement ces informations — comme il le ferait en situation réelle de dialogue — pour pouvoir répondre correctement à la question. Ici encore, les données D correspondant aux interventions du système fictif doivent être aussi succinctes que possible, puisque l'objectif n'est pas la compréhension des réponses du système de dialogue !

## 6.4. Niveau inférence

Dans le cadre d'une CHM finalisée, les inférences de nature pragmatique sont bien sûr prééminentes au cours du dialogue. On donnera donc trois exemples de raisonnement pragmatique :

(13) D *Bonjour, je voudrais un aller-retour pour Paris*
Q *Vouloir billet ?*
R *Oui*
(14) D *Bonjour, je voudrais un aller-retour pour Paris*
Q *Aller-retour Paris-Grenoble ?*
R *Oui*
(15) D *Je voudrais une chambre avec bain*
D *Ah, et une avec douche pour mon collègue*
Q *Chambre avec douche ?*
R *Oui*

Les tests (13) et (14) correspondent à une inférence pragmatique sans ellipse ni anaphore. Dans le premier cas, le système de compréhension doit associer le concept d'*aller-retour* avec celui de *billet*, tandis que dans le second cas, c'est la connaissance de la localisation de la borne de réservation (gare de Grenoble) qui lui permet de répondre par l'affirmative[7]. La donnée (15) comprend une anaphore pronominale qui peut être éventuellement résolue à un niveau strictement structurel. La connaissance du monde de l'application (dans notre exemple : *une chambre peut comporter un bain, une douche ou un cabinet de toilette*) facilite néanmoins la tâche du système.
On peut de même définir des jeux de tests pour des inférences de sens commun. Par exemple:

(16) D *Pierre se rend à un meeting.*
*Il doit le présider.*
Q *Est-ce que Pierre doit présider un meeting ?*
R *Oui*

Où la connaissance que "un meeting ça se préside" ne faisant pas a priori partie du monde de l'application, relève du sens commun et vient en renfort pour répondre à la question; alors qu'une résolution au niveau structurel ne pourrait guère aboutir que par défaut.

## 6.5 Préalable aux niveaux de dialogue (4 à 7)

Dès que l'on envisage l'évaluation du dialogue en tant que tel, la difficulté se trouve intrinsèquement augmentée. En effet tout dialogue suppose au moins deux interlocuteurs, et pour prendre corps au moins deux tours de parole (un pour chaque interlocuteur) c'est-à-dire un échange. Ainsi là où la compréhension pouvait se ramener à un problème de représentation de données ou de connaissances, le dialogue doit les englober dans une représentation d'une situation de communication, celle où des tours de parole existent, où les signaux viennent d'un interlocuteur ou de l'autre. Quelle que soit la forme que prenne l'évaluation, dans la mesure où elle n'est pas observation humaine, elle implique que le système "reconnaisse" et attribue à chacun des interlocuteurs ce qui lui revient, ce qui constitue le fondement d'une "méta-connaissance" car non-extraite des énoncés des interlocuteurs. Toute question se rapportant ensuite à des notions de but illocutoire, d'intention, de pertinence, de stratégie, peut se formuler par : "Quel est le but illocutoire de cette réplique ?", "Est-ce que l'utilisateur X a obtenu une information ?, "Quel est la demande de X ?", "Qu'elle est le but préalable de X ?", "Est-ce que l'utilisateur est satisfait ?", " Est-ce que le but de X est atteint ?" (1) et relève de ce méta-niveau. Il est en effet rare qu'un usager se présente en demandant: "Mon but est de vous demander de me dire les horaires de bus de la ligne 3". C'est donc bien au système qu'incombe la détermination du but de l'usager et son préalable est bien que le système sache que l'usager est un interlocuteur qui a habituellement un but. La notion de but est un concept implicite, parmi d'autres, à la situation de communication.
Ces concepts implicites sont-ils accessibles aux systèmes de traitement automatique ?
Les systèmes de dialogue actuels raisonnent déjà peu ou prou en terme de but de l'utilisateur. Rien n'interdit donc qu'ils puissent répondre prochainement à des questions du type (1) dans la mesure où les notions de satisfaction de but, de pertinence, y seront implantées. Et ce d'autant plus que c'est à l'aulne de ces prises en compte qu'ils seront évalués en fin de compte par les usagers. Un système qui répondrait aux questions (1) devient un système de dialogue qui fait en même temps de l'analyse de dialogue, qui a un regard sur le déroulement du processus de communication. Dès lors la question "Est-ce que cette réplique est pertinente ?" pourrait aussi et devrait pouvoir *in fine* lui être posée.
Exemple de dialogue entre un client (C) et un agent (A)

| | | |
|---|---|---|
| C- | *Bonjour,* | ouverture |
| | *je cherche la piscine municipale...* | requête_1 |
| | *et puis les horaires...* | requête_2 |
| A- | *La piscine municipale est 2, rue Machin,* | réponse-requête_1 |
| | *mais il y en a une plus près.* | complétive |
| C- | *Ah bon ?* | expressif-continuatif |
| A- | *Oui, il y a la piscine Truc tout près d'ici.* | information_1 |
| C- | *C'est par où ?* | requête_3 |
| A- | *Vous prenez la première à gauche et c'est un peu plus loin* | réponse-requête_3 |
| C- | *A gauche ?* | requête-clarification |
| A- | *Oui c'est cela.* | confirmation |
| C- | *Très bien merci.* | satisfaction |

[7] Dès qu'intervient le contexte pragmatique de l'application, les jeux de test perdent bien évidemment en généralité. Cela ne remet pas en cause la généricité de la méthodologie proposée.

*Au revoir.* clôture
A- *Au revoir.* clôture-réactive

Les questions suivantes pourraient être posées au système (s'il joue le rôle d'agent dans ce dialogue) :

Q *C veut-il aller à la piscine municipale ?*
R *oui*
ou
Q *C a-t-il formulé une requête à propos de la piscine Truc ?*
R *oui*

Aussi nous pensons que la méthode DQR doit pouvoir se généraliser à l'évaluation du dialogue D (la séquence QR étant insérée au cours du dialogue ou posée à la fin). Nous donnons ci-après quelques exemples pour les niveaux 4 à 7.

## 6.6 Niveau Interprétation du Type d'acte illocutoire

Les tests de ce niveau portent sur une réplique. Ils doivent permettre de diagnostiquer si le système de dialogue reconnaît les types d'actes de dialogue (ou buts illocutoires), notamment pour les actes directs et indirects, elliptiques, etc. L'exemple (17-17') montre l'effet contextuel qui influe sur la réponse R (la question Q porte sur la réplique antécédente).

(17) D *Un billet pour Paris SVP*
Q *Est-ce une demande ?*
R *Oui*
une demande indirecte
(17') D *Qu'est-ce que vous m'avez demandé ?*
D *Un billet pour Paris SVP*
Q *Est-ce une demande ?*
R *Non*
une réponse clarificatrice

Les tests de ce niveau doivent permettre de diagnostiquer également les incidences comme la clarification, l'argumentation, l'exemplification, la réparation, etc. Par exemple (18) teste la réparation :

(18) D *Je m'appelle Dupont*
D *Est-ce que votre nom est Durand ?*
D *Non, moi c'est Dupont*
Q *Est-ce que le client s'appelle Durand ?*
R *Non*

## 6.7 Niveau Reconnaissance des Intentions

A ce niveau, les tests portent sur un échange ou sur l'ensemble du dialogue (appelé transaction dans ce contexte applicatif). Les questions portent sur les buts et intentions de l'utilisateur, sur la situation initiale et sur la situation finale (satisfaction, informations acquises, etc.). Nous avons par exemple :

(19) D *Vous m'avez demandé un billet aller-retour ?*
D *Oui, pour Paris, SVP*
Q *Le client veut-il aller à Paris ?*
question sur l'intention préalable
R *Oui*
(20) D *Il y a un premier feu en face de la gendarmerie et un second près de l'école*
D *Dois-je tourner à gauche après le feu ?*
D *Oui, le premier*
D *D'accord et là, je trouve la Mairie*
Q *Le client sait-il aller à la Mairie ?*
question sur la satisfaction du but
R *Oui*

## 6.8 Niveau Pertinence de la Réponse

Ce niveau de test concerne les réponses de la machine vis-à-vis de l'utilisateur. Les questions s'adressent donc en principe à l'utilisateur en cours ou à la fin du dialogue. Mais elles peuvent aussi s'adresser à un observateur extérieur et cet observateur pourrait d'ailleurs être une machine à qui on fournirait les critères de pertinence nécessaires.
Le problème de la pertinence se pose sur plusieurs dimensions :
(a) vis-à-vis des ressources linguistiques : pour un même contenu propositionnel, est-ce que l'acte est bien formulé ? Est-ce que sa force illocutoire est bien ajustée ?
(b) vis-à-vis de l'utilisateur : y a-t-il toutes les informations pertinentes dans la réponse, c'est-à-dire nécessaires et suffisantes ? Quelles sont les informations superflues ?

(19)D *Je voudrais une chambre double avec WC et téléphone*
D *Ne savez-vous pas que l'hôtel est complet aujourd'hui?*
Q *Cette question est-elle agressive ?*
R *Oui* (degré de force mal dosé)
(21)D *Vous m'avez demandé un billet aller-retour ?*
D *Oui, pour Paris, SVP*
D *Pour aller à Paris ?*
Q *Cette question est-elle nécessaire ?*
R *Non*
(21')D *Vous m'avez demandé un billet aller-retour ?*
D *Oui, pour Paris, SVP*
D *Je vous propose une place en seconde classe fumeur*
Q *Cette proposition est-elle possible à cet instant ?*
R *Non*
(car il manque le jour et l'heure on ne peut donc savoir s'il y aura des places disponibles)

Un système pourrait répondre automatiquement à ce type de question (si on lui définit ce qu'est être "pertinent"), par exemple dans un script "réservation" hôtelière: la demande de l'utilisateur remplit le slot "chambre", mais laisse les slots "date", "nombre de lits", "bain" etc. vides. Si la réponse du système peut contribuer à remplir les slots manquants, alors elle est pertinente... encore ne faudrait-il pas poser une question avec tous les slots dans un même énoncé car cela deviendrait incompréhensible. Où s'arrête alors la pertinence ?

## 6.9 Niveau Pertinence de la Stratégie

La stratégie peut-être testée au moment des ruptures que nous définissons génériquement comme l'ensemble des impasses, incompréhensions, incidences, abandons, changements brutaux de thème, remises en cause, remises en question, déroutements, etc. Lorsque le dialogue se déroule normalement, la stratégie peut ne pas être pertinente parce que simplement trop longue ou trop sinueuse. Pour tester la longueur d'un dialogue, le critère le plus simple est de compter le nombre de tours de parole qui permet d'atteindre et de satisfaire le but. Mais le cas

des ruptures est plus intéressant car il permet de diagnostiquer le système. Les ruptures se produisent pour deux raisons :
(a) la communication était fondée sur des informations non partagées (implicites d'arrière-plan, concepts inexistants, etc.), et l'on s'aperçoit tout d'un coup qu'on ne se comprend plus et que ce qu'on a dit jusque là ne servait à rien,
(b) la stratégie menée par l'un des partenaires est inadéquate (lenteur du déroulement, incidences trop nombreuses, clarifications à répétition, directivité trop grande, etc.).
Il est alors facile de faire un diagnostic sur ces ruptures. Par exemple à l'aide de tests tels que (22) et (23)

(22) D *Vous m'avez demandé un billet aller-retour ?*
D *Oui, pour Paris, SVP*
D *Je vous propose une place en seconde classe fumeur*
D *Non je préfère en première car j'ai 50% de réduction*
D *Pour quelle heure ?*
D *Tôt le matin ne me dérange pas, je dois seulement arriver à 12 h*
D *Voilà, une place première au train de 5 h 44, TGV avec supplément, 900 F*
D *Ah non ! Je n'ai pas de réduction sur ce train ?* <..... Rupture
Q *Le client est-il mécontent ?*
R *Oui*
Q *L'information « 50% de réduction » a-t-elle été utilisée à temps ?*
R *Non*
(il aurait fallu utiliser l'information sur la réduction de 50% plus tôt)

(23) D *Je voudrais une chambre pour mon fils*
D *Oui, pour qui ?*
D *Pour mon fils Paul, il vient me voir jeudi*
D *Une chambre simple ou double ?*
D *Simple, c'est pour mon fils*
D *D'accord, je réserve pour quand ?*
D *Ecoutez, vous le faites exprès ?* <..... Rupture
Q Le client est-il content ?
R Non
Q *L' information « pour quand » était-elle déjà disponible ?*
R *Oui*

Une question plus pertinente étant bien sûr ici: « Y a-t-il trop de questions de confirmations indirectes ? ».
Nous avons vu en 6.8 et 6.9 quels étaient les objectifs de ces niveaux. L'utilisation du paradigme DQR y suppose l'implantation de notions de plus en plus complexes au sein des systèmes (pertinence, satisfaction, bon sens...) auxquelles sera conditionnée la possibilité d'une évaluation automatique. Il permet néanmoins d'objectiver l'évaluation à ces niveaux pour, par exemple, une enquête utilisateur précise.

## Conclusion

L'évaluation des systèmes de compréhension et de dialogue oral peut se situer dans le cadre général de l'évaluation des systèmes de traitement du langage et en particulier dans le prolongement de l'évaluation de la compréhension de l'écrit. Nous avons proposé de mettre au point une méthodologie d'évaluation basée sur des tests génériques de type DQR qui devraient permettre d'améliorer nos analyses des phénomènes dialogiques et de diagnostiquer les systèmes de compréhension et de dialogue. Nous avons défini sept niveaux de test :

- *Information explicite* (niveau 1) — Repérage d'une information explicitée dans l'énoncé.
- *Information implicite* (niveau 2) — Résolution des références implicites à une information.
- *Inférence* (niveau 3) — Construction du sens complet de l'énoncé.
- *Interprétation du type d'acte illocutoire* (niveau 4) — Identification du but illocutoire d'une réplique
- *Reconnaissance des intentions (ou du but)* (niveau 5) — Satisfaction du but de l'échange
- *Pertinence de la réponse* (niveau 6) — Pertinence des informations délivrées par la machine
- *Pertinence de la stratégie* (niveau 7) — Satisfaction de l'utilisateur dans le déroulement du dialogue.

## Références


Antoine J.Y. (1995), *Conception de dessins et CHM : améliorer l'interaction orale au niveau linguistique*, in Caelen J. et Zreik K. (ed), *Le Communicationnel pour concevoir*, Europia, Paris.

Antoine J., Caelen J. (1996) *Améliorer la reconnaissance de la parole par l'intégration de contraintes linguistiques robustes : le modèle microsémantique ALPES*, 21° Journées d'Etudes sur la Parole, JEP'96, Avignon.

Antoine J.Y. (1996), *Parsing spontaneous speech without syntax*, COLING'96, Copenhague, Danemark.

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

Blanche-Benveniste, C. et al. (1990). *Le français parlé*, CNRS Editions, Paris, France.

Caelen J. et al, (1997). *Les corpus pour l'évaluation du dialogue homme-machine, ARC B2*, Journées JST-FRANCIL, Avignon.

Cole, R.A. Mariani, J. Uszkoreit, H. Zaenen, A. Zue, V. Varile, G. Zampoli, A. (1996). Chapitre *Evaluation* du livre "Survey of the State of the Art of Human Language Technology ". (http://www.cse.ogi.edu/CSLU/HLTsurvey/)

Cole, R. Hirschman, L. ..., Zue, V. (1995). The challenge of Spoken Language Systems: Research Directions for the Nineties. *IEEE transactions on Speech and Audio processing*, Vol. 3, N°1, Jan 95.

De Mori R. (1994*) Apprentissage automatique pour l'interprétation sémantique*, 20° Journées d'Etudes de la Parole, JEP'94, Trégastel, 11:19.

EAGLES, (1997). (Expert Advisory Group for Language Engineering Standards). The Handbook of Spoken Language Systems. Mouton de Gruyter Ed., à paraître.

Estival D et al. (1994), *Survey of existing Test Suites*, report du LRE 62-089 D-WP1, University of Essex.

Fouvry, F. Balkan, L. (1996). Test Suites for Quality Evaluation of NLP Products. Proceedings of *Natural Language processing and Industrial Applications*, Moncton, New-Brunswick, Canada..

Jokinen, K. (1996). Adequacy and Evaluation. Proceedings of the ECAI 96 Workshop *Gaps and Bridges: New directions in Planning and Natural Language Generation*. Budapest.

Lamel, L. et al (1995). Development of spoken language corpora for travel information. EUROSPEECH'95, Madrid, p.1961-64.

Lehmann, S. Estival, D. Oepen, S. (1996). TSNLP – Des jeux de phrases-test pour l'évaluation d'applications dans le domaine du TAL, *Actes de la conférence TALN 96*, (pp. 97-103), Marseille.

Minker, W. Bennacef, S. (1996). Compréhension et évaluation dans le domaine. 21° Journées d'Etudes sur la Parole, JEP'96, Avignon, p. 417-21.

Pérennou G (1996), *Compréhension du dialogue oral. Rôle du lexique dans le décodage conceptuel*, actes séminaire lexique du GDR-PRC CHM, Toulouse.

Rolbert, M. et Sabatier, P. (1996). Evaluation des systèmes de compréhension de textes: *Travaux sur l'évaluation des systèmes de traitement automatique du langage naturel :Etude de l'existant.* Rapport de recherche, ARC Informatique, Linguistique et Corpus écrits, AUPELF-UREF.

The FRACAS consortium (Cooper, R.et al) (1994). Harmonising the approaches. *Public Deliverables of the FRACAS Project (A Framework for Computational Semantics))*, LRE 62-051, Deliverable D7.

The FRACAS consortium (Cooper, R.et al) (1996). Using the framework. *Public Deliverables of the FRACAS Project (A Framework for Computational Semantics))*, LRE 62-051 , Deliverable D16, (Chapitre 3).