Vers l'évaluation de systèmes de dialogue homme-machine : de l'oral au multimodal

Frédéric Landragin

CNRS – Laboratoire LaTTICe (UMR 8094)
1 rue Maurice Arnoux – 92120 Montrouge
frederic.landragin@linguist.jussieu.fr

Résumé L'évaluation pour le dialogue homme-machine ne se caractérise pas par l'efficacité, l'objectivité et le consensus que l'on observe dans d'autres domaines du traitement automatique des langues. Les systèmes de dialogue oraux et multimodaux restent cantonnés à des domaines applicatifs restreints, ce qui rend difficiles les évaluations comparatives ou normées. De plus, les avancées technologiques constantes rendent vite obsolètes les paradigmes d'évaluation et ont pour conséquence une multiplication de ceux-ci. Des solutions restent ainsi à trouver pour améliorer les méthodes existantes et permettre des diagnostics plus automatisés des systèmes. Cet article se veut un ensemble de réflexions autour de l'évaluation de la multimodalité dans les systèmes à forte composante linguistique. Des extensions des paradigmes existants sont proposées, en particulier DQR/DCR, sachant que certains sont mieux adaptés que d'autres au dialogue multimodal. Des conclusions et perspectives sont tirées sur l'avenir de l'évaluation pour le dialogue homme-machine.

Abstract Evaluating human-machine dialogue systems is not so efficient, objective, and consensual than evaluating other natural language processing systems. Oral and multimodal dialogue systems are still working within reduced applicative domains. Comparative and normative evaluations are then difficult. Moreover, the continuous technological progress makes obsolete and numerous the evaluating paradigms. Some solutions are still to be identified to improve existing methods and to allow a more automatic diagnosis of systems. The aim of this paper is to provide a set of remarks dealing with the evaluation of multimodal spoken language dialogue systems. Some extensions of existing paradigms are presented, in particular DQR/DCR, considering that some paradigms fit better multimodal issues than others. Some conclusions and perspectives are then drawn on the future of the evaluation of human-machine dialogue systems.

Mots-clés : Dialogue finalisé, multimodalité, évaluation pour le dialogue homme-machine, paradigme d'évaluation, test utilisateur, diagnostic, paraphrase multimodale

Keywords: Task-driven dialogue, multimodality, evaluating human-machine dialogue, evaluation paradigm, user test, diagnosis

1 Introduction

La communauté internationale voit paraître un nombre grandissant d'articles sur l'évaluation des systèmes de dialogue oraux et multimodaux. Outre les livres de référence en dialogue homme-machine qui incluent désormais systématiquement un chapitre sur l'évaluation (Gibbon et al., 2000 ; puis López-Cózar, Araki, 2005), on trouve un grand nombre de propositions ciblées sur le dialogue oral ou sur une modalité particulière d'un système multimodal. Des paradigmes d'évaluation sont proposés, de plus en plus larges et complexes, regroupant entre autres des ensembles de métriques, des tests utilisateurs et des méthodes d'analyse de questionnaires remplis par des sujets après leur utilisation du système à évaluer. Dans la communauté française, les propositions se cantonnent pour le moment à l'oral ou à certains aspects du dialogue multimodal comme le comportement d'agents animés, et il n'y a pas encore de chapitre sur l'évaluation de la multimodalité dans des ouvrages tels que (Chaudiron, 2004) ou (Caelen, Xuereb, 2007). Ces efforts sont pertinents et louables, mais ne doivent pas faire oublier plusieurs constats récurrents qui restent particulièrement valables.

Premier constat : contrairement aux systèmes de recherche d'information, de reconnaissance de la parole ou d'analyse syntaxique, les systèmes de dialogue homme-machine restent toujours au niveau de prototypes de recherche difficiles à réaliser et à faire fonctionner, ainsi que très sensibles au comportement des utilisateurs. A part quelques exemples marginaux, ludiques par exemple, il n'existe à l'heure actuelle aucun système fiable commercialisé et utilisé de manière profitable par une population de taille importante. Autrement dit, le passage à l'échelle reste un problème majeur en dialogue homme-machine et les évaluations réalisées s'en tiennent à des prototypes de recherche ou à des systèmes professionnels tellement finalisés (militaires par exemple) qu'ils ne s'adressent qu'à un nombre extrêmement réduit d'utilisateurs. Ce qui est valable pour les systèmes oraux l'est encore plus pour les systèmes multimodaux. L'évaluation pour le dialogue homme-machine se cantonne donc à un périmètre limité qui, sans remettre en cause son utilité, nuance quelque peu sa portée.

Deuxième constat : il risque d'exister bientôt autant de méthodologies d'évaluation que de systèmes proprement dits. Ce n'est pas un problème en soi, mais cela soulève des interrogations. En particulier, on peut s'interroger sur le bien-fondé d'une méthode d'évaluation proposée par les concepteurs d'un système dans le but d'évaluer ce seul système, la méthode étant elle-même évaluée par son application au système en question. Cette description peut sembler caricaturale, elle reflète pourtant une certaine réalité, ou en tout cas elle s'en approche. Cette situation est tout à fait normale et inévitable, compte tenu du nombre réduit de systèmes et, face aux avancées de chaque système, de la nécessité de prendre en compte des aspects qui ne sont pas traités par les méthodologies d'évaluation existantes. Ainsi, on en vient à proposer ou à étendre une méthodologie d'évaluation en vue de pouvoir évaluer les avancées d'un nouveau système. Les avancées technologiques rapides ne font qu'augmenter ce phénomène. L'évaluation pour le dialogue homme-machine semble ainsi perpétuellement en retard par rapport à son objectif.

Troisième constat : l'évaluation sert non seulement à améliorer le développement d'un système particulier (en passant par des mesures, des diagnostics et des questionnaires de satisfaction), mais aussi à comparer des systèmes les uns par rapport aux autres. Plusieurs campagnes ont été lancées, et ce qu'il en ressort finalement, c'est qu'il est très difficile de comparer plusieurs systèmes de dialogue, même s'ils ont été réalisés pour des contextes applicatifs comparables, par exemple le renseignement ferroviaire ou hôtelier pour ne citer que ces deux applications largement exploitées. Pour cet aspect, le domaine du dialogue

homme-machine semble poser un problème plus délicat que les autres domaines du TAL, et contribue à l'image de fragilité attachée à son évaluation.

Face à ces constats, on peut s'interroger sur la faisabilité de l'évaluation pour le dialogue homme-machine. Dans ce but, la section 2 passera en revue les principaux problèmes et quelques méthodologies qui nous semblent prometteuses. La question de la faisabilité nous semble constituer un problème de fond qui n'est pas assez discuté et pour lequel nous essaierons d'apporter quelques pistes de réflexion. Les critiques que nous venons de porter avec les constats précédents ne nous empêcheront pas, dans un second temps, de proposer des pistes pour une meilleure prise en compte de la multimodalité dans des paradigmes existants. La section 3 s'attachera ainsi à faire un point sur les possibilités d'extension à la multimodalité des méthodologies prévues pour l'oral, en particulier les méthodes DQR et DCR, et sur la prise en compte de phénomènes pour l'instant ignorés dans les méthodologies déjà prévues pour la multimodalité. A défaut d'évaluer nos propositions sur des systèmes existants ou sur de nouveaux systèmes (c'est l'une des perspectives de ce travail), nous les expliciterons sur des exemples typiques tels que le classique « mets ça ici » (Bolt, 1980).

2 Méthodologies existantes

Dans le contexte du dialogue homme-machine oral, beaucoup de méthodologies ont été proposées (Antoine, Caelen, 1999 ; Bonneau-Maynard et al., 2006 ; Devillers et al., 2004 ; Dybkjær et al., 1998 ; Eckert et al., 1998 ; Litman, Pan, 1999 ; Möller et al., 2007). Elles constituent une sorte de cadre de référence comprenant des recommandations pour mettre en œuvre des tests d'interaction avec des utilisateurs, des méthodes pour analyser automatiquement ou semi-automatiquement les traces d'interaction obtenues, des repères pour déterminer des métriques d'évaluation, ou encore des principes pour constituer et analyser des questionnaires remplis *a posteriori* par les utilisateurs. Chaque évaluateur peut ainsi piocher dans ce stock pour déterminer la ou les méthodes qu'il va appliquer. En fait, un seul test semble insuffisant et une véritable évaluation semble devoir grouper plusieurs types de test. Les campagnes d'évaluation (EVALDA/MEDIA), les groupes de travail (groupe MADCOW, groupe « compréhension de parole » du GdR I3) et les divers consortiums de projets européens exploitent largement ce principe. Lorsque plusieurs systèmes sont en jeu et que l'évaluation est comparative, des règles de fonctionnement peuvent être définies de manière à mieux contrôler la qualité de l'évaluation. La campagne d'évaluation par défi avec sa gestion croisée des rôles des concepteurs des systèmes en jeu (Antoine, 2003) en est un exemple.

Les principales propositions de méthodologie s'accompagnent chacune d'une idée originale qui vient simplifier la mise en œuvre d'un type de test en lui apportant un moyen d'être opérationnalisé dans un contexte déterminé. Le paradigme du groupe MADCOW (Hirschman, 1992) apporte ainsi la notion de gabarit qui caractérise les solutions minimales et maximales à une requête et rend ainsi son évaluation plus rigoureuse. Le paradigme PARADISE – PARAdigm for DIalogue System Evaluation (Walker et al., 2001) se focalise sur la maximisation de la satisfaction de l'utilisateur et propose de prendre la satisfaction de la tâche comme référence. Autre exemple d'idée originale, (López-Cózar et al., 2003) propose d'évaluer un système en générant automatiquement des énoncés utilisateurs de test, c'est-à-dire en modélisant le comportement de l'utilisateur, y compris ses erreurs. En France, la méthodologie DQR – Donnée–Question–Réponse (Zeiliger et al., 1997) introduit le principe de questionner le système sur le point à évaluer, avec l'avantage de déplacer ainsi l'objet de l'évaluation de la donnée vers la question, et donc ni sur les réponses ou réactions du système (méthode « boîte noire », qui ne nécessite pas d'explorer les structures internes au système, mais qui manque de précision), ni sur les structures sémantiques du système (méthode « boîte

transparente », précise et conduisant facilement à un diagnostic, mais qui nécessite de disposer de représentations sémantiques de référence). Encore faut-il que le système soit capable de répondre aux questions Q de DQR. Le paradigme adapté DCR – Demande–Contrôle–Réponse/Résultat/Référence (Antoine, Caelen, 1999) minimise ce problème en remplaçant la question par un contrôle qui est une simplification ou une reformulation de la demande utilisateur initiale. Pour sa part, le paradigme PEACE – Paradigme d'Evaluation Automatique de ComprEhension (Devillers et al., 2002) apporte l'idée originale de modéliser l'historique du dialogue par une paraphrase, ce qui permet de rester dans le mode « boîte noire » tout en permettant une évaluation de la compréhension en contexte.

Dans le contexte du dialogue homme-machine multimodal, les propositions sont loin d'être aussi pertinentes. Le paradigme PROMISE – PROcedure for Multimodal Interactive System Evaluation (Beringer et al., 2002) est présenté comme une extension de PARADISE à la multimodalité, avec des principes pour affecter des scores aux entrées et sorties multimodales. La proposition reste en fait à un niveau très approximatif, bien en deçà de la variété des phénomènes multimodaux. Les aspects intéressants de l'article concernent le dialogue oral, avec des considérations sur le niveau de complétude de la tâche et le niveau de coopération de l'utilisateur. Les travaux de Bernsen et Dybkjær, qui font pourtant référence dans le milieu du dialogue multimodal, sont plutôt décevants en ce qui concerne l'évaluation. (Bernsen, Dybkjær, 2004) présente ainsi une méthodologie prévue pour un système, avec une focalisation sur la méthode du questionnaire rempli *a posteriori* par les utilisateurs (la raison donnée est d'ailleurs que les autres méthodes ne sont pas encore bien établies). Malheureusement, les questions du questionnaire restent à un niveau très superficiel pour ce qui concerne la multimodalité : « avez-vous utilisé la souris ou avez-vous pointé sur l'écran ? », « quelles étaient vos impressions en produisant un geste ? », et « auriez-vous aimé en faire plus avec le geste ? si oui, pour faire quoi ? ». Les réponses qui ont été fournies par les utilisateurs semblent également très pauvres, d'autant plus qu'une des conclusions des auteurs est que les utilisateurs ont préféré parler plutôt qu'exploiter les possibilités multimodales… Pour sa part, (Dybkjær et al., 2004) est plus une revue de méthodologies et de projets qu'une proposition de nouvelle méthodologie pour la multimodalité : le propos reste au niveau de recommandations très générales. Par ailleurs, une des remarques finales de l'article rejoint notre point de vue : « The field is moving rapidly beyond the standard task-oriented, speech-only SLDS [Spoken Language Dialogue System] towards multimodal SLDSs, mobile systems, situation-aware systems, location-aware systems, internet access systems, educational systems, entertainment systems, etc. In fact, technology development may appear to speed further ahead of the knowledge we already have on evaluation, usability and standards, increasing the proportion of what we do not know compared with what we do know. ». Dans un autre registre, (Vuurpijl et al., 2004) présente un outil, appelé « µeval », pour la transcription des données multimodales et l'évaluation d'un système. Or l'évaluation ne concerne que les tours de dialogue et ne traite pas les phénomènes multimodaux. Enfin, (Walker et al., 2004), malgré son titre, se focalise sur les modèles utilisateur et les stratégies de dialogue (oral) mais quasiment pas sur les aspects multimodaux. D'une manière générale pour l'évaluation des systèmes multimodaux, on ne retrouve donc pas les principes appliqués dans les campagnes d'évaluation des systèmes oraux. C'est ce que nous allons contribuer à faire en nous focalisant sur les méthodologies qui nous semblent les plus prometteuses.

3 Extension au dialogue multimodal

(Zeiliger et al., 2000) ont retenu une méthodologie de type « boîte noire » permettant de faire un diagnostic du système, méthodologie qui repose sur des tests génériques pour l'évaluation de la compréhension d'un énoncé isolé. Les aspects contextuels ont été négligés (nous y

reviendrons avec PEACE), mais c'était le prix à payer pour obtenir une méthodologie simple et bien délimitée. Le principe est de procéder à des évaluations ponctuelles, chacune d'entre elles étant centrée sur un phénomène particulier. Ainsi, dans la matérialisation DQR, l'évaluation ponctuelle prend la forme d'une question Q adressée au système et permet de vérifier sa bonne compréhension de la demande initiale D. Un des exemples donnés concerne la résolution des anaphores, avec la demande, la question et la réponse suivantes :

- D = « Vous prenez la rue à droite et vous la suivez sur 300 mètres » (énoncé initial, tel qu'il a été adressé au système dans le but de faire avancer la tâche) ;

- Q = « Suivre rue à droite ? » (question adressée au système juste après l'énoncé D et destinée à évaluer la bonne compréhension de D) ;

- R = « oui » (réponse du système montrant que l'anaphore a été bien comprise et rendant l'évaluation positive).

Les auteurs spécifient sept niveaux caractérisant la portée des questions posées. Nous allons reprendre ces niveaux en indiquant à chaque fois comment étendre le paradigme pour pouvoir l'exploiter en dialogue multimodal.

3.1 DQR multimodal

Niveau 1 = « information explicite ». Il s'agit du repérage d'une information explicitée dans l'énoncé, l'intérêt étant de tester la bonne compréhension de l'énoncé littéral compte tenu de la grande variabilité du langage spontané. Les exemples donnés par les auteurs se contentent de reprendre une partie de l'énoncé et de demander une confirmation de la compréhension de cette partie : D = « vous prenez à droite après les bâtiments blancs aux volets bleus » puis Q = « volets blancs ? » ou « volets bleus ? ». L'extension de ce principe à la multimodalité consiste à poser des questions sur les éléments de l'énoncé multimodal. Avec D = « mets ça ici » + geste en (x_1, y_1) + geste en (x_2, y_2), on peut tester les capacités de capture de la multimodalité en posant les questions Q suivantes : « ça ? » + geste en (x_1, y_1) ; « mettre ici ? » + geste en (x_2, y_2) ; « mettre ça ? » + geste en (x_2, y_2) ; « mettre ça ici » + geste en (x_2, y_2) + geste en (x_1, y_1) ; etc. La procédure peut sembler naïve, mais elle permet de tester de manière très simple le bon appariement des gestes avec les expressions référentielles, ce qui constitue un processus non négligeable de la fusion multimodale. Une attention particulière sera donnée à la synchronisation temporelle entre les mots prononcés et les gestes produits. Ainsi, un décalage temporel entre « ça » et le geste dans la question Q pourra conduire selon le système à une réponse positive reflétant sa robustesse pour l'appariement multimodal même quand les conditions de production sont déviantes, ou au contraire à une réponse négative reflétant l'incapacité du système à sortir d'un certain intervalle temporel.

Niveau 2 = « information implicite ». Ce niveau concerne la résolution des anaphores, des ellipses, des incomplétudes et autres informations implicites mais récupérables aux niveaux syntaxique et sémantique. Un exemple fait intervenir : D = « donnez-moi un billet pour Paris et aussi pour Lyon » et Q = « billet pour Lyon ? ». La résolution de la référence étant l'un des principaux aspects de la multimodalité spontanée, un DQR multimodal devra bien entendu en rendre compte. Ainsi, en reprenant comme D la primitive universelle de la multimodalité, « mets ça ici » avec deux gestes de désignation, les questions Q pourront introduire des précisions sur les référents, en partant par exemple de la mention de leur catégorie et en allant jusqu'à donner leur identifiant unique tel qu'il est géré par le système : « mettre cet objet ? » + geste en (x_1, y_1) ; « mettre ce fichier ? » + geste en (x_1, y_1) ; « mettre 'submis.tex' ? » (sans

geste) ; « mettre obj_{4353} ? » (sans geste) ; etc. La procédure d'évaluation inclut donc la paraphrase en langage naturel d'une référence multimodale. Ce qui reste simple pour le geste déictique l'est beaucoup moins pour les autres types de gestes co-verbaux. Imaginons par exemple que « mets ça ici », ou plutôt « déplace ça ici » pour ne pas trop compliquer l'exemple[1], s'accompagne d'un seul geste qui part de l'objet à déplacer et aboutit au lieu de destination. Selon une première hypothèse, cette trajectoire gestuelle est considérée comme la matérialisation de la nécessaire transition entre la désignation d'objet et la désignation de lieu. Dans ce cas, seules les extrémités de la courbe sont utilisées lors des analyses sémantiques : le point (x_1, y_1) puis l'objet présent en ce point ou dans un voisinage immédiat sont unifiés avec « ça », et le point (x_2, y_2) est unifié avec « ici ». Autrement dit on revient au cas précédent. Selon une seconde hypothèse, la trajectoire est considérée comme la combinaison de ces deux désignations avec un geste co-verbal illustratif apportant une caractéristique de l'action de déplacement, à savoir le chemin (ou points de passage) à suivre. La trajectoire est alors analysée d'un point de vue temporel (courbe produite de manière régulière, sans point d'arrêt significatif) et d'un point de vue structurel (arc de cercle), avant d'être unifiée à « déplace », c'est-à-dire d'être interprétée comme un chemin de déplacement. Si l'on veut tester cette fonctionnalité du système multimodal, il suffit de poser une question Q supplémentaire : « suivre cette trajectoire ? » ou « déplacer selon ces points de passage ? », en reprenant dans un cas comme dans l'autre le geste complet. Le seul inconvénient reste celui qui est valable pour l'ensemble de la méthodologie DQR, à savoir la nécessité pour le système de traiter de telles questions.

Niveau 3 = « inférence ». Il s'agit ici de la construction du sens complet de l'énoncé, la difficulté étant l'identification des sous-entendus, identification qui fait appel à des raisonnements de sens commun et à des inférences pragmatiques. Avec D : « je voudrais un aller-retour pour Paris », les auteurs proposent Q : « vouloir billet ? ». Cet aspect est indépendant des modalités, et reste valable dans l'état pour le dialogue multimodal. Même si des inférences peuvent être identifiées lors de l'utilisation consécutive de plusieurs gestes, il est vrai que ce niveau concerne surtout la langue orale.

Niveau 4 = « interprétation du type d'acte illocutoire ». On entre ici dans les niveaux de dialogue, avec un premier aspect concernant les actes de langage et la capacité du système à identifier le bon type d'acte, même en cas d'acte de langage indirect. Avec D : « un billet pour Paris », qui peut faire suite à une question ou qui peut correspondre à une demande initiale, la question Q : « est-ce une demande ? » permet d'évaluer l'acte identifié par le système. Il est nécessaire ici de distinguer deux types de dialogue multimodal. Dans le premier type, les gestes et les autres modalités de communication restent co-verbaux, c'est-à-dire que l'information qu'ils apportent s'ajoute à celle portée par l'énoncé en langage naturel, et l'acte de langage de l'énoncé multimodal est celui de l'énoncé oral. C'est le cas des exemples étudiés jusqu'à présent avec « mets ça ici » et « déplace ça ici ». Un autre type de dialogue multimodal autorise des gestes quasi-linguistiques ou, d'une manière générale, un message effectué avec une modalité autre que le langage naturel et portant en lui-même un acte de communication (ou acte de dialogue), similaire à un acte de langage. C'est le cas par exemple lorsque l'on étend une interface graphique et que l'on autorise des gestes ayant des formes telles qu'une croix ou une flèche, chaque forme étant associée à un déclenchement d'action. La croix est un équivalent du « supprimer » en langage naturel et prend comme argument

[1] La différence relève de la résolution de la référence aux actions : « déplace » ne peut référer qu'à une action de déplacement, alors que « mets » peut référer soit à une action de déplacement, soit à une action de création. Les considérations suivantes concernent l'existence de plusieurs primitives pour « déplace » : `'déplacer(objet,lieu)'` et `'déplacer(objet,lieu,chemin)'`.

l'objet ciblé par le geste. Quant à la flèche, elle équivaut au « déplace ça ici » avec les aspects dont nous avons parlé. Dans un tel type de système multimodal, l'analyse en termes d'actes de langage et d'actes de dialogue met en jeu plusieurs processus : l'identification de l'acte de l'énoncé oral et du prédicat associé, l'identification de l'acte du geste et du prédicat associé, ainsi que l'analyse de la compatibilité ou de l'incompatibilité entre les diverses hypothèses de manière à aboutir à un seul acte de dialogue qui sera à l'origine de la réaction du système. Par exemple, un geste en forme de flèche effectué en même temps que l'énoncé oral « déplace ça ici » ne posera pas de problème, alors qu'un geste en forme de croix effectué en même temps que le même énoncé oral conduira à une incohérence. Selon le système, cette incohérence pourra être interprétée soit comme une erreur soit comme l'exécution de deux tâches parallèles. Tous ces aspects peuvent être évalués grâce aux questions Q suivantes : « ce geste est-il une demande ? » + geste ; « ce geste accompagne-t-il la parole ? » + geste ; « l'énoncé multimodal est-il une demande ? » ; etc. A ce stade, nous avons fait le tour des principaux problèmes qui se posent pour le traitement des entrées en dialogue multimodal.

Niveau 5 = « reconnaissance des intentions ». Il s'agit ici de déterminer les intentions ou les buts sous-jacents aux énoncés de l'utilisateur, donc à un niveau plus profond que le niveau 4. Le principe est d'interroger explicitement les états intentionnels, avec des questions Q telles que : « l'utilisateur sait-il/veut-il… ? ». De tels états intentionnels sont indépendants des modalités, et l'extension de DQR à la multimodalité ne change rien à ce niveau.

Niveau 6 = « pertinence de la réponse ». L'objet de la question évaluative est ici assez large puisqu'il s'agit de tester la pertinence des réponses du système. Les aspects couverts sont donc *a priori* les capacités linguistiques (et donc multimodales) dont font preuve les réponses, leur adéquation par rapport à l'énoncé initial de l'utilisateur, par rapport aux connaissances de l'application, par rapport aux moyens de communication, par rapport au profil de l'utilisateur, etc. Dans (Zeiliger et al., 2000), les exemples de questions Q sont les suivants : « cette question est-elle agressive ? » ; « cette question est-elle nécessaire ? » ; « cette proposition est-elle possible à cet instant ? ». Ces exemples interrogent à la fois la forme et le contenu de la réponse. Réaliser un système de dialogue capable de répondre à de telles questions n'est pas simple. Cela suppose que le système (chacun de ses modules) soit capable d'évaluer la pertinence de ses propres décisions, un peu comme dans le modèle du carnet d'esquisses de (Sabah, 1996). En dialogue multimodal, il faudrait ajouter tous les aspects liés à la multimodalité en sortie, c'est-à-dire aux décisions que le système a prises lors de la détermination du contenu et de la forme de la réponse multimodale. Ainsi, des exemples possibles pour Q sont : « le choix de la ou des modalités de sortie est-il pertinent ? » ; « le message est-il surchargé ? » ; « le message est-il redondant ? » ; « le message est-il bien synchronisé ? » ; « les informations présentées sont-elles pertinentes ? » ; etc. Ces questions font le tour des principaux problèmes qui se posent en sortie dans le dialogue multimodal. Elles intègrent des aspects métalinguistiques qui ne sont généralement pas implantés dans le modèle conceptuel et le lexique des systèmes. Ces aspects métalinguistiques, en plus des aspects métacognitifs vus précédemment, constituent clairement une limite à la faisabilité de ce sixième niveau, ainsi qu'à celle du niveau suivant.

Niveau 7 = « pertinence de la stratégie ». Ce dernier niveau teste la qualité de la stratégie de dialogue, c'est-à-dire si elle a été efficacement menée et si elle est réussie. En fait, les questions couvrent non seulement la stratégie de dialogue, mais également la stratégie de gestion de la tâche : « le client est-il content ? » ; « y a-t-il trop de questions de confirmation indirectes ? » ; etc. Peuvent également être interrogés la lenteur, le nombre d'incidences, les raisons possibles d'une rupture, c'est-à-dire tous les aspects que le système de dialogue peut (théoriquement) calculer. Ces aspects étant indépendants des modalités, nous les gardons et nous obtenons un DQR multimodal complet.

3.2 DCR multimodal

Comme nous l'avons déjà évoqué, une autre matérialisation est le paradigme DCR (Antoine, Caelen, 1999) qui, en remplaçant la question évaluative par un contrôle C, minimise le problème de la capacité du système à répondre à cette question parfois métalinguistique. Le contrôle consiste en une simplification ou une reformulation de la demande utilisateur initiale. Ainsi, en reprenant quelques-uns des exemples précédents, on fera intervenir les contrôles multimodaux C suivants : « mets 'submis.tex' en (x_1, y_1) » ; « déplace obj_{4353} de (x_1, y_1) à (x_2, y_{12} » ; « déplace obj_{4353} selon les points de passage (x_3, y_3), (x_4, y_4), (x_5, y_5)… ». Passer du DQR multimodal à un DCR multimodal nécessite donc la paraphrase de manière simple et non ambiguë des références multimodales, avec la description en langage naturel de coordonnées spatiales. Les autres aspects ne posent pas de problème particulier, en tout cas pas plus de problème que le passage de DQR à DCR.

3.3 Vers un PEACE pour le dialogue multimodal ?

Les principes de PEACE (Devillers et al., 2002) sont la reformulation de l'historique en une phrase unique, l'utilisation de cette phrase pour une évaluation contextuelle de l'énoncé courant, et l'exploitation de représentations sémantiques de référence. Nous avons déjà parlé de la difficulté d'appliquer ce dernier principe au dialogue multimodal, et c'est donc l'idée de reformulation de l'historique qu'il s'agit d'étudier ici.

La modélisation de l'historique du dialogue est un problème récurrent en dialogue homme-machine, et s'avère particulièrement complexe en dialogue multimodal (Landragin, 2004). En nous focalisant sur le problème de la référence aux objets, l'historique doit conserver à la fois l'identifiant des référents (pour ressortir ceux-ci lors de l'interprétation d'une anaphore) et les mentions utilisées pour y référer (non seulement pour interpréter les éventuelles références mentionnelles ou métalinguistiques, mais aussi et surtout pour interpréter les ellipses, en particulier les ellipses nominales). En multimodal, il en est de même et les formes référentielles multimodales doivent donc être conservées. (Landragin, 2004) montre que l'état de la scène visuelle doit également être sauvegardé à chaque étape, conduisant ainsi au moins à un historique linguistique, un historique gestuel et un historique visuel. Une chaîne de référence faisant appel aux modalités utilisées ou aux souvenirs de l'utilisateur est alors interprétable, par exemple : « l'objet que je viens de désigner », « les deux objets groupés un peu plus loin », « celui de gauche », « celui qui était à droite », « le dernier ». Ces expressions référentielles montrent d'elles-mêmes que la paraphrase d'un historique multimodal est une tâche impossible à réaliser, ou alors au prix de simplifications telles que le biais introduit enlèvera toute plausibilité à l'évaluation. En effet, le seul processus de paraphrase automatisable est l'utilisation systématique des identifiants des référents, or cette solution semble plus destructrice en dialogue multimodal qu'en dialogue oral : elle met en effet de côté l'ensemble des aspects multimodaux. Il nous apparaît donc difficile d'appliquer les principes de PEACE au dialogue multimodal.

4 Conclusion et perspectives

L'évaluation des systèmes de dialogue multimodaux s'avère en fin de compte plus complexe que celle des systèmes oraux (pourtant déjà bien délicate), surtout quand la multimodalité est considérée comme l'association complémentaire du langage naturel et d'autres modalités de communication sur lesquelles s'appuie le langage. Dans cet article nous avons proposé une extension du paradigme DQR/DCR à la multimodalité. Nos illustrations ont porté sur des

dérivations de l'exemple à l'origine de l'essor des travaux sur la multimodalité (« mets ça ici »), et sont ainsi applicables à la majorité des phénomènes de référence multimodale. Si nos propositions et les remarques afférentes constituent un produit de recherche en soi, elles peuvent peut-être aussi être utiles à la réalisation de systèmes de dialogue, de par les questions soulevées et les préoccupations détaillées.

Plusieurs aspects restent à étudier pour obtenir une méthodologie couvrant le champ occupé actuellement par le dialogue multimodal. Un aspect concerne les pistes qui sont explorées en ce moment pour simplifier la réalisation de systèmes multimodaux : comme les phénomènes sont de plus en plus nombreux et les processus de plus en plus complexes, une solution consiste à faciliter le travail des développeurs en leur fournissant des environnements de développement capables d'automatiser certaines phases de conception. Un exemple en est l'approche par dérivation et génération de modèles : en simplifiant un peu, les développeurs écrivent des modèles, et l'environnement de développement génère automatiquement des modules du système à partir de ces modèles. L'évaluation des systèmes obtenus doit alors reposer non seulement sur le comportement du système face à des utilisateurs, mais aussi sur la qualité des modèles initiaux et sur celle de la chaîne de dérivation de modèles. D'autre part, quand les systèmes de dialogue multimodaux seront suffisamment nombreux, il nous apparaît utile de revenir sur la méthode d'évaluation par défi. Son principe, que ce soit l'étape de dérivation d'énoncés à partir d'un ensemble d'énoncés initiaux ou l'échange des rôles entre différents concepteurs, nous apparaît en effet tout à fait pertinent pour le dialogue multimodal.

Références

ANTOINE J.-Y. (2003). Pour une ingénierie des langues plus linguistique. HDR informatique, Université de Bretagne Sud, Vannes.

ANTOINE J.-Y., CAELEN J. (1999). Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat). *Langues*, vol. 2, n° 2, 130-139.

BERINGER N., KARTAL U., LOUKA K., SCHIEL F., TÜRK U. (2002). PROMISE – A procedure for multimodal interactive system evaluation. *Proceedings of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, 77-80.

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

BOLT R.A. (1980). Put-That-There: Voice and gesture at the graphics interface. *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, Seattle.

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C., VILLANEAU J. (2006). Results of the French Evalda-Media Evaluation Campaign for Literal Understanding. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Gênes, Italie.

CAELEN J., XUEREB A. (2007). *Interaction et pragmatique*. Paris : Hermès-Lavoisier.

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

DEVILLERS L., MAYNARD H., PAROUBEK P. (2002). Méthodologies d'évaluation des systèmes de dialogue parlé : réflexions et expériences autour de la compréhension. *Traitement Automatique des Langues*, vol. 43, n° 2, 155-184.

DYBKJÆR L., BERNSEN N.O., DYBKJÆR H. (1998). A Methodology for diagnostic evaluation of spoken human-machine dialogue, *International Journal of Human Computer Studies*, vol. 48, 605-625.

DYBKJÆR L., BERNSEN N.O., MINKER W. (2004). Evaluation and usability of multimodal spoken language dialogue systems, *Speech Communication*, vol. 43, n° 1-2, 33-54.

ECKERT W., LEVIN E., PIERACCINI R. (1998). Automatic evaluation of spoken dialogue systems. *Proceedings of the 2nd Workshop on Formal Semantics and Pragmatics of Dialogue*, University of Twente, Enschede, The Netherlands, 99-110.

GIBBON D., MERTINS I., MOORE R.K. (2000). *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*. Kluwer Academic Publishers.

HIRSCHMAN L. (1992). Multi-Site Data Collection for a Spoken Language Corpus: MADCOW. *Proceedings of the DARPA Speech and Natural Language Workshop*, New York.

LANDRAGIN F. (2004). *Dialogue homme-machine multimodal*. Paris : Hermès-Lavoisier.

LITMAN D.J., PAN S. (1999). Empirically evaluating an adaptable spoken dialogue system. *Proceedings of the 7th International Conference on User Modeling*.

LOPEZ-COZAR R., ARAKI M. (2005). *Spoken, multilingual and multimodal dialogue systems: Development and assessment*. John Wiley & Sons, Ltd.

LOPEZ-COZAR R., DE LA TORRE A., SEGURA J.C., RUBIO A.J. (2003). Assessment of dialogue systems by means of a new simulation technique, *Speech Communication*, vol. 40, 387-407.

MÖLLER S., SMEELE P., BOLAND H., KREBBER J. (2007). Evaluating spoken dialogue systems according to de-facto standards: A case study, *Computer Speech and Language*, vol. 21.

SABAH G. (1996). Le « carnet d'esquisses » : une mémoire interprétative dynamique. *Actes du colloque AFCET – AFIA*, Rennes.

VUURPIJL L.G., TEN BOSCH L., ROSSIGNOL S., NEUMANN A., PFLEGER N., ENGEL R. (2004). Evaluation of multimodal dialog systems. *Proceedings of the LREC Workshop on Multimodal Corpora and Evaluation*, Lisbon, Portugal.

WALKER M.A., PASSONNEAU R., BOLAND J.E. (2001). Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems. *Meeting of the Association of Computational Linguistics*.

WALKER M.A., WHITTAKER S., STENT A., MALOOR P., MOORE J., JOHNSTON M., VASIREDDY G. (2004). Generation and Evaluation of User Tailored Responses in Multimodal Dialogue. *Cognitive Science*, vol. 28, n° 5, 811-840.

ZEILIGER J., ANTOINE J.-Y., CAELEN J. (2000). La méthodologie DQR d'évaluation qualitative des systèmes de dialogue oral homme-machine, Dans Mariani J., Masson N., Néel F., Chibout K. (Ed.) *Ressources et Evaluations en Ingénierie de la Langue*, AUF et De Boeck Université.