Les corpus pour l'évaluation du dialogue homme-machine

J. Caelen(1), J. Zeiliger(2), M. Bessac(1), J. Siroux(3), G. Perennou(4)
(1) CLIPS-IMAG Grenoble, (2) ICP-INPG Grenoble, (3) IRISA Lannion, (4) IRIT Toulouse

Résumé
Le dialogue oral homme-machine commence à dépasser le cadre des systèmes prototypes de laboratoire mais n'a pas encore atteint le terrain des applications professionnelles ou grand public. On ne le rencontre guère que dans les serveurs vocaux, où il n'est souvent qu'une succession de questions directives de la machine et de réponses brèves de l'usager, quand ce ne sont pas des réponses codées sur le clavier DTMF du combiné téléphonique. Cependant des études sont actuellement en cours pour intégrer dans des applications réelles des dialogueurs plus souples, à initiatives partagées —voir par exemple le projet ARISE (Automatic Railway Information System for Europe) de la communauté Européenne [ARISE, 96].
N'étant pas encore en possession de vrais systèmes s'adressant à de vrais utilisateurs placés dans de réelles situations d'usage, il faut le plus souvent créer ces situations pour étudier les systèmes de dialogue en développement dans les laboratoires. Le problème de l'évaluation de ces systèmes est donc délicat et c'est peut-être ce qui en fait l'intérêt scientifique. En effet on a autant besoin de critères **prédictifs** que de critères de **performances** pour l'évaluation prospective des systèmes. Les critères **prédictifs** sont même à certains égards presque plus utiles dans la phase de conception, dans la mesure où ils évitent de s'engager dans des méthodes qui n'ont aucune chance de succès et pour lesquelles les coûts de développement sont élevés.
C'est pourquoi il est utile de faire :
- des **analyses d'usage** afin de cerner les pratiques et les besoins des utilisateurs, puis
- des **simulations** pour mettre une première maquette dans les mains d'utilisateurs, et enfin,
- des **tests** de performance, de robustesse et d'utilisabilité sur le système final (et sur les maquettes intermédiaires).

Ces trois phases posent le problème plus général de la conception de systèmes de dialogue que nous discutons dans cet article. Pour ces trois phases nous avons réuni des corpus de dialogue dans le domaine du renseignement touristique à savoir :
- un corpus pilote enregistré en situation réelle (maison de tourisme de Grenoble),
- un corpus en Magicien d'Oz (mettant en situation de laboratoire deux interlocuteurs),
- un corpus de dialogue homme-machine (en situation d'utilisation à la SNCF).

Nous montrons les apports respectifs de ces trois sources de données à la conception et à l'évaluation de systèmes de dialogue.

1. Introduction

L'évaluation des systèmes de dialogue homme-machine se heurte à plusieurs niveaux de difficulté :

a) du point de vue de l'usage

On se situe dans un contexte non stabilisé d'utilisation car les systèmes de dialogue homme-machine sont à la fois innovants et à la fois concurrents par rapport à d'autres systèmes interactifs

Ces systèmes n'ont pas encore une signification d'usage (on ne sait pas bien non plus à quels types d'usagers ils s'adressent) car ils sont innovants (on pense qu'ils apporteront des fonctionnalités nouvelles par rapport aux systèmes existants), mais on ne dispose pas vraiment de systèmes de dialogue commercialisés. Les seuls systèmes dont on dispose sont encore dans les laboratoires à l'état de développement.

Ces systèmes sont déjà, même avant leur commercialisation, en concurrence avec des systèmes existants utilisant des interfaces graphiques ou multimodales (bornes interactives de renseignement par exemple).

b) du point de vue de la conception

Ces systèmes n'ont pas été conçus la plupart du temps sur un cahier des charges précis. Soit ils sont encore immatures dans les laboratoires de recherche et il est donc impossible de les valider par rapport à des fonctionnalités pré-définies; soit il s'agit de systèmes existants que l'on tente d'adapter et d'intégrer à de nouvelles applications. Dans ce dernier cas des cycles évaluation / re-spécification / réévaluation sont possibles, comme dans le projet ARISE, mais demandent des moyens importants,

c) du point de vue de la complexité

Le dialogue met en jeu de très nombreuses variables et s'adresse à des usagers qui réagissent aux tests souvent de manière subjective. Quels sont alors les critères les plus importants : la fiabilité, la robustesse aux conditions de l'environnement, l'adéquation à la tâche, l'attrait de la communication verbale, la facilité d'usage, etc. ?

d) du point de vue de l'utilisabilité

Enfin le paradigme même du dialogue homme-machine n'est pas encore très clair: en quoi se distingue-t-il du dialogue humain ? Quelles sont ses spécificités ? Ses attraits ? Quand peut-on dire qu'un système est réellement utilisable ?

En fait on ne peut dire qu'un système " marche " que s'il est vendu et utilisé. Mais on voit bien *qu'ici ce n'est pas encore le cas*. C'est donc dans une

perspective de *diagnostic* que nous nous plaçons : nous cherchons à aider le concepteur à évaluer son système à l'aide de critères aussi objectifs que possible. Pour cela nous fondons notre approche sur un ensemble de corpus.

Le cadre d'application choisi dans l'ARC[1] B2 est le renseignement touristique (horaires de transport, réservation hôtelière, recherche d'itinéraires, etc.).

2. Cycle de développement d'un système interactif

Un système de dialogue homme-machine est un système interactif à interface vocale. Le Génie Logiciel a depuis longtemps proposé des méthodes de développement des systèmes informatiques qui ont été appliquées aux interfaces graphiques. Parmi ces méthodes, la méthode en V préconise de faire les tests de validation ou d'évaluation après chaque étape de développement (et si ce n'est pas possible, de définir au moins les *critères a priori* que devra satisfaire le système à la fin de chacune de ces étapes).

Le cycle de vie est une modélisation de la succession des étapes régissant la réalisation d'un système. Dans la plupart des approches, on identifie classiquement quatre étapes au cours de ce processus :

- la spécification externe : identification et formalisation du besoin, qui conduit à la rédaction du cahier des charges,
- la conception ou spécification interne, permettant de passer de la spécification externe à une représentation liée à des contraintes de réalisation (en fonction des matériels et logiciels choisis/disponibles). C'est lors de cette étape qu'est définie l'architecture système, ainsi que la nature des données transitant entre les différentes parties,
- la réalisation proprement dite, qui implique en dernier ressort le codage effectif sur la base de spécifications détaillées établies lors de l'étape précédente,
- enfin, l'intégration et les tests, par assemblage progressif des parties des constituants, puis des constituants eux-mêmes. Il est important de noter que du fait de son positionnement en aval des trois étapes précédentes, l'évaluation, ici mentionnée, ne permet de statuer que *globalement* sur le cycle de conception, et non pas sur un point d'une de ces étapes. C'est pourquoi, encore une fois la méthode en V préconise une évaluation de type diagnostic à chaque étape.

C'est par rapport à ce cycle que nous allons examiner l'apport des corpus.

---

[1] Evaluation des systèmes de dialogue oral. Action de Recherche Concertée, AUPELF-UREF.

3. Les corpus

Le but d'un corpus est de fournir des données objectives, communes pour tous les systèmes à évaluer. Comme nous l'avons dit ci-dessus, l'évaluation est étroitement liée à la conception. Dès la première étape on peut évaluer si le cahier des charges est cohérent et répond bien au besoin. Pour cela faut-il encore avoir étudié les usages et ciblé correctement les besoins. En dialogue homme-machine cela doit être le rôle du(es) *corpus-pilote*(s). Puis dans l'étape de conception, il faut déjà évaluer ce que sera le système final : pour cela on a recours à *des corpus-simulés*, enregistrés par exemple avec des techniques de magicien d'Oz. Dans le cycle de développement d'un système les versions successives sont évaluées à partir d'enregistrements de dialogues avec des usagers plausibles ou réels. L'évaluation prend en compte différentes données sociales et en particulier celles qui sont relatives aux usages possibles du système et de systèmes concurrents. Les évaluations portent sur le comportement technique des systèmes et sur les appréciations des usagers (confort, utilisabilité …).

On peut aussi penser à des corpus-tests préenregistrés contenant certaines difficultés qu'un système devrait surmonter. Comme il ne pourra s'agir de dialogues réels qui par définition ne peuvent être préenregistrés, seuls certains aspects du dialogue oral peuvent faire l'objet de tests à partir de corpus de ce type, par exemple la compréhension des énoncés du dialogue. C'est ce que prévoit la première partie du programme de l'action B2.

3.1 Le corpus-pilote pour l'analyse d'usage

3.1.1 Présentation générale et objectifs

Le but du Corpus-Pilote est de recueillir, pour une tâche générique de renseignements touristiques, des corpus réels (situation réelle H-H) linguistiquement riches, afin d'y mettre en évidence les comportements et les attitudes des usagers dans ce contexte ainsi que les phénomènes linguistiques et dialogiques inhérents naturellement à ce type de tâche. Les phénomènes visés sont donc de type linguistique (référenciation spatiale, référenciation temporelle, parole spontanée, etc.) et dialogiques (stratégies de dialogue, en particulier dans des situations de négociation ou de coopération, comportements tels que prise de parole et interruptions).

Les connaissances retirées de l'étude de ce corpus serviront à mettre au point les scénarios nécessaires à la réalisation des corpus ultérieurs (en Magicien d'Oz),

afin que lesdits phénomènes y soient présents. Ils serviront également d'exemplification pour le dialogue homme-machine ainsi que pour le choix des critères de test.

Il apparaît donc important pour ce Corpus-Pilote de se situer dans le cadre et dans la visée prévus pour les évaluations ultérieures: le renseignement touristique comme cadre d'application.

Les scénarios envisagés pour l'application-test portent sur:
- la négociation d'une chambre d'hôtel (emplacement, commodités, confort),
- la définition d'un emploi du temps pour compléter un agenda de rendez-vous (gestion du temps libre : durées de visite, des déplacements)
- le besoin de connaître les activités et les horaires d'ouverture des lieux publics ou de loisirs,
- l'obtention d'information sur les localisations, trajets et points de repères de ces lieux

Le Corpus-Pilote a été enregistré à la Maison du Tourisme de Grenoble, appelée ci-dessous l'agence.

3.1.2 Conditions expérimentales

Les clients et l'agent ne sont soumis à aucune consigne particulière. Les conditions d'enregistrement sont celles du bureau de l'agence. On dispose deux microphones directifs orientés l'un vers le client, l'autre vers l'agent et reliés à un enregistreur D.A.T. L'expérimentateur assiste à la prise de son et prend des notes sur le comportement des interlocuteurs, la situation, etc. Il s'assure du respect des règles déontologiques.

Les sessions ont été de 8 heures par jour, à raison de 5 jours consécutifs, du lundi au vendredi (compris). On a recueilli environ 40 h d'enregistrement après s'être assuré au préalable que les niveaux sonores étaient suffisants et que la qualité des données restait audible en moyenne.

Ces 40 h d'enregistrement ont été réduites aux parties utiles de la conversation par :
1) élimination des silences longs entre les transactions,
2) élimination des séquences inaudibles,
3) élimination des séquences accidentelles,
4) élimination des conversations mettant en scène des locuteurs non francophones,
5) élimination des conversations non autorisées.

On a obtenu alors un enregistrement utile d'une durée d'environ 12 heures.

3.1.3 Annotations

Les enregistrements sont segmentés en conversations, chaque conversation comprenant :

- l'ouverture de la conversation (ex : "*Bonjour ...*")
- la conversation entière (comprenant une transaction unique ou plusieurs transactions successives, suivie(s) de(s) la réponse(s))
- la clôture de la conversation (ex : "*Merci. Au revoir*.")

L'annotation permet de typer et de repérer la conversation. Cette annotation comprend :

- des lignes d'en-tête décrivant les conditions expérimentales et le chemin d'accès aux fichiers associés.
- des lignes de description concernant la conversation entière:
    - qualité sonore
    - stratégie de dialogue
    - phénomènes linguistiques
    - nombre de transactions
    - nombre de transactions prévues
    - nombre de transactions induites
    - nombre de transactions enchâssées *
- des lignes de description concernant chaque transaction:
    - nature
    - type d'objet
    - objet
    - portée spatiale
    - recours à un document ou à un geste
    - issue.

3.1.4 L'analyse d'usage

Qu'entend-on par analyse d'usage ?

C'est une approche du problème de la conception de systèmes (interactifs pour ce qui nous concerne) qui fait appel aux méthodes de l'ethnométhodologie (action située notamment), de la psychologie cognitive, de la psycho-sociologie et de l'ingénierie des systèmes. Elle s'appuie sur l'analyse du contenu d'usage à partir de tests auprès des usagers réalisés de deux manières :

a) par enquête directe,

b) par enquête indirecte.

L'enquête directe se fait en observant et en mesurant le comportement d'usagers en situation de travail ou en les questionnant pendant leur travail. L'enquête indirecte se fait en questionnant des personnes tierces.

A l'issue d'une analyse d'usage on obtient :
a) une signification d'usage,
b) des valeurs d'usage,
c) des profils d'usagers.

Il faut tout d'abord distinguer le client de l'usager, ce dernier n'ayant pas obligatoirement fait lui-même le choix du système qu'il utilise, ni participé à sa conception. Nous nous intéresserons dans la suite à cette catégorie d'usager. Un usager est un individu qui participe à la construction de la signification de l'usage ; celui-ci n'est pas prédéterminé par des catégories traditionnelles d'analyse sociologique. D'où la nécessité de procéder par enquêtes de type *interviews cliniques* (diagnostics).

La *signification d'usage* se construit sur 4 plans :
1- la confrontation aux techniques déjà utilisées par l'usager
2- la confrontation aux pratiques d'information et de communication coutumières de l'usager
3- la confrontation à l'identité sociale de l'usager (notamment son rôle, son statut, sa culture dans son interaction avec son environnement)
4- la confrontation à l'évolution de l'environnement de l'usager.

Selon le résultat de cette confrontation, le sens de l'usager se construit par la *négociation* des valeurs nouvelles ou par leur *imposition*. Cela se manifeste sur les quatre axes :
- banalisation / idéalisation
- hybridation / substitution
- identité active / identité passive
- accompagnement / rupture

La *valeur d'usage* se fonde sur l'image sociale qu'a un individu plongé dans un groupe social (la voiture peut être valorisante pour lui par exemple). Mais l'individu agit plus ou moins sur son environnement social à partir de ses préférences et de ses représentations. Ce double mouvement d'intériorisation et d'extériorisation fait ainsi émerger des valeurs et des systèmes de valeurs (les *profils*) auxquels se rattachent les individus.

3.1.5 Résultats obtenus

Nous avons concentré notre étude sur la réservation hôtelière et plus particulièrement sur deux groupes de variables :

• NN+D = Nombre de Nuitées et Dates (arrivée et départ) car ces variables sont liées dans les processus inférentiels que nous avons examinés,

• TC+FCI+FCE = Type de Chambre (simple, double, etc.), Facteurs de Confort Intérieur (bain, WC, etc.) et Facteurs de Confort Extérieur (vue, bruit, etc.) car ces variables sont liées entre elles pour les mêmes raisons.

Pour ces variables, nous avons observé que les processus inférentiels pour le calcul de leur valeur au cours du dialogue sont de complexité variable. Nous en donnons ci-après un aperçu pour la variable NN :

référence directe dans l'énoncé : " pour 3 nuits "
référence indirecte par la date : " pour le mardi 15 "
référence indirecte par liste de dates : " pour les 20, 21 et 22 "
référence par intervalle : " du 16 au 20 juin "
référence par durée : " pour la semaine " (on remarquera qu'une semaine de voyageur de commerce est de 5 jours et que celle d'un touriste est de 7 jours)
référence relative : " pour demain et après-demain "
référence avec arrière-plan : " pour le week-end de la Pentecôte "
référence dialogique : " non, une nuit de plus "
référence discursive : " j'arrive vendredi et je repars lundi "
référence implicite : la nuit où le client arrive à l'hôtel
référence indirecte implicite : " comme la semaine dernière "

On notera que ces références sont plus ou moins précises et/ou *ambiguës*. On pourrait faire les mêmes remarques pour la variable D : " en milieu de semaine ", " le 23, pour 3 nuits ", " j'arrive demain à 20h mais je ne prendrai la chambre que le lendemain ", " tout de suite ", " dans une semaine ", etc.

Pour le deuxième groupe de variables nous notons des énoncés encore plus complexes et qui prennent appui sur plusieurs tours de parole. Elles sont liées de manière plus forte ; par exemple un hôtel " Formule1 " est généralement près de l'autoroute (donc supposé bruyant) avec confort standard et prix type. On retrouve

les mêmes difficultés de traitement référentiel que pour le premier groupe de variables, auquel se sur-ajoute une difficulté de nature dialogique, dans la mesure où la portée référentielle couvre plusieurs tours de parole. Nous donnons ci-après quelques exemples de difficultés rangés par ordre croissant du nombre de tours de parole impliqués pour résoudre la référence :

1 tour de parole : " une chambre double, avec WC et bain, donnant sur la rue " avec toutes les variantes donnant souvent des ellipses, jusqu'à " une chambre sympathique "

2 tours de parole :

A : " pour combien de personnes ? "
C : " pour moi " ou " c'est pour M. et Mme Dupont " ou " une chambre avec 2 lits jumeaux "

C : " c'est pas trop bruyant ? "
A : " toutes nos chambres ont un double-vitrage " ou " la rue n'est pas passante "

C : " A ce prix-là, la chambre a un bain ? "
A : " oui et les toilettes "

3 tours de parole :

C : " je viens en vacances pour 8 jours "
A : " il vous faudra un accès facile sur le jardin... "
C : " oui et 2 lits jumeaux "

On pourrait également multiplier les exemples d'expressions connotées comme " de bon standing ", " ce que vous avez de mieux ", " je ne suis que de passage (=> pas besoin de télévision) ", " nous sommes loin de la gare (=> pas bruyant) ", " c'est la gare qu'on voit là-bas ? (=> c'est bruyant ?) ", " j'ai une grosse commande (=> je veux des prix) ", " je suis sur un fauteuil roulant (=> normes pour handicapés) ", " pour deux personnes, comme la semaine dernière ", etc. Voici un cas plus complexe encore à 5 tours de parole :

C : " vous reste-t-il des chambres ? "
A : " oui, pour 2 personnes seulement "
C : " est-ce beaucoup plus cher ? "
A : " 50 F de plus seulement "
C : " bon je vais la prendre quand même... "

dont on doit comprendre que le client est seul.

De cette étude nous pouvons conclure quelques valeurs d'usage et quelques profils d'usagers, et de là la signification générale d'usage :

Valeurs d'usage

voyage d'affaire ou transit / vacances (les différences sont marquées essentiellement sur les variables NN, FCI, FCE)

seul / famille (TC)

Profils d'usagers

habitué / occasionnel (marqué par le degré d'implicite pour l'habitué)

pour soi / intermédiaire (le degré de précision est plus grand de la part de l'intermédiaire)

personne âgée / jeune / handicapée (TC,FCI) (les jeunes cherchent des chambres peu chères, les handicapés des chambres spécialement équipées et les personnes âgées un confort intérieur comme ascenseur, télévision, toilettes, etc.).

Signification d'usage

la conception d'un système de dialogue oral pour la réservation hôtelière ou le renseignement touristique doit prendre en compte le fait qu'il existe déjà des modes institués et des technologies de communication comme le contact direct, le téléphone, les bornes interactives, les panneaux indicateurs, les agences et plus récemment les services sur Internet. Du point de vue des fonctionnalités, la technologie dite " serveur vocal touristique " ne sera pas innovante ; sa signification d'usage risque de s'hybrider avec des technologies existantes.

3.2 Les corpus simulés pour la conception

Qu'entend-on par simulation ?

Pour concevoir et modéliser le dialogue homme-machine, seule une méthode itérative est possible puisqu'on ne dispose pas de machine *a priori*. On est obligé de simuler totalement ou en partie des situations de dialogue homme-machine de plus en plus réalistes, en partant de dialogues humains enregistrés dans la phase précédente (si le système envisagé est totalement innovant cette phase n'est pas nécessaire, mais nous venons de voir le contraire dans le contexte d'application choisi). Le problème est donc maintenant de faire des choix dans le cahier des charges du système, pour intégrer au mieux les valeurs d'usage et les profils d'usager. Il est souhaitable de procéder par étapes en s'appuyant sur des corpora

de dialogue homme-machine simulé (par une technique dite de Magicien d'Oz par exemple), pour aboutir au dialogue homme-machine proprement dit.

Rien ne prouve d'ailleurs que l'étude du dialogue humain ou du dialogue simulé soit pertinente pour le dialogue homme-machine et justifie cette démarche. De leur côté, les éthnométhodologistes ne conseillent pas de recourir à de telles méthodes qui biaisent la communication. La psychologie expérimentale préconise au contraire des protocoles précis pour analyser les variables du problème. Les conditions de mise en situation sont bien connues en psychologie expérimentale et permettent d'isoler un certain nombre de variables au moyen d'expériences finement contrôlées. Dans le dialogue les variables sont nombreuses et il n'est pas envisageable d'espérer obtenir des données significatives à l'aide d'un seul type d'expérience. Généralement on peut distinguer des situations dans lesquelles :

• l'observateur est dans la boucle d'interaction et participe au dialogue — c'est le cas des méthodes de verbalisation ou d'élicitation de connaissances (on distingue en outre la verbalisation en cours de tâche de la verbalisation en dehors de toute activité),

• l'observateur est hors de la boucle d'interaction et ne participe pas au dialogue proprement dit. Cette deuxième méthode est apparemment la moins biaisée mais ne permet pas d'orienter la conversation, ce qui peut avoir deux effets opposés chez les sujets : la prolixité pour les uns, un certain mutisme pour les autres.

Les variables que nous pouvons retenir sont :
- la complexité de la tâche,
- *les usages langagiers*,
- la richesse sémantique du thème du dialogue,
- les rôles des partenaires dans l'interaction.

La première variable *complexité de la tâche* permet de mesurer l'influence de la tâche sur le dialogue (le raisonnement sous-tendu par la planification de la tâche induit-il des structures de dialogue particulières ? Ou inversement, en quoi le dialogue peut-il offrir un cadre structurant au raisonnement ? Retrouve-t-on des marqueurs, des points d'articulation, des stratégies utilisés dans la tâche pour le dialogue ?). La deuxième variable *richesse sémantique* permet de mesurer les effets des connaissances d'arrière-plan dans la compréhension d'un dialogue, et la troisième variable *rôle des partenaires* permet de relativiser les observations obtenues sur des dialogues à des groupes ou des classes d'individus ou d'usagers. Les consignes qui leur sont données en début de session expérimentale jouent en

effet pour une bonne part dans les résultats obtenus. Un entretien après la session est donc nécessaire avant l'interprétation des données enregistrées.

Les données que l'on peut recueillir à l'issue de ces simulations portent sur les points suivants :

- vocabulaire
- marqueurs de discours
- syntaxe
- concepts de la tâche
- activité et plan de la tâche
- évolution des connaissances partagées
- thèmes du dialogue
- buts
- ruptures et changement de monde de référence
- incompréhensions
- stratégies de dialogue

On dispose pour le français d'un certain nombre de corpus concernant des tâches de renseignement touristique : corpus "SNCF" (Du GDR-PRC Communication Homme-Machine, enregistré en 1985), corpus "Air France" (enregistré dans le cadre du projet Sundial en 1989), corpus "Maison du Tourisme" et "Réservation Hôtelière" (enregistrés respectivement en 1993 et 1996 par l'équipe GEOD, CLIPS-IMAG). Nous n'avons donc pas jugé utile dans le cadre de l'action B2 d'enregistrer de nouveau corpus. Chaque équipe est donc censée disposer de ces corpus à partir desquels il est facile d'établir des listes de critères d'évaluation à partir des données recueillies pour le système réel final à tester.

3.3 Le corpus-test pour l'évaluation de la compréhension

Dans le cadre de l'ARC B2, il est prévu d'enregistrer un corpus de dialogue homme-machine pour l'évaluation. Si l'on place un utilisateur devant un logiciel de dialogue et que l'on recueille ses énoncés et ceux de la machine, le corpus recueilli sera obligatoirement dépendant de la " machine " mise en face de l'utilisateur. On court le risque que le corpus ne soit pas assez générique pour tester des dialogues homme-machine diversifiés. D'autre part l'usager vraiment motivé va s'adapter à la machine et réduire de lui-même son champ d'action et l'étendue de ses requêtes: ainsi les phénomènes d'incompréhension vont disparaître petit à petit au fur et à mesure de l'apprentissage du système par l'usager. On ne saura donc pas en fin de compte si le système a un niveau de compréhension suffisant ou si l'utilisateur s'en contente et détourne la difficulté en s'appropriant le système d'une autre manière. En conséquence, les phénomènes

que l'on souhaite tester risquent d'être absents du corpus, ce qui revient à une certaine négation du corpus lui-même pour la tâche de test en question.

Le corpus-test qui nous intéresse ici devrait donc avoir un certain nombre de caractéristiques qui dépendent précisément des critères de tests d'évaluation que doivent remplir les systèmes de dialogue. Pour le moment les discussions méthodologiques sont en cours : elles s'orientent vers deux types de méthodes (a) soit procéder par test DQR, voir (Zeiliger et al., 1997), (b) soit par annotation des corpus de manière à évaluer les sorties du système à tester par rapport à des références types. Nous décrivons ci-après les contraintes induites par ce dernier type de méthode au niveau des annotations.

3.3.1 Une approche de l'évaluation de la compréhension

Deux niveaux principaux de compréhension peuvent en général être distingués dans les systèmes de dialogue.

Au premier niveau, la compréhension est littérale (CL) en ce sens qu'elle n'a pas accès aux informations sur l'état du dialogue; en revanche elle prend en compte les connaissances générales et statiques de la tâche.

Au deuxième niveau, la compréhension précise le sens en fonction de l'état du dialogue, qui est déterminé par l'historique des tours de parole et des événements liés à la tâche ou à l'usager. Sont évaluées à ce niveau les références (que le premier niveau avait peut-être programmé en terme de sens instructionnel, l'exécution étant différée au deuxième niveau) et les inférences mettant en jeu le contexte dynamique du dialogue.

Pour évaluer le premier niveau de compréhension la solution proposée actuellement consiste à annoter le corpus-test. Nous en suggérons ici les éléments essentiels (basés sur (Pérennou, 1997)).

Le corpus de test consisterait en couples (E, C) où E est un énoncé et C la représentation sémantico-pragmatique issue de CL. Pour beaucoup de systèmes, la représentation de C en tant que suite de structures de traits est de la forme ([Acte=a ; Prop=[r]])+. Elle prend en compte les actes illocutionnaires de dialogue et les actes propositionnels (prédicatifs et référentiels) définis par une sous-structure de traits [r]. Dans un énoncé il peut évidemment y avoir plusieurs actes illocutionnaires, chacun commandant une partie propositionnelle.

L'approche adoptée dans le projet ATIS (Air Travel Information System) d'ARPA (Advanced Research Project Agency), est un peu différente (voir par

exemple (Bates et al. 1992)) : la compréhension du système est jugée sur les informations délivrées à l'usager. Cette méthode à l'avantage de ne pas demander de consensus sur une représentation interne de la compréhension. Elle a l'inconvénient de ne bien prendre en compte que le traitement d'énoncés complets en eux-mêmes. De plus elle exige que les systèmes de dialogue soient connectés à une même base de données. Ces deux inconvénients, qui seraient préoccupant dans le cadre de B2, n'existent pas dans la méthode proposée ici.

a) Annotation en actes illocutionnaires

Dès lors que l'on a convenu d'un ensemble d'actes de dialogue pour l'application visée, l'acte sera choisi de manière adéquate dans cette liste. Par exemple (à titre indicatif) cet ensemble pourrait être {Assert, Question(attribut), Question(vérité), Question(accord)…, Confirmation, Acceptation, Dénégation, Refus, Rectification(attribut), Précision(attribut…) :

à quelle heure arrive-t-il? → [Acte=Question(HeureAr)]
non, pas celui-là → [Acte=Refus]
non, mais à 8 heures → [Acte=Dénégation] ;
[Acte=Rectification(horaire)…]

b) Annotation de la partie propositionnelle

[r] consiste en structures de traits de la forme [([attribut= [modifieur=Q1; (argi=vi)+])+] illustrées par les exemples suivants où I, Int, ~, < sont respectivement l'opérateur neutre et les constructeurs d'intervalles, d'ensembles flous et de demi-intervalles :

le train Paris Lille → [[LieuDép=[Mod=I;arg1="Paris"];
[LieuDest=[Mod= I; arg1="Lille"]]
départ entre 10 et 11 h. → [[HeureDép=[Mod=Int;arg1=10;arg2=11]]
départ vers 10 h. → [[HeureDép=[Mod= ~ ; arg1=10]]
arrivée avant 10 h. → [[HeureDép=[Mod= < ; arg1=10]]
départ dans la matinée → [[HeureDép="matinée"]]
départ le matin → [[HeureDép="matinée"]]

Dans les deux derniers exemples, l'énoncé réfère globalement à un ensemble de valeurs; cet ensemble a pour nom "matinée" et sa signification pourrait être la structure de traits suivante représentant un intervalle flou (IntFl) : [Mod=IntFl ; arg1=7; arg2=12]

Les références peuvent être négatives (exclusion de certaines valeurs). Dans ce cas le modificateur devra être précédé d'un symbole de négation qui ici sera '-' par exemple (étant entendu que -< est aussi ≥) :

pas avant 10 heures → [[HeureDép=[Mod= -<;arg1=10]]

Dans les modèles de compréhension en usage, les attributs peuvent correspondre aux rôles sémantico-pragmatiques introduits sous des formes diverses dans les modèles à base de schémas ou de structures de traits. Il ne paraît pas difficile de trouver un consensus sur le type d'annotation ci-dessus.

Une des différences entre les deux niveaux de compréhension tient au fait que certaines références (déictiques ou anaphoriques) ne pourront être déterminées sans la connaissances de l'état du dialogue. Cependant, dès le premier niveau, il est possible de mettre en place des instructions qui, exécutées en contexte, achèverons le calcul référentiel. Voici quelques exemples

départ demain soir	→	[[JourDép="demain"];[HeureDép="soir"]]
départ ce soir	→	[[JourDép="aujourd'hui"];[HeureDép="soir"]]
non, un peu plus tard	→	[[HeureDép="un_peu_plus_tard"]]

"soir" est traité de manière analogue à "matinée"; "demain" renvoie à l'opération date_courante+1; "un_peu_plus_tard" est une instruction qui recherche l'horaire précédemment indiqué et relance une requête en conséquence.

Pour évaluer le premier niveau de compréhension, il suffit de vérifier que les instructions ont bien été placées dans la structure de traits.

Remarque - Les actes prédicatifs sont inclus implicitement dans les attributs. L'attribut "HeureDest" par exemple contient les prédicats "aller à" ou "arriver à"… Dans un contexte de tâche bien déterminé cette simplification est souvent possible.

c) Exemples d'annotation complète

Montrons maintenant comment les actes illocutionnaires et propositionnels sont combinés dans le modèle d'annotation. Soit l'énoncé suivant :

Non non pas le train de Toulon… celui de Toulouse… qui part dans la matinée

Il contient trois actes : une dénégation commandant une référence négative, une rectification et une précision. Selon le modèle proposé l'énoncé serait annoté par C suivant :

[Acte=dénégation ; Réf=[LieuDest=[Mod=-I ; Arg1="Toulon"]]] ;

[Acte=rectification ; Réf=[LieuDest=[Mod=I; Arg1="Toulouse"]]] ;

[Acte=précision ; Réf=[HeureDép="matinée"]]

A noter le caractère facultatif de la référence négative; l'exemple est équivalent au suivant dont la première partie de la représentation se réduirait à [Acte=dénégation] —ce qui est nié devenant elliptique (et du ressort du deuxième niveau de compréhension) :

Non non … celui de Toulouse … qui part dans la matinée

d) Scores

Les énoncés du corpus test étant annotés, les scores d'un système pourront être obtenus en comparant les sorties du module de compréhension et les annotations. Les pourcentages suivants seront à considérer: omission ou insertion de référents, erreurs de substitution.

Il est possible de distinguer les scores portant sur les référents, sur les modificateurs et sur les actes.

Conclusion

Le recours aux corpus durant le cycle de vie du développement d'un système de dialogue n'est pas sans poser de problèmes méthodologiques et nécessite quelques précautions.

En ce qui concerne le corpus pilote, l'aspect évolutif, diachronique porté par toute réalisation doit être considéré. En effet, une réalisation à partir d'un corpus pilote peut modifier les valeurs et la signification d'usage. Par exemple, des processus inférentiels manquants peuvent amener à supprimer des composants de la valeur d'usage; ou au contraire des caractéristiques de réalisation (temps de réponse très rapides, fourniture d'information particulièrement pertinente) sont susceptibles d'introduire une part d'innovation et une nouvelle signification d'usage. Il serait donc intéressant de confronter en permanence durant le développement la signification d'usage prédite aux conséquences potentielles de la réalisation afin d'éviter de désagréables surprises à la mise en place du système.

L'enregistrement de tout corpus (de conception ou même d'évaluation) de dialogue s'effectue dans une situation en fait toujours compliquée, même si on tente d'en maîtriser le maximum d'aspects. En effet, les principes constitutifs et l'environnement d'un dialogue forment un ensemble complexe, dans lequel les

interactions entre composants sont fortes, quelquefois indirectes et peu perceptibles sans analyse très fine. Dans l'état actuel de nos connaissances sur les dialogues homme-machine et notamment sur l'alchimie cognitive et perceptive humaine mise en jeu durant la confrontation avec le système, les exploitations d'un corpus doivent se faire avec beaucoup de précaution : données, enseignements et conclusions (surtout pour des usages comparatifs) sont à relativiser et à ne pas généraliser trop rapidement. Ceci explique que d'une part il existe une assez grande quantité de corpus dans des domaines assez proches, créés parce que leurs concepteurs avaient perçus des difficultés d'exploitation de corpus pré-existants, et que d'autre part, certains corpus n'ont jamais pu être entièrement utilisés vis-à-vis du but déclaré pour leur création.

Dans tous les cas (corpus-pilote, corpus simulés, corpus-test) l'annotation est une opération a mener avec la plus grande circonspection pour deux raisons :

a) c'est une méthode lourde et longue qui est très coûteuse en temps et en moyens humains,

b) elle fige les données dans des théories, méthodologies ou principes valables à un instant donné et pour un objectif donné. Un système d'annotation est rarement neutre, souvent peu évolutif et les annotations dépendent en grande partie de l'annotateur lui-même.

Nous pensons donc qu'il est important de réfléchir à la généricité d'un corpus et à la notion de distance entre corpus (de même niveau ou de niveaux différents, à constituer ou déjà constitués), notions que l'on pourrait aussi appliquer par extension de la même façon aux systèmes. Elles permettraient d'évaluer le bien-fondé de certaines utilisations et/ou comparaisons, ou bien de montrer la nécessité, compte tenu par exemple d'une valeur de distance trop grande, de constituer d'autres corpus/systèmes selon les buts visés. Cette distance pourrait s'exprimer sur les constituants clés du système (différents selon le point de vue retenu - par exemple ergonomique: convivialité, technique: vocabulaire, grammaire) et pourrait prendre en compte les paramètres tels que: le domaine d'application, le type d'opération, les modalités d'entrée et de sortie, la stratégie de dialogue, les algorithmes de base...

Nous pensons que la mise au point d'une telle métrique devrait permettre d'affiner les méthodologies fondées sur les corpus avec un *effet de bord* intéressant: celui d'expliciter les principes sous-jacents aux systèmes et à la constitution de corpus.

Bibliographie des équipes

ARISE, (1996). Automatic Railway Information System for Europe, projet LE3-4229.

Bates M., Boisen S., Makhoul J., (1992). Developping an Evaluation Methodology for Spoken Language Systems. In : Proceedings of DARPA Speech and Natural Language Workshop, 102-8.

Bessac, M. & Caelen-Haumont, G. (1995). Analyses pragmatique, prosodique et lexicale d'un corpus de dialogue oral homme-homme. In : JADT 1995, Actes des IIIèmes journées internationales d'analyse de données textuelles, Vol. 1, Rome, 11-13 décembre, 1995. Rome : Eurograf 2000, p. 363-370.

Bessac, M., Colineau N., Caelen-Haumont, G. (1996). Actes de dialogue et prototypes mélodiques. In JEP'96, actes des Journées d'Etude sur la Parole, Avignon, 9-11 juin.

Caelen, J., A.L. Fréchet, A.L. (1992). Attitudes cognitives et actes de langage. Revue "Recherches sur la philosophie et le langage", n° 14, Vrin éd., Paris, pp. 19-48.

Caelen, J. (1994). Analyse de dialogue finalisés et simulés. Actes du séminaire TALN'94, GDR-PRC Communication Homme-machine, Marseille, avril 94, pp. 119-133.

Colineau, N. & Caelen, J. (1995). Etude de marqueurs dialogiques dans un corpus de conseption. in Le Communicationnel pour concevoir, J. Caelen & K. Zreik éds. Europ'IA, Paris, p. 203-222.

Fréchet, A.L. (1992). Analyse linguistique d'un corpus de dialogue oral homme-machine. Thèse Linguistique, Paris-Sorbonne.

Fréchet, A.L. & Caelen, J. (1993). Cognitive Attitudes and Speech Acts in Situations of Man-Machine Communication. Selected Proceedings of the 5rd International Conference Work With Display Units (WWDU'92), Berlin, Germany, H. Luczak, A. Cakir and G. Cakir eds., Amsterdam, Elsevier, pp. 340-344.

Larrey, P. & Pérennou, G. (1995). Report on conceptual level and assessment of the French version of the Philips understanding and dialogue components Deliverables 2222 and 2223 du contrat MLAP 63-036 MAIS.

Ozkan, N. & Caelen, J. (1994a). Design Issues for adaptative multimodal interfaces. ERCIM Workshop Reports, N. Carbonel éd., INRIA, pp. 89-98.

Ozkan, N. (1994b). Analyses communicationnelles de dialogues finalisés. Thèse SC, INPG Grenoble.

Pérennou, G. (1996). Lexique et Dialogue oral. Séminaire Lexique et communication parlée du GDR PRC CHM, pp.169-178.

Pérennou, G. (1997). Note sur l'évaluation des dialogueurs - La compréhension dans le dialogue oral spontané. AUPELF-UREF, ARC thème B2 (14p.).

Zeiliger, J., Antoine, J.Y., Caelen, J., (1997), Vers une méthodologie qualitative d'évaluation des systèmes de compréhension et de dialogue oral homme-machine, proc. JST-FRANCIL'97, Avignon, France.