Étude des situations problématiques dans le dialogue oral homme-machine

Caroline Bousquet

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 route de Narbonne– 31062 Toulouse Cedex 4, France
Tél.: +33 (0)5 61 55 72 01 - Fax: +33 (0)5 61 55 62 58
Mél: bousquet@irit.fr

ABSTRACT

This paper concerns the problematic of the error and unexpected situation handle in the framework of spoken dialogue systems. For this, we present an annotation methodology of dialogue problematic situations to study the system and user's behaviour when a problematic situation occurs. The methodology proposed was applied on 242 dialogues recorded in the framework of the ARISE project (spoken dialogue system about train schedules).

1. INTRODUCTION

L'intérêt des recherches sur les systèmes de dialogue oraux finalisés n'a cessé d'augmenter depuis la dernière décennie. Malgré l'amélioration indéniable des divers modules composant un système de dialogue, et tout particulièrement les moteurs de reconnaissance de la parole, ces systèmes de dialogue rencontrent encore trop de difficultés. Du point de vue de l'utilisateur, ces problèmes apparaissent généralement comme une mauvaise compréhension du système. Cette incompréhension peut être totale, dans le sens où le système n'a absolument rien compris aux propos de l'usager et doit lui demander de répéter. Elle peut être aussi partielle et apparaît alors comme une mauvaise compréhension d'une information particulière (par exemple la confusion d'une ville avec une autre) ; c'est ce que nous appellerons dans la suite de ce document une erreur. L'usager essayera alors de corriger l'erreur en faisant appel à un *sous-dialogue correctif* [4]. Ces incompréhensions sont dues à des mauvaises performances du système (le plus souvent engendrées par des erreurs de reconnaissance) ou bien par l'utilisation du système d'une manière trop différente des conditions de laboratoire (milieu trop bruité, mots inconnus…). Plusieurs recherches ont été menées dans l'objectif de détecter les mots inconnus et les erreurs de reconnaissance afin de limiter ce type d'erreurs [6], [7], [3], [5]. A ces incompréhensions, s'ajoutent tous les comportements inattendus de l'usager qui va compliquer la tâche du système (par exemple, lorsqu'il est ambigu dans ses propos ou qu'il fait des demandes hors-périmètres). Afin de les éviter, il est nécessaire que l'usager sache ce qu'il peut faire et ce qu'il ne peut pas faire. C'est la notion de *transparence* introduite par Karsenty dans [9]. Nous qualifierons toutes ces situations (incompréhensions et comportements inattendus) de *problématiques*. Lorsque le dialogue se trouve être dans une situation problématique, il est nécessaire de gérer cette situation afin d'en sortir le plus rapidement possible. Par exemple, en présence d'une erreur, il faut inciter l'usager à la corriger (voir [10]). Nos objectifs à long terme sont, d'une part, de limiter autant que possible la présence de ces situations problématiques et d'autre part, de proposer des stratégies de rémédiation. Dans le second cas, il est nécessaire de pouvoir identifier automatiquement la présence d'une situation problématique afin d'appliquer la stratégie de dialogue adéquate. Or cet automatisme du processus nécessite au préalable une étude linguistique et une annotation de ces situations. Peu de recherches ont été réalisées dans ce domaine : les annotations en actes de dialogue [13] peuvent concerner quelque unes des situations problématiques (désaccord, correction…) ; Hirschberg [8] a proposé une annotation des corrections et des *aware sites* (lorsque l'usager se rend compte de la présence d'une erreur). Nous exposons dans ce papier notre méthode d'annotation des comportements des deux interlocuteurs (le système et l'usager) lors de situations problématiques. Celle-ci a été appliquée sur un corpus de 242 dialogues enregistrés sur un système de dialogue oral homme-machine donnant les horaires de train [2]. Enfin, nous présentons les premiers résultats obtenus qui montrent comment les interlocuteurs réagissent face à une situation problématique. L'issue de ce travail préliminaire est une catégorisation des situations problématiques au travers des comportements des interlocuteurs.

2. MÉTHODOLOGIE D'ANNOTATION

2.1. Principes de base

Nous proposons d'annoter les dialogues en utilisant le formalisme XML. Nous avons pour cela créé trois principales balises pour délimiter un dialogue, un tour de parole du système et celui de l'usager. Nous verrons dans la suite comment ajouter des attributs à ces balises afin d'annoter les divers comportements et réactions du système et de l'usager. Afin de pouvoir interpréter aisément les réactions de l'usager, nous avons choisi de prendre les énoncés correspondant à ce qu'il a réellement dit et non la sortie du système de reconnaissance de la parole. Nous nous limiterons à l'annotation des tours de parole (systèmes et usager) jugés problématiques. La (Figure 1) montre un exemple de dialogue annoté.

2.2. Comportement du système

Nous distinguons plusieurs comportements qui peuvent éventuellement être raffinés en plusieurs sous-cas. Ces

distinctions seront retranscrites à l'aides des attributs *comportement* et *cas*. Certains peuvent être provoqués par le comportement de l'usager au tour de parole précédent qui lui-même peut parfois s'expliquer par le comportement précédent du système. Cet *historique des comportements* sera annoté, si besoin est, à l'aide des attributs *usr* pour le comportement de l'usager et *sys* pour le système. Il est apparu que l'historique remontant à l'échange précédent est suffisant pour l'annotation de notre corpus. Le format de l'annotation sera le suivant :

<sys comportement="xxx" usr="xxx" sys="xxx" cas="xxx">

De l'analyse du corpus de dialogue, nous distinguons les six comportements suivants :

- **Le système fait une erreur** – l'erreur peut être due au système lui-même (c'est le cas par défaut car de loin le plus fréquent) ou provoquée par un comportement particulier de l'usager (par exemple, si ce dernier s'est exprimé de manière ambiguë) qui sera alors annoté dans l'historique des comportements ;
- **Il ne comprend pas ou n'a pas entendu pas ce que dit l'usager** – cette incompréhension du système se traduit généralement par une **demande de répétition** explicite ou bien par la répétition (telle quelle ou en reformulant) de ce que le système avait dit lors de l'échange précédant et éventuellement accompagné d'une explication (du style "Je ne vous ai pas entendu") et / ou d'une formule de politesse. Ce comportement peut s'expliquer par le comportement de l'usager au tour de parole précédent (absence de réponse, réponse ambiguë…) ;
- **Il relance l'utilisateur après avoir détecté la présence d'une erreur** que l'usager n'a pas encore corrigée ; ce tour de parole fait alors partie d'un sous-dialogue correctif ;
- **Il poursuit son dialogue normalement sans tenir compte d'une intervention de l'usager** qu'il n'a pas compris ou ne peut pas traiter comme les demandes hors-périmètres ;
- **Il clôt le dialogue alors que l'usager ne s'y attend pas** – Nous distinguons alors trois cas : 1) le système croyait que l'usager avait fini (mais il a mal compris), 2) il abandonne le dialogue car il n'arrive décidément pas à comprendre/entendre l'usager (il peut alors lui conseiller d'utiliser un autre service afin d'obtenir les renseignements voulus), 3) le système se rend compte d'une erreur fatale (par exemple la base de données est inaccessible) ;
- **Il fait n'importe quoi :** Ce sont les "bugs" du système (et non des erreurs classiques de reconnaissance ou de compréhension) qui peuvent apparaître en particulier lorsque le système est en cours de développement.

2.3. Comportement de l'usager

Les comportements, éventuellement leurs cas particuliers et l'historique des comportements, sont annotés à l'aide d'attributs de la manière suivante :

<usr comportement="xxx" sys="xxx" cas="xxx">

Nous distinguons 12 comportements :

- **L'usager corrige** suite à une erreur du système (sous-dialogue correctif) ;
- **Il réfute** une information sans la corriger ;
- **Il ne corrige pas** alors que le système a fait une erreur et accepte cette erreur ;
- **Il ne répond pas** ou tout au moins il ne dit rien de pertinent – Cette absence de réponse peut avoir plusieurs explications : elle peut faire suite à une erreur ou à une incompréhension du système qui déstabilise l'usager, elle peut arriver en tout début de dialogue alors que l'usager est un peu perdu ou encore elle peut avoir lieu alors que le dialogue se passe sans écueil ;
- **Il est illogique** dans ses propos (par exemple, dans le cadre d'un trajet, les villes de départs et d'arrivée sont identiques) ;
- **Il est ambigu** dans ses propos ;
- **Ses croyances sont erronées** ;
- **Il fait une demande hors-périmètre** ;
- **Il fait une demande ingérable** pour le système – Nous appelons demandes ingérables les demandes que le système est capable de traiter mais pas dans l'état où il se trouve lors de cette demande. Par exemple, dans certains systèmes il n'est pas prévu de corriger une information qui a été précédemment confirmée. Ce type de comportement dépend des choix ergonomiques faits lors du développement du système ;
- **Il change d'avis** suite à des erreurs ou à des incompréhensions de la part du système ou bien de son plein gré ;
- **Il relâche ses contraintes** : comme pour le changement d'avis, ce comportement peut être provoqué par une erreur ou une incompréhension du système ;
- **Il abandonne le dialogue,** soit parce qu'il y a eu bien trop de situations problématiques, soit pour une autre raison a priori inconnue (peut-être lorsque l'usager se rend compte que c'est une machine qui lui répond ?).

<dialogue numero="XXX">
<sys> *Veuillez indiquer les villes de départ et d'arrivée du trajet souhaité* </sys>
<usr> *Paris Toulouse* </usr>
<sys comportement="erreur"> *Vous partez de Paris pour Mulhouse. En cas d'erreur corrigez-moi sinon indiquez le jour de votre départ* </sys>
<usr comportement="repond_pas" sys="erreur"> *euh* </usr>
<sys comportement ="dmde_repet" usr="pas_repondu" sys="erreur"> *Excusez-moi, je ne vous ai pas entendu. Veuillez répéter je vous prie.* </sys>
<usr comportement ="abandon" cas="erreur"></usr>
</dialogue>

Figure 1 : Exemple d'annotation.

2.4. Discussion

Nous remarquons que les principes d'annotation proposés sont parfois redondants entre l'annotation des comportements de l'usager et ceux du système. En effet,

les comportements d'un locuteur sont souvent dépendants du comportement de son interlocuteur. Nous avons cependant tenu à conserver cette redondance car elle permet de pouvoir ensuite analyser les corpus facilement en considérant chaque tour de parole de manière indépendante. Ce type d'annotation permet de dénombrer de manière automatique les tours de parole correspondant à un cas particulier.

3. RÉSULTATS

3.1. Description de l'application

Le domaine d'application concerne la demande d'informations sur les horaires de train de la SNCF. Le corpus est issu de l'enregistrement de dialogues réels recueillis sur la plate-forme DEMON développé à l'IRIT dans le cadre du projet européen ARISE [2]. DEMON a été développé à partir de la plate-forme de Philips [1] pour le français. Ce corpus comporte 242 dialogues correspondant à 3110 échanges. Le taux d'erreur de reconnaissance est très important (40.5%).

3.2. Déroulement général du dialogue

Environ 70% des dialogues aboutissent aux buts de l'usager. Cependant le nombre de dialogues que nous pourrions qualifier de *parfaits,* c'est-à-dire ne comportant aucune des situations problématiques recensées dans la section 2, sont très rares (3.31%). Près de 30% des dialogues sont abandonnés avant que l'usager n'obtienne les informations qu'ils souhaitent. Ces abandons se traduisent par le fait que l'usager ou le système raccroche avant la fin du dialogue. Dans environ 61% des abandons c'est l'usager qui met fin au dialogue suite à des erreurs ou à des incompréhensions du système ou encore sans raison apparente. Dans 39% des cas, c'est le système qui abandonne. La (Table 1) montre la répartition de ces abandons.

Table 1 : Répartition des abandons (72 dialogues).

L'usager abandonne		Le système abandonne		
À cause des erreurs	Raison inconnue	Il ne comprend rien	Il croit que c'est fini	Erreur fatale
48.61%	12.50%	25.00%	11.11%	2.78%

3.3. Comportement du système

Nous avons recensé 573 erreurs de la part du système soit en moyenne 2.37 erreurs par dialogue. Les causes de ces erreurs sont détaillées dans la (Table 2). Nous observons que généralement ces erreurs sont dues aux mauvaises performances du système et rarement à un comportement inattendu de l'usager. Le corpus comporte un total de 613 demandes de répétitions (soit en moyenne 2.5 fois par dialogues). Comme le montre la (Figure 2), les demandes de répétitions sont provoquées dans plus de la moitié des cas suite à un comportement inattendu de l'usager au tour de parole précédent (absence de réponse, réponse ambiguë ou illogique, demande hors-périmètres). Les incompréhensions classiques dues seulement au système représentent près de 45 % des demandes de répétitions.

Table 2 : Catégorisation des erreurs du système.

Uniquement due au système	97.20%
Suite à une réponse ambiguë de l'usager	1.05%
Suite à une demande ingérable	1.75%

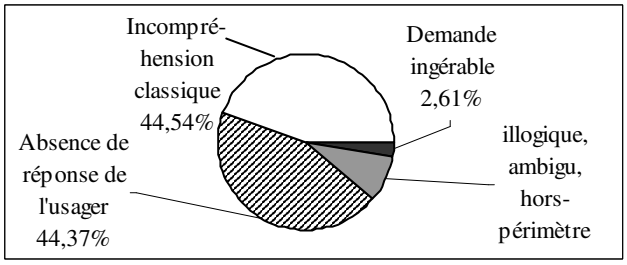

Figure 2 : Répartition des demandes de répétitions.

3.4. Comportement de l'usager

Les différents comportements problématiques de l'usager sont décrits dans la (Table 3). Nous observons, qu'excepté les corrections et les absences de réponses qui sont très fréquentes (10%), les autres comportements sont relativement rares. Enfin, il y a très peu de changements d'avis ou de relâchement de contraintes classiques c'est-à-dire non provoqués par des erreurs ou des incompréhensions de la part du système (6 occurrences).

Table 3 : Comportements de l'usager.

	Occurrences	Moyenne / tours de parole
Corrections	343	11.96%
Absence de réponses	287	10.01%
Réfutations sans correction	100	3.49%
Absence de correction	51	1.78%
Demandes ingérables	48	1.67%
Changement d'avis	42	1.46%
Ambigu	33	1.15%
Relâchement de contraintes	15	0.52%
Croyance erronée	11	0.38%
Demande hors-périmètre	7	0.24%
Illogique	6	0.2%

Nous avons cherché à savoir dans quelles situations l'usager ne répondait pas (Table 4). Dans plus d'un tiers des cas, il n'y a pas d'explication plausible car le dialogue se passait correctement avant l'absence de réponse.

Table 4 : Causes des absences de réponse.

Après une erreur du système	23.34%
Après une incompréhension du système	22.30%
En début de dialogue	17.42%
Autres	36.94%

L'objectif des concepteurs de systèmes est bien entendu de limiter les erreurs mais aussi d'inciter les usagers à corriger. En effet, un système capable de dialoguer ne se distingue pas par l'absence d'erreur mais par sa capacité à les gérer [12]. Aussi nous estimons qu'il est important

d'examiner si l'usager corrige ou non lorsqu'il y a une erreur. Le système de dialogue utilisé fait très régulièrement des demandes de confirmation implicites ou non (voir [10]) afin de permettre à l'usager de détecter les erreurs éventuelles. Nous observons (Figure 3) que dans plus des trois quarts des cas de demandes de confirmation lors d'une erreur, l'usager tente de corriger ou tout au moins réfute l'information erronée (appel à un *sous-dialogue correctif*). Cependant il arrive souvent que l'usager ne tente pas de corriger et accepte l'erreur (≈10%) ou bien change d'avis ou relâche ses contraintes en espérant se faire comprendre plus facilement (≈6%). De plus il arrive régulièrement que la présence de l'erreur trouble l'usager au point qu'il ne réponde rien (≈11%).

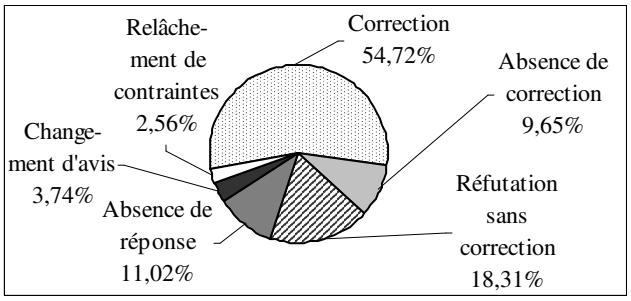

Figure 3 : Comportement de l'usager suite à une erreur.

4. CONCLUSION ET PERSPECTIVES

Nous avons présenté dans ce papier une méthodologie d'annotation des situations problématiques afin d'étudier les comportements des interlocuteurs (respectivement usager et système). Nous avons ensuite procédé à des analyses statistiques qui montrent que ces situations sont très fréquentes et malheureusement pas toujours bien gérées. Notre objectif à long terme est de détecter automatiquement les situations problématiques afin de définir la conduite de dialogue la plus appropriée. L'étape suivante consistera donc à étudier les phénomènes caractérisant ces situations : marqueurs lexico-syntaxiques (en nous inspirant de Morel [11]), présence accrue de phénomènes propres à la parole spontanée et d'erreurs de reconnaissance, contradiction avec les croyances et attentes du système… Cette analyse effectuée, nous serons alors en mesure de proposer une caractérisation et une modélisation de ces situations problématiques pour les identifier de manière automatique et améliorer alors la gestion du dialogue, ultime but de ce travail d'analyse.

5. BIBLIOGRAPHIE

[1] H. Aust, M. Oerder, F. Seide et V. Steinbiss. The Philips automatic trainTimetableinformation system. Speech Communication, 17, pages 249-62, 1995.

[2] P. Baggia et al.. Language Modelling and Spoken Dialogue Systems – the ARISE Experience. In *Proceedings of European conference on speech communication and technology*, Vol. 4, pages 1767-1770, Budapest, 5-9 septembre 1999.

[3] M. Boros et al. Semantic Processing of Out-of-vocabulary Words in a Spoken Dialogue System. In *Proceedings of European conference on speech communication and technology*, pages 1887-1890, Rhodes, 22-25 septembre 1997.

[4] C. Bousquet, R. Privat et N. Vigouroux. Error handling in spoken dialogue systems: toward corrective dialogue. In *Proceedings of ISCA workshop on Error Handling in Spoken Dialogue Systems*, pages 41-45, Château-d'Oex-Vaud, 28-31 août 2003.

[5] C. Bousquet et N. Vigouroux. Recognition Error Handling by the Speech Understanding System to Improve Spoken Dialogue Systems. In p*roceedings of ISCA workshop on Error Handling in Spoken Dialogue Systems*, pages 113-118, Château-d'Oex-Vaud, 28-31 août 2003.

[6] G. Chung. Automatically incorporating unknown words in JUPITER. In p*roceedings of International Conference on Spoken Language Processing*, Vol. 4, pages 520-523, pékin, 16-20 octobre 2000.

[7] T.J. Hazen et al. Recognition Confidence Scoring for Use in Speech understanding Systems. In p*roceedings of Automatic Speech Recognition – Challenges for the millennium*, pages 213-220, Paris, 18-20 septembre 2000.

[8] J. Hirschberg, M. Swertz et D. Litman. Labeling Corrections and Aware Sites in Spoken Dialogue Systems. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue,* pages 72-79, Aalborg, Aout 2001.

[9] L. Karsenty. Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology*, 5(2):147-158, 2002.

[10] A. Lavelle, M. de Calmès et G. Pérennou. Confirmation strategies to improve rates in a telephonic inquiry dialogue system, In p*roceedings of European conference on speech communication and technology,* Vol. 3, pages 1399-1402, Budapest, 5-9 septembre 1999.

[11] M.A. Morel. *Analyse linguistique d'un corpus de dialogues homme/machine - Tome 1*. Publication de la Sorbonne Nouvelle,1988.

[12] D. Perlis et K. Purang. Conversational adequacy: Mistakes are the essence. In *proceedings of AAAI workshop: Detecting, Repairing and Preventing Human-Machine Miscommunication*, pages 47-56, Portland, juillet 1996.

[13] A. Strent. The MONROE Corpus. *Technical report 728/TN 99-2*, mars 2000.