Traitement de désignations orales dans un contexte visuel

Ali CHOUMANE
IRISA/Cordial, Université de Rennes1
6 rue de Kerampont, BP80518, 22305 Lannion, France
choumane@irisa.fr

Résumé. Nous nous intéressons aux systèmes multimodaux qui utilisent les modes et modalités suivantes : l'oral (et le langage naturel) en entrée et en sortie, le geste en entrée et le visuel en sortie par affichage sur écran. L'usager échange avec le système par un geste et/ou un énoncé oral en langue naturelle. Dans cet échange, encodé sur les différentes modalités, se trouvent l'expression du but de l'usager et la désignation des objets (référents) nécessaires à la réalisation de ce but. Le système doit identifier de manière précise et non ambiguë les objets désignés par l'usager. Nous traitons plus spécialement dans cet article les désignations orales, sans geste, des objets dans le contexte visuel. En effet, l'ensemble du contexte multimodal, dont le mode visuel, influe sur la production de l'entrée de l'usager. Afin d'identifier une désignation produite en s'appuyant sur le contexte visuel, nous proposons un algorithme qui utilise des connaissances « classiques » linguistiques, des connaissances sur les objets manipulés, et des connaissances sur les aspects perceptifs (degré de saillance) associés à ces objets.

Abstract. We are interested about multimodal systems that use the following modes and modalities : speech (and natural language) as input as well as output, gesture as input and visual as output through displaying on the screen. The user exchanges with the system by a gesture and/or an oral statement in natural language. This exchange, encoded on the different modalities, contains the goal of the user and also the designation of objects (referents) necessary to the realization of this goal. The system must identify in a precise and non-ambiguous way the objects designated by the user. In this paper, our main concern is the oral designations, without gesture, of objects in the visual context. Indeed, the whole of the multimodale context including visual mode, influences the production of the user input. In order to identify a designation based on the visual context, we propose an algorithm which uses « traditional » linguistic knowledge, knowledge about manipulated objects and perceptive aspects (degree of salience) associated to these objects.

Mots-clés : communication homme machine multimodale, référence, saillance.

Keywords: multimodal human computer communication, reference, salience.

1 Introduction

Cette étude [1] se situe dans le contexte des systèmes de communication personne-machine multimodaux. Le but de tels systèmes est de permettre aux utilisateurs d'obtenir la réalisation de

[1] Ces travaux sont partiellement financés par le contrat 211-B2-9/ARED 1800 du conseil régional de Bretagne, France.

services. Par exemple actuellement des systèmes multimodaux sont conçus pour fournir des renseignements sur des horaires de vols aériens, pour élaborer des itinéraires ou encore pour aider la réalisation de maquettes ou de plans.

L'interaction entre l'utilisateur humain et le système pour la réalisation de service nécessite d'atteindre un consensus sur le but à réaliser. Cet accord concerne une compréhension mutuelle des intentions qui peuvent apparaître (et qu'il faut satisfaire) durant les différentes phases de l'interaction mais aussi une vue partagée sur toutes les entités (paramètres, objets, ...) manipulées et nécessaires à l'accomplissement de la tâche.

L'utilisateur désigne ces entités en recourant aux modes et modalités à sa disposition : oral, langue, geste... On parle d'activités référentielles. Un rôle important d'un système est de reconnaître et de comprendre ces activités référentielles.

Ce travail est ardu car le système est confronté à de nombreuses difficultés. La performance de l'utilisateur dans l'activité de désignation n'est pas sûre, elle peut être entachée d'ambiguïtés, d'erreurs, d'hésitation et conduit à des « bruits » ou des malentendus qui sont susceptibles d'être aggravés par les dispositifs matériels et les programmes du système. Enfin, bien que la multimodalité soit normalement utilisée pour améliorer la communication et diminuer le nombre d'ambiguïtés, l'usage conjoint de plusieurs modes multiplie les problèmes techniques et peut dégrader les performances des usagers.

Dans cet article nous nous intéressons aux entrées des systèmes multimodaux de communication personne-machine. Pour comprendre le but de l'usager, le système doit effectuer correctement la fusion des entrées qui parviennent des différentes modes. Un des points critiques de cette fusion est la résolution des expressions référentielles (ER). Nous proposons un algorithme de résolution de désignations orales aux objets du contexte visuel. Cet algorithme est fondé d'une part, sur des connaissances linguistiques liées à l'expression référentielle utilisée et d'autre part sur la saillance des objets du contexte visuel. Nous pensons que l'attention de l'usager peut être influencée par la présentation des objets dans le contexte visuel (notion de saillance) et nous prenons en compte cette notion dans le processus de résolution des références.

Après une section dans laquelle nous présentons notre contexte de travail et le problème traité, nous exposons les principaux éléments de la solution proposée. Nous commençons par analyser les différents cas possibles à prendre en compte dans la solution puis nous montrons la rôle de la saillance dans le traitement des ERs. Enfin nous détaillons l'algorithme de traitement des ces désignations que nous illustrons par un exemple d'application qui souligne l'intérêt de cet algorithme.

2 Problématique

Notre cadre de travail est le système prototype Géoral tactile (Siroux *et al.*, 1997) qui est implémenté sur la plateforme multiagent DORIS (L'Hour *et al.*, 2004). C'est un système multimodal pour une application de renseignements géographiques et touristiques. L'usager peut demander des informations et la localisation de sites touristiques comme plage, église, camping en précisant un endroit, une zone (dessinée ou située par rapport à un élément géographique ou cartographique particulier : rivière, route, côte) ; il peut également demander la distance et l'itinéraire entre deux localités.

Les modes et modalités mis à la disposition de l'usager dans le système Géoral sont les suivants :

– l'oral en entrée et en sortie du système. L'usager peut formuler ses demandes ou ses réponses aux questions du système par la voix et en langage naturel (LN) de manière spontanée (pas de consignes particulières d'élocution). Certaines réactions du système sont aussi transmises oralement à l'utilisateur.
– le mode visuel : le système affiche sur un écran une carte de la région ; cette carte contient des informations géographiques et touristiques habituelles : routes, rivières, fleuves. Des effets de zoom, de surlignage, de clignotement permettent au système de focaliser l'attention de l'usager.
– le mode gestuel par l'intermédiaire d'un écran tactile : l'usager peut désigner par différents types de geste des éléments sur la présentation affichée à l'écran.

Un dialogue avec Géoral est composé d'un ou plusieurs échanges. Un échange est constitué des tours de communication de l'usager et du système (Bilange, 1992). Un tour de communication de l'usager consiste en un énoncé oral et/ou une entrée gestuelle et celui du système consiste en une sortie orale (par la synthèse de la parole) et un affichage sur l'écran. Par exemple, un échange simple contient deux tours de communication : un tour de l'usager (question) et un tour du système (réponse). Notons qu'un échange peut être imbriqué en contenant d'autres échanges (dans le cas des questions de clarification). Dans un tour de communication d'un usager, le problème pour le système est de résoudre les ERs, c'est-à-dire, trouver le référent d'un symbole dans une modalité en utilisant des informations présentes dans la même ou dans d'autre modalité.

Dans ce cadre, nous avons proposé une nouvelle définition d'un modèle général de traitement des ERs (Choumane & Siroux, 2006). Ce modèle est fondé sur deux principes : la définition de langages associés à chaque modalité (LN, geste, visuel) plus un langage pivot et la création de fonctions reliant certains objets de chaque modalité permettant d'identifier les référents. A chaque tour de communication, les langages encodent les objets issus des modalités et mémorisent aussi certaines parties du contexte courant et de l'historique de l'interaction. Des traitements seront associés à chaque modalité (par exemple : traitement des anaphores pour la LN) et des traitements spécifiques seront mis en place pour la détermination des référents désignés de manière multimodale.

L'objet de l'article est de présenter une partie du traitement du modèle général de résolution des ERs. L'objectif est de trouver l'objet désigné dans le contexte visuel commun entre l'usager et le système (l'écran), et qui représente le référent d'une ER. Ce traitement est lié à deux des langages du modèle général : le langage qui encode le mode oral et celui qui encode le mode visuel.

Il existe plusieurs types d'ER : ERs qui font référence à des entités de l'historique LN comme l'anaphore (Mitkov, 2002), des ERs qui n'ont pas d'antécédents linguistiques parce qu'elles sont employées en première mention (Vieira & Poesio, 2000), (Manuélian, 2003) et/ou qui font référence à des objets dans une autre modalité qui correspond, par exemple, au contexte visuel dans le système Géoral. Ce dernier type d'ER est produit :
– conjointement avec un geste. Dans ce cas, ce sont des ERs à usage déictique dans lequel le référent est l'objet désigné par le geste.
– ou sans geste.

Dans cet article nous nous intéressons uniquement aux ERs en première mention qui sont produites sans geste. Nous discutons dans la suite les circonstances de production de telles ERs et une méthodologie de traitement.

3 Analyse et solution proposée

Nous traitons ici plus particulièrement le cas d'une désignation par l'oral (sans geste) d'un objet du contexte visuel commun entre l'usager et le système (il s'agit de l'écran dans le système Georal). Nous proposons une solution pour les entrées représentées génériquement par un expression régulière « je veux X (le long | à gauche | près de | à l'embouchure |...) de Y » où X $\in$ {les campings, les hôtels, ... } et Y $\in$ {la rivière, la route, ...}. Cependant la portée de cette solution est plus large, en effet, on trouve ce genre d'expressions référentielles dans d'autre tâches multimodales comme dans (Landragin, 2005) et (Qu & Chai, 2006) dans lesquelles on peut déplacer à l'aide de la LN des objets situés.

3.1 Différents cas possibles

Nous pouvons nous trouver face à plusieurs possibilités de référencement illustrées par les exemples 1, 2, et 3.

Exemple 1 je veux les campings le long de la rivière

Exemple 2 je veux les campings le long de la rivière le Léguer

Exemple 3 je veux les campings le long des rivières

Pour l'exemple 1, trois cas de figure sont envisageables :
- il n'existe aucune rivière dans le contexte visuel (pas de résolution). Dans ce cas la réponse à l'usager est une décision de dialogue. Un message d'information et une question de clarification seront suffisants.
- il existe une seule rivière dans le contexte visuel. Dans ce cas le référent de « la rivière » est l'objet sur l'écran qui représente la seule rivière existante.
- il existe n rivières ($n \geq 2$). Dans ce cas nous appliquons une stratégie fondée sur la saillance pour trouver le référent (détails ci-dessous).

Dans l'exemple 2, qui est un cas particulier de l'exemple 1, l'objet désigné est explicitement nommé dans l'énoncé. C'est-à-dire que l'objet désigné par « la rivière » n'apporte pas d'ambiguïté, il s'agit d'une précision par le nom de l'objet affiché sur l'écran qui est la rivière appelée « le Léguer ».

Dans l'exemple 3, l'ER « des rivières » porte la marque du pluriel, elle désigne les objets du contexte visuel de type « rivière ». Il n'y a pas d'ambiguïté et la résolution ne nécessite pas un traitement spécial autre que celui de la détection de l'ER lors de l'analyse syntaxique.

Les exemples 1, 2 et 3 [2] illustrent des débuts de dialogue dans lesquels les ERs sont employées en première mention. Au cours de dialogue, toute ER doit être évaluée en terme de coréférence (reprise) ou de première mention. Pour cette détermination on utilise les liens lexicaux (synonymie, hypernoymie, et méronymie) ainsi que la liste des items lexicaux représentant les objets sur l'écran. Notons que nous traitons les ERs dans un cadre applicatif précis qui permet de restreindre certains traitements.

La question qui se pose est pourquoi un usager est capable de demander des informations par une entrée multimodale aussi « vague » (exemple 1) et sans geste ? Il faut d'abord noté que

[2] Exemples extraits d'une expérimentation légère avec Georal (Siroux *et al.*, 1997).

malgré l'absence d'un geste conjointement à la parole, nous classons cette entrée comme multimodale. En effet, l'usager fait référence à un objet dans une modalité (LN dans l'exemple 1) en s'appuyant sur des informations présentes dans une autre modalité (les objets du contexte visuel dans l'exemple 1). D'une part, dans la production d'une entrée multimodale comme dans l'exemple 1, l'usager s'est appuyé sur le contexte visuel pour désigner son objet parce que cet objet est « suffisamment saillant », pour l'usager, que ce dernier n'a pas complété son entrée par un geste (notons aussi que l'usager peut être confronté à des problèmes de performance : crainte de toucher l'écran, ...). D'autre part, le système prépare et affiche le contexte visuel en fonction de l'importance applicative des objets et calcule donc leur saillance pour cet affichage.

3.2 Notations

La formalisation dans les prochaines sections est fondée sur les notations suivantes :

CVC_c = contexte visuel commun courant entre le système et l'usager.
e un échange donné entre l'usager et le système.
t un tour de communication donné. t de l'usager comprend normalement un énoncé oral et/ou une entrée gestuelle. Dans cet article, il s'agit uniquement d'un énoncé oral.

R={r_k, $1 \leq k \leq K$/ ER(r_k)}, le(s) ER(s) prononcée(s) par l'usager au tour de communication t.

O_k={o_j, $1 \leq j \leq J$}, l'ensemble des objets candidats référents de l'expression référentielle r_k.

$S_c(o_j)$ est la saillance de l'objet o_j dans le contexte visuel courant c (CVC_c).

Pour simplifier la présentation de l'algorithme (cf. section 3.4), nous supposons que dans un tour donné t d'un usager, |R| = 1 (K=1). C'est-à-dire que t ne contient qu'une seule ER qui désigne un objet dans CVC_c. Donc :
– r_1 est la seule ER dans le tour de communication t. Dans l'exemple 1, r_1 est « la rivière ».
– O_1 contient les objets candidats référents de l'expression référentielle r_1.

3.3 Saillance et son utilisation

La saillance intervient fortement lors de l'interprétation d'un énoncé en situation de dialogue ou lors de la compréhension d'un texte : en mettant en avant un élément, elle dirige l'attention sur cet élément et rend sa prise en compte prioritaire dans le processus de résolution des références et des coréférences (Landragin, 2005). Nous trouvons dans la littérature deux types de saillance : la saillance linguistique et la saillance visuelle.

La saillance linguistique, qui dépend uniquement de la modalité LN, constitue, par exemple, une aide à la résolution des anaphores (Lappin & Leass, 1994). Dans toute communication homme-machine dont le mode visuel fait partie, la saillance visuelle constitue un critère d'identification de l'objet désigné et perçu de manière prioritaire (Landragin, 2005).

Notre approche est fondée sur ce que nous appelons la « saillance contextuelle ». Nous visons ainsi une notion plus large que celle de la saillance visuelle. En effet, la saillance contextuelle d'un objet sera affectée durant l'interaction selon que l'objet est désigné ou non par l'usager tout en prenons en compte les caractéristiques visuelles des objets qui peuvent capter son attention. Désormais, nous faisons référence à la saillance contextuelle par le mot « saillance » uniquement. Nous allons montrer, dans la suite de cet article, comment cette saillance permet la

résolution d'une ER sans antécédent linguistique et sans geste accompagnant l'énoncé oral.

Nous distinguons deux moments d'utilisation de la saillance :
– au début d'un dialogue nous attribuons des valeurs par défaut aux objets du CVC. Cette attribution n'est pas nécessairement équiprobable, les valeurs peuvent être en effet liées à l'application. La détermination d'une méthode de calcul numérique de la saillance (Landragin, 2005) ne fait pas partie de cet article. Notons simplement l'existence de plusieurs facteurs qui contribuent à rendre saillant un objet et qui interviennent donc pour la quantification de la saillance de cet objet. Ces facteurs pourront être la couleur, la taille, la complexité, etc. d'un objet. Cette phase d'initialisation sera appliquée à chaque début de dialogue (au moins un échange).
– durant l'interaction, on modifie les saillances des objets qui sont dans O_1 à la fin de chaque interprétation de l'énoncé de l'usager. Ainsi la (les) saillance(s) de(s) référent(s) augmente(nt) et la (les) saillance(s) de(s) autre(s) objet(s) de O_1 diminue(nt). Nous rappelons que l'ensemble O_1 d'un tour de communication d'un usager contient l'objet (les objets) candidat(s) référents de r_1. Soit $S_c(o_j)$ la saillance de l'objet o_j dans le CVC_c. A la fin de l'interprétation de l'énoncé de l'usager, et après avoir utilisé les saillances de CVC_c, nous allons modifier les saillances des objets de O_1. L'interprétation par le système de tout énoncé de l'usager va prendre en compte l'ensemble des informations contextuelles (visuelles, linguistiques, ...).

Voici l'algorithme simplifié de distribution de saillance (a et b sont des constantes à régler, avec $a > 0$ et $b \leq 0$) :

si c'est un début de dialogue **alors**
  pour tout $o_j \in CVC_c$ **faire**
    $S_c(o_j) \leftarrow S_0(o_j)$ (initialisation des saillances)
  fin pour
sinon {c'est la fin de l'interprétation d'un énoncé de l'usager et nous disposons d'un ensemble O_1. Nous allons modifier les saillances des objets de O_1}
  pour tout $o_j \in O_1$ **faire**
    si o_j est un référent **alors**
      $S_c(o_j) \leftarrow S_c(o_j) + a$
    sinon {c'est-à-dire que c'est un objet à pénaliser}
      $S_c(o_j) \leftarrow S_c(o_j) + b$
    finsi
  fin pour
finsi

Cet algorithme est appelé par l'algorithme de recherche du référent ci-dessous (cf. section 3.4). Si l'algorithme de distribution de saillance a un ensemble O_1 en entrée, alors les saillances de ces objets o_j seront modifiées d'une valeur a ou b selon que l'objet est un référent ou non. Les constantes a et b seront réglées lors d'expériences ultérieures à mener.

3.4 L'algorithme de recherche de référent

L'algorithme que nous proposons (organigramme de la figure 1) est constitué de plusieurs phases :

– La première phase est celle de détermination de l'ensemble O_1 des objets candidats référents de r_1. Le critère de candidature est le type de l'objet. Dans l'exemple 1, O_1 est l'ensemble

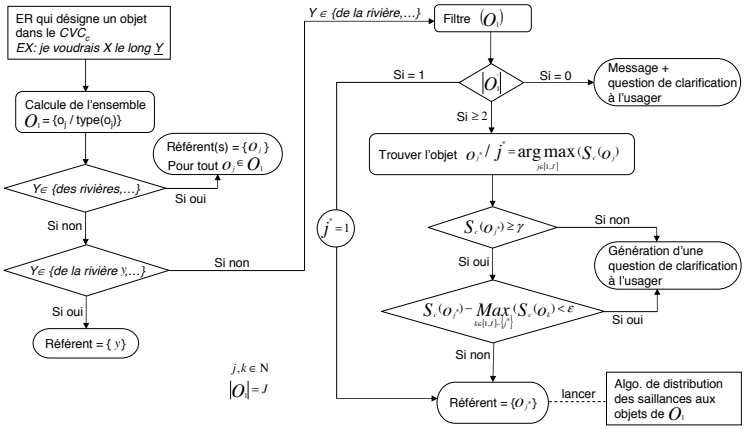

FIG. 1 – Algorithme de recherche de référent, dans le contexte visuel, désigné par l'oral

des objets de type « rivière » et de tous les objets dont leur type appartient à la classe des relations lexicales de « rivière » comme les synonymes, les hypernoymes, et les méronymes. Ces connaissances de synonymie, d'hypernoymie, et de méronymie peuvent être trouvées dans une base de données lexicales comme WordNet lors de l'analyse syntaxique de l'énoncé. Donc O_1 contient les objets de types « rivière », « fleuve », « ruisseau », etc. Pour prendre en compte des éventuels problèmes liés aux performances de l'usager (le cas d'un usager qui prononce « rivière » pour désigner un « ruisseau », ...), nous avons choisi de déterminer O_1 selon la classe lexicale entière. Cependant d'autres choix sont possibles :

1. Dans le cas d'objets qui ne sont pas explicitement typés sur l'écran [3], la détermination de l'ensemble O_1 selon toutes les informations lexicales est nécessaire. Dans l'exemple « je veux les hôtels le long de la route », l'ensemble O_1 contient tous les objets o_j de CVC tel que le type de o_j est égal à « route », « voie », etc.
2. Dans le cas des objets typés explicitement sur l'écran (le cas des rivières dans Géoral) d'une manière efficace qui garantit une meilleur visibilité de ces types, la détermination de l'ensemble O_1 pourrait dépendre uniquement du type de l'ER, d'une partie de sa classe lexicale, etc.

– Dans la deuxième phase nous testons si l'ER en question correspond à un cas particulier tel que ceux des exemples 2 et 3. Dans les deux cas, le(s) référent(s) est (sont) l' (les) objet(s) directement nommé(s) dans l'énoncé oral ou l'ensemble O_1 entier (cf. section 3.1).
– Si l'ER n'est pas un cas particulier (c'est alors l'un des cas de l'exemple 1), nous appliquons un filtre sur O_1. Le critère de filtrage dépend du type des objets. Par exemple, « la qualification » forme le critère de filtrage pour les objets de type rivière. Dans le cas de l'énoncé, « je veux les campings le long de la grande rivière », nous filtrons O_1 en ne choisissant que « les grandes rivières ». Dans le système Géoral, les caractéristiques des objets du CVC comme la taille, la couleur, etc. sont codées en représentation interne.
– Ensuite, nous testons le nombre d'élément de O_1, et trois cas sont à prendre en considération :

[3] Les routes dans le système Géoral ne sont pas étiquetées par « route » sur l'écran, mais par D11, D21, etc. Mais, dans la représentation interne du système, D11, D21, etc. sont codés comme route, voie, etc. Les choix des représentations sur l'écran sont liés à des problèmes de clarté, d'efficacité d'affichage, et d'usage.

1. Si $|O_1| = 0$, alors le message d'information et la question de clarification seront lancés. Si la question est accompagnée d'affichage graphique spécial (zoom, clignotement, etc) les saillances des objets mis en cause par le système seront modifiées. Notons que dans ce cas, il y aura imbrication d'échange.
2. Si $|O_1| = 1$ alors le référent de r_1 est l'objet représenté par o_1.
3. Si $|O_1| \geq 2$ alors nous recherchons l'objet le plus saillant parmi les objets de O_1, c'est l'objet o_{j^*} tel que :

$$j^* = \arg\max_j S_c(o_j)$$

$$avec\ S_c(o_{j^*}) \geq \gamma\ et\ S_c(o_{j^*}) - \max_{k \in [1,J]-\{j^*\}} S_c(o_k) \geq \epsilon$$

γ est un indice de confiance, il est nécessaire pour éviter le choix d'un objet peu saillant. Il dépend de la moyenne des saillances de tous les objets de O_1 et peut être calculé par :

$$\gamma = \frac{\lambda}{J} \sum_{j=1}^{J} S_c(o_j)$$

avec λ est un réel. Nous comptons affiner le calcul de γ par des expériences ultérieures à mener. ϵ est un réel suffisamment grand pour dire que r_1 n'a désigné que l'objet qui a la plus grande saillance o_{j^*}. C'est pour détecter les cas d'ambiguïté.
Si

$$S_c(o_{j^*}) < \gamma\ ou\ S_c(o_{j^*}) - \max_{k \in [1,J]-\{j^*\}} S_c(o_k) < \epsilon$$

Alors le système génère une question à poser à l'usager pour choisir un des objets O_1. Et après la résolution de l'expression référentielle en question, on modifie les saillances des objets de O_1 en faisant appel à l'algorithme de distribution de saillance montré ci-dessus (cf. section 3.3).

Notons qu'après avoir trouvé l'objet désigné, nous devons prendre en compte la préposition prononcée avant Y (Vandeloise, 1986) (« le long de » dans l'exemple 1) pour trouver la partie exactement visée par l'usager comme étant l'espace physique de recherche du thème (les campings dans l'exemple 1). L'influence de cette préposition pourrait être plus intéressante dans le cas de « à l'embouchure de », « à gauche de », etc.

3.5 Scénario d'application

Nous présentons dans cette section un exemple de dialogue entre un usager (U) et le système Georal (S). Les énoncés de U et de S qui sont en italique, correspondent respectivement à des entrées orales et à des sorties orales (par synthèse de la parole). Notons que U_t (S_t) correspond à un énoncé de U (S) au tour de communication numéro t.
Échange $e = 1$
U_1 : *Quel est le nom de ça + geste de pointage sur l'écran*
S_1 : *Bois de Lann ar Waremm*
Dans cet échange, qui se déroule dans un contexte visuel courant CVC_c, l'expression référentielle « ça », à usage déictique, a comme référent l'objet désigné par le geste démonstratif qui accompagne l'énoncé oral. C'est l'objet de CVC_c qui représente le *Bois de Lann ar Waremm*.
Échange $e = 2$
U_1 : *Je veux les campings le long de la rivière*

Le CVC_c de ce tour de communication est issu du contexte visuel précédent après certaine modification dans son contenu comme les saillances des objets. L'ER « la rivière » ne peut pas être résolue linguistiquement, il s'agit d'une première mention qui n'a aucun antécédent dans le discours. Pour cela, nous appliquons l'algorithme proposé ci-dessus pour trouver l'objet référent du CVC_c de l'ER « la rivière ».
r_1 = *la rivière*.
L'ensemble des candidats référents de r_1 est $O_1 = \{o_1, o_2, o_3, o_4\}$, avec o_1 = *Le Léguer*, o_2 = *Ruisseau de Gruguil*, o_3 = *Ruisseau de Kerhuel*, et o_4 = *Rivière des Traouïero*.
Nous remarquons qu'il n'y a pas de qualification de r_1, donc O_1 ne change pas après la phase de filtrage.
Comme $|O_1| = 4$, alors nous allons chercher l'objet o_{j^*}.
Supposons que dans CVC_c nous avons les données suivantes :

$$\begin{cases} S_c(o_1) = 2,\ S_c(o_2) = 1,\ S_c(o_3) = 1,\ et\ S_c(o_4) = 1.5 \\ \lambda = 1 \\ \epsilon = 0.3 \end{cases}$$

alors $j^* = 1$ et $\gamma = 1.375$.
Nous passons maintenant à la phase de validation de l'objet o_1 comme référent. Comme

$$S_c(o_{j^*}) = S_c(o_1) = 2 > \gamma\ et\ S_c(o_{j^*}) - \max_{k \in \{2,3,4\}} S_c(o_k) = S_c(o_1) - S_c(o_4) = 0.5 > \epsilon$$

alors le référent de « la rivière » est l'objet o_1 (le Léguer).
S_1 : *voulez vous les campings le long de la rivière le Léguer ? Merci de confirmer. Si non, veuillez préciser une autre rivière.*
U_2 : *oui*
S_2 : *il y a un camping qui répond à votre requête : camping de Beg Léguer*.

Après la confirmation de l'usager, l'expression référentielle r_1 est considérée, par le système, comme résolue. Ainsi la saillance de l'objet o_1 sera augmentée d'une valeur a. Les autres objets de O_1 (o_2, o_3, et o_4) seront pénalisés en baissant leur saillance d'une valeur b puisqu'ils n'étaient pas désignés.

L'échange numéro 2 n'est pas simple, il contient un sous-échange (S_1, U_2) mené par le système qui pourrait affecter aussi les saillances de O_1. Notons que les saillances des objets du CVC_c seront réinitialisées dès le premier échange du prochain dialogue.

4 Conclusion

Nous avons proposé une solution pour le traitement des désignations orales des objets dans le contexte visuel commun entre l'usager et le système. Cette solution est fondée sur des connaissances sur la modalité langue naturelle, des connaissances sur les objets manipulés, et des connaissances sur les aspects perceptifs (degré de saillance) associés à ces objets. Elle met en valeur un algorithme constitué de plusieurs phases. La phase générale consiste à chercher le référent le plus saillant dans une liste des candidats référents (appelée O_1) avec un indice de confiance et une vérification de validité pour détecter les cas d'ambiguïté. La terminaison du processus de résolution provoque, quelque soit le résultat, la modification des saillances pour le

tour de communication suivant. Nous pensons affiner certains paramètres en menant des expériences avec le système Georal. L'état du module de reconnaissance de la parole de Georal ne permet pas actuellement de mener des expérimentations de validation.

L'apport principal de notre proposition réside dans les possibilités de stratégie de dialogue offerte au système pour faire préciser les référents en cas de problème. En effet, si la liste des objets candidats référents est relativement importante, nous pensons qu'il est fastidieux de proposer à l'usager de choisir un des objets de la liste en lui présentant tous les objets. Le choix de proposer l'objet le plus saillant à l'usager est issu du fait que ce dernier s'est appuyé sur le contexte visuel dans son activité de désignation.

Remerciements

L'auteur remercie Jacques Siroux pour sa collaboration et son aide précieuse pour les travaux exposés.

Références

BILANGE E. (1992). *Dialogue personne-machine. Modélisation et réalisation informatique*. 2-86601-324-7. Hermes.

CHOUMANE A. & SIROUX J. (2006). Toward a generic model including knowledge and treatments for multimodal reference resolution. In V. P. GUERRERO-BOTE, Ed., *Proceedings Inscit2006*, volume 2, p. 298 – 302, Mérida - Spain.

LANDRAGIN F. (2005). Traitement automatique de la saillance. In *Douzième conférence sur le traitement automatique des langues*, p. 263 – 272.

LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**, 535 – 561.

L'HOUR J., BOËFFARD O., SIROUX J., MICLET L., CHARPENTIER F. & MOUDENC T. (2004). Doris, a multiagent/ip platform for multimodal dialogue applications. *ICSLP*.

MANUÉLIAN H. (2003). *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes*. PhD thesis, Université Nancy 2.

MITKOV R. (2002). *Anaphora Resolution*. 0-582-32505-6. Pearson Education.

QU S. & CHAI J. Y. (2006). Salience modeling based on non-verbal modalities for spoken language understanding. In *ICMI '06 : Proceedings of the 8th International Conference on Multimodal Interfaces*, p. 193–200, New York, NY, USA : ACM Press.

SIROUX J., GUYOMARD M., MULTON F. & RÉMONDEAU C. (1997). Multimodal references in georal tactile. In *Workshop Referring Phenomena In a multimedia Context And Their Computational Treatment, 35th Meeting Of the ACL*, Madrid - Spain.

VANDELOISE C. (1986). *L'espace en français*. Editions du Seuil, Paris.

VIEIRA R. & POESIO M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, **26 (4)**, 539–593.