FUSION DE CONTRAINTES POUR LA SYNCHRONISATION DES MODALITES ET POUR LA RESOLUTION DES REFERENCES DANS UN ENONCE MULTIMODAL

Frédéric Landragin, Bertrand Gaiffe, Nadia Bellalem et Laurent Romary

LORIA — UMR 7503
Campus scientifique, BP 239
54506 Vandœuvre-lès-Nancy
{landragi,gaiffe,nbell,romary}@loria.fr

Résumé *— Avant de résoudre les références aux objets et aux lieux d'un énoncé multimodal associant parole et geste de désignation, il est nécessaire de mettre en correspondance chaque geste avec le segment linguistique qui lui correspond. Pour cela, nous exprimons les contraintes syntaxiques, sémantiques et pragmatiques apportées par (a) les expressions référentielles et (b) les gestes, compte tenu de la disposition des objets et de leurs éventuels regroupements perceptifs. Une formalisation logique de ces contraintes et leur fusion permet d'aboutir à des hypothèses de correspondances puis à l'identification des référents. Nous exprimons alors la forme propositionnelle de l'énoncé permettant au système de retrouver les fonctions de l'application à exécuter. Malgré les nombreuses hypothèses faites sur chaque expression prise individuellement, nous montrons que le contenu propositionnel dans son ensemble est souvent le même. Cette méthode nous semble particulièrement prometteuse, tant parce qu'elle permet d'identifier les facteurs précis intervenant dans l'intégration de la langue et du geste, qu'en ce qu'elle permet d'envisager très facilement une implantation informatique.*

1. Introduction

Nous nous plaçons dans le cadre de l'interaction multimodale spontanée comportant des énoncés oraux en langue naturelle et des gestes de désignation, le geste venant en complément de la parole pour faciliter les références aux objets de l'application (éventuellement à des lieux). La compréhension de ces références se base sur un contexte d'ordre applicatif, perceptif et dialogique.

Nous nous appuyons sur un corpus multimodal établi suite à une expérimentation de type magicien d'Oz. Cette expérimentation, appelée Magnét'Oz [14], consistait à placer les sujets face à une simulation d'interface homme-machine dotée d'un microphone et d'un écran tactile. Les sujets étaient invités à s'exprimer et à utiliser le stylet de manière parfaitement naturelle, ceci dans un *contexte applicatif* très ciblé imposant le rangement d'objets dans des boîtes. Chaque boîte n'acceptait qu'un type particulier d'objet (au sens de leur forme visuelle), entraînant ainsi un regroupement catégoriel des actes de référence. Les formes des objets étaient volontairement complexes et abstraites afin d'empêcher une nominalisation directe. L'accent était ainsi porté sur l'utilisation du geste de désignation dans le *contexte perceptif* : les objets étaient répartis en suivant ou en contrecarrant les principes de groupement de la Gestalttheorie (proximité, similarité, continuité dans la disposition), ce qui incitait à une grande variété de gestes. Quant au *contexte dialogique*, il consistait à simuler la gestion d'un historique du dialogue permettant d'interpréter les ellipses et les anaphores.

L'objectif de cet article est d'exprimer sous une forme logique des contraintes précises liées à l'intégration des informations venant des trois types de contexte.

2. Synchronisation des modalités

Avant de résoudre la ou les références incluses dans l'énoncé courant, il est nécessaire de synchroniser les modalités, c'est-à-dire d'associer chaque geste à l'expression référentielle (ER) qui lui correspond sémantiquement. Ce problème peut s'avérer très complexe lorsque les situations suivantes, tirées du corpus Magnét'Oz, sont prises en compte :

1. Pour la plupart des références multimodales, ni le geste seul ni l'ER seule ne permettent d'identifier les référents. Les sujets utilisent les deux modalités de manière complémentaire, émettant ainsi des énoncés et des gestes simples mais par conséquent souvent imprécis ou ambigus. Chaque modalité, à elle seule, porte donc peu d'informations sémantiques utiles pour notre problème.
2. Le type du geste (c'est-à-dire les caractéristiques de sa trajectoire) ne permet pas d'identifier l'ER qui lui correspond. Par exemple, un geste de pointage peut désigner un lieu, un objet ou un groupe d'objets, et dans ce dernier cas peut présenter une ambiguïté de portée. C'est ici que le contexte perceptif va jouer un rôle primordial (cf. [4], [14] et [15]).
3. Le type de l'ER (groupe nominal défini, démonstratif, etc.) ne permet pas d'affirmer si un geste lui est associé ou non. Un démonstratif peut en effet être utilisé sans geste lors d'une anaphore, ou avec un geste lors d'une référence démonstrative (au sens d'Anne Reboul [11]). Comme nous le montre notre corpus de référence, une description définie peut de même être associée ou non à un geste (si c'est le cas, le caractère démonstratif de la référence résultante est porté par le geste, éventuellement par un marqueur déictique).

4. Plusieurs gestes peuvent correspondre à une même ER *(par exemple : « ces objets » + trois gestes, un par référent)*. Plusieurs ERs peuvent correspondre à un même geste *(par exemple : « cet objet et celui-ci » + un geste entourant lentement les deux référents)*. Dans certains cas, plusieurs gestes et plusieurs ERs se combinent pour constituer une même référence indissociable.

Dans les travaux existants, la mise en correspondance des gestes et des ERs se fait essentiellement sur un critère de proximité temporelle. David McNeill [8] pose l'hypothèse que la phase préparatoire du geste précède le segment linguistique accentué, et que sa phase significative précède ou se termine au moment de l'accent maximum. Dans le domaine de l'ingénierie des interfaces, l'aspect prosodique est parfois négligé et cette hypothèse se retrouve souvent associée à des seuils déterminés expérimentalement, comme dans [2], [10] et [6]. Sharon Oviatt montre ainsi que le geste intervient dans un intervalle de trois à quatre secondes avant l'ER, ces valeurs observées dans [10] étant reprises dans [6] pour un algorithme de synchronisation et d'intégration multimodale. Les autres critères généralement cités pour l'intégration multimodale sont la complémentarité logique et la compatibilité de type (cf. [2] et [3]). Ils sont néanmoins souvent utilisés dans un but de gestion des manques d'informations. On retrouve en particulier ce principe dans les interfaces pour lesquelles le geste peut venir prendre la place de la parole (par exemple : *« colorie » + geste de désignation d'objets + « en rouge »*, exemple tiré de [3]).

Dans notre cadre de l'interaction spontanée et en particulier dans le corpus Magnét'Oz, le geste vient toujours en complément de la parole : on a systématiquement une trace linguistique associée au geste. Reste le critère de proximité temporelle : on observe dans le corpus que les hypothèses de McNeill et les différentes valeurs de seuil proposées ne sont pas toujours respectées. Ceci est peut-être dû à l'utilisation du dispositif de vis-à-vis qu'est l'écran tactile, ou au contexte applicatif très fort qui autorise une certaine imprécision. En tout cas, même lorsque les écarts temporels sont importants, les énoncés restent toujours compréhensibles. Un système doit donc comprendre ce type d'énoncés et nous partons de l'hypothèse que la proximité temporelle n'est pas un critère prépondérant pour la mise en correspondance des gestes et des ERs. Par contre, suite à l'observation que l'ordre des gestes est identique à celui des ERs dans tous les énoncés du corpus, nous partons de cette hypothèse qui nous semble en accord avec l'aspect naturel de l'interaction.

3. Intégration des modalités

On peut envisager au moins deux types de stratégies de combinaison des informations portées respectivement par les gestes de désignation et par la langue :

1. La première solution consiste à calculer pour chaque ER présente dans l'énoncé un ensemble de candidats référents ; à effectuer le même travail en ce qui concerne les candidats aux gestes de désignation ; puis à tenter la fusion ER par ER. Chaque fois que pour une ER donnée, l'intersection entre l'ensemble des hypothèses langagières et l'ensemble des hypothèses gestuelles se réduit à un singleton, on considère qu'on a obtenu le coréférent entre langue et geste pour l'expression considérée. Cette solution, utilisée dans beaucoup de systèmes, conduit à des architectures logicielles caractérisées par la présence d'un interpréteur autonome et indépendant pour chaque modalité (les avantages souvent évoqués sont l'extension facile du système à d'autres modalités et la parallélisation des processus d'interprétations).
2. La seconde solution consiste à ne pas travailler directement sur des ensembles d'objets, mais sur les contraintes visant à l'identification des référents.

La différence entre ces deux types de stratégies tient à la différence entre « forme logique » et « contenu propositionnel » que nous explicitons ci-dessous.

Forme logique et contenu propositionnel

Le corpus que nous considérons ne contient que des ordres. Sans entrer dans le détail de la théorie des actes de langage [12], nous admettrons que chaque énoncé peut être représenté sous la forme : ORDRE (P) avec P une proposition (le contenu propositionnel).

La réaction attendue de la part du système de dialogue face à un énoncé est de rendre vraie cette proposition P. Ainsi, un ordre tel que *« mets cet objet dans la troisième boîte »* se formalise comme :

```
∃e e-mettre(Système,o42,b43) and e>now
```

Le système doit rendre vraie cette proposition et doit pour cela exécuter une action (future par rapport au moment de l'énoncé) de mettre l'objet précis (représenté par une constante `o42`) mentionné dans l'énoncé, dans la boite (`b43`) à laquelle réfère l'expression *« la troisième boîte »*.

Le contenu propositionnel est donc lié au contexte : on y trouve de véritables objets de l'application. *A contrario*, l'analyse linguistique de l'énoncé (hors contexte) nous livre sur cet exemple une *forme logique* telle que :

```
∃e ∃o ∃b e-mettre(Interlocuteur,o,b) and
objet(o) and démonstratif(o) and boite(b)
and défini(b) and numéro(b,3)
```

La différence entre les deux stratégies évoquées au début du paragraphe 3 peut donc se reformuler de la façon suivante :

1. Première stratégie : fusion sur le contenu propositionnel (ou plutôt sur des hypothèses concurrentes quant à ce contenu).
2. Deuxième stratégie : fusion sur la forme logique.

De la forme logique au contenu propositionnel

Nous avons illustré la différence entre forme logique et contenu propositionnel en ne considérant que des ERs démonstratives et définies (*« cet objet »*, *« la troisième boîte »*). Parmi les expressions à résoudre lors du passage de la forme logique au contenu propositionnel, on trouve également des expressions anaphoriques, et en particulier des pronoms de troisième personne. Or, sur l'ensemble du corpus considéré, **on n'observe aucun cas d'expression à la fois anaphorique et cogestuelle sur un individu** (le cas fréquent de

« celui / celle-ci » correspond à une reprise anaphorique de la classe et non pas d'un individu particulier : *« cet objet et celui-ci » = « cet objet et cet [objet]-ci »*). La première stratégie autoriserait de tels cumuls et nous semble donc devoir être rejetée.

La question qui se pose alors est de savoir comment cumuler des contraintes sémantiques en provenance respectivement de la langue et du geste.

Contraintes sémantiques issues du geste

Un geste est caractérisé par une trajectoire, c'est-à-dire une succession de coordonnées, l'exemple le plus simple étant le geste de pointage caractérisé par un seul couple de coordonnées.

La prise en compte du geste dans l'énoncé multimodal nécessite en premier lieu une étude fine de la trajectoire gestuelle [1]. Cette étude consiste en une analyse hors contexte perceptif et linguistique de la trajectoire. Il s'agit donc à partir d'un signal d'extraire les parties significatives du geste. Chacune de ces parties peut ensuite être catégorisée et donc se voir affecter une étiquette sémantique indiquant par exemple s'il s'agit d'un entourage ou d'un ciblage. L'ambiguité sémantique de la trajectoire est prise en compte par la proposition d'une liste d'hypothèses correspondant aux différents découpages possibles de la trajectoire. La superposition de ces hypothèses avec le contexte perceptif permet de rattacher les objets candidats à la co-référence.

Le module de calcul du geste travaille indépendamment de l'éventuelle expression linguistique co-occurente au geste. Dans cette mesure, même dans le cas du pointage sur un objet simple, au moins une hypothèse de désignation d'ensemble d'objets est calculée (qu'on songe à des expressions linguistiques telles que *« ces objets »* ou *« ce groupe de trois triangles »*). Les ensembles d'objets candidats au niveau perceptif sont déterminés grâce à l'implantation de l'algorithme de Thórisson [13] modélisant la notion de groupement perceptif à partir des principes de la Gestalttheorie.

On dispose donc des informations suivantes pour chaque hypothèse de partie significative : ses caractéristiques physiques (listes des coordonnées, centre de gravité, rectangle englobant, etc.), une étiquette sémantique (`TrajectoireOuverte`, `Entourage`, `Pointage`, etc), et les objets candidats. Les gestes peuvent être éventuellement associés à des événements plus sémantiques pour exprimer par exemple l'orientation ou la vitesse.

4. Interprétation de l'énoncé multimodal

Dans ce qui précède, nous avons fait la différence entre forme logique (= sens de l'énoncé hors contexte) et contenu propositionnel. Dans le cadre qui nous intéresse, le contenu propositionnel est une proposition que le système doit rendre vraie. De ce fait, il reste dans ce contenu propositionnel des variables quantifiées (en l'occurrence, sur nos exemples simples, quantifiées existentiellement). Une des difficultés du passage de la forme logique vers le contenu propositionnel est d'identifier les expressions référentielles (pour lesquelles on a une constante dans la forme logique) et les expressions non référentielles qui restent donc quantifiées dans le contenu propositionnel. Dans le cas d'un ordre en particulier, l'action elle-même est non référentielle.

A titre d'illustration, considérons l'énoncé : *« déplace cet objet dans la troisième boîte » + geste de pointage sur un objet + trajectoire C allant de l'objet à la boîte*. Cet énoncé se traduit en un contenu propositionnel de la forme :

```
∃e e-déplacer(Système,o42,b43) and
trajectoire(e,C)
```

Les gestes associées doivent être interprétés respectivement comme :

1. Un geste de désignation participant à l'identification de `o42`.
2. Un geste sémantique précisant des caractéristiques de l'événement de déplacement.

La difficulté est donc d'associer ou non des informations gestuelles aux variables de la forme logique qui dans notre cas était :

```
∃e ∃o ∃b e-déplacer(Interlocuteur,o,b) and
objet(o) and démonstratif(o) and boîte(b)
and défini(b) and numéro(b,3)
```

Comme notre exemple l'illustre, seuls `e` et `o` se sont vus ici associés des gestes.

Du côté des gestes, on avait ici `g1` et `g2` tels que :

```
Pointage(g1)[objets possibles:o42]

TrajectoireOuverte(g2)[objets possibles:
o42,...,b43]
```

Pour la combinaison de la langue et du geste, cet exemple illustre trois points :

- Le modèle sémantique nous permet d'associer une trajectoire à un événement de déplacement. (ce ne serait pas le cas avec un événement de destruction par exemple).
- L'écriture logique nous fait perdre toute synchronisation temporelle entre expressions et gestes ; si comme nous l'avons dit, la datation temporelle précise est non pertinente, l'ordre relatif des ERs et l'ordre relatif des gestes sont eux essentiels.
- La combinaison ER–geste permet d'ignorer certaines hypothèses issues du traitement gestuel (l'ensemble des référents possibles dans notre cas).

La solution que nous proposons est exactement fondée sur les constatations précédentes :

- Nous avons une représentation de la forme logique de l'énoncé (issue de l'analyseur syntaxique TAG [7]). A cette représentation est associée une liste qui correspond à l'ordre des variables de la forme logique dans la forme linéaire de l'énoncé.
- Le module d'analyse des gestes fournit une liste de catégories de gestes et d'hypothèses concernant les objets désignés.
- Le module de fusion de références multimodales, qui est en fait le module de calcul de référence, assure le passage de la forme logique au contenu propositionnel expression après expression. Selon

la nature de l'expression, il peut décider : (a) de ne pas associer de geste à l'expression ; (b) d'associer un geste vu comme une désignation ; ou (c) d'associer un geste vu de façon purement sémantique.

Algorithme de fusion

Notre algorithme de fusion est donc un algorithme de passage de la forme logique au contenu propositionnel. Comme nous l'avons précédemment expliqué, ce passage n'est pas limité aux co-références entre ERs et geste.

L'algorithme est le suivant :

```
Pour chaque variable apparaissant dans la
forme logique faire les 4 hypothèses
suivantes:
   1. autonomie référentielle
   2. anaphore
   3. co-référence avec un geste (i.e. en
      n'utilisant que les ensembles
      d'objets fournis par le module de
      traitement du geste et en oubliant la
      partie sémantique gestuelle)
   4. co-sémantique du geste (i.e. ajout de
      la catégorie sémantique du geste en
      tant que prédicat sur la variable
      considérée)
Fin pour
```

On impose les contraintes de compatibilité supplémentaires suivantes :

- Les indéfinis ne sont pas anaphoriques ni co-référentiels avec des gestes.
- Les pronoms de troisième personne sont nécessairement anaphoriques.

Cet algorithme semble au premier abord engendrer énormément d'ambiguïtés. Nous allons montrer sur un exemple qu'on obtient en fait le plus souvent le même contenu propositionnel en de multiples exemplaires. Reprenons ainsi notre exemple : *« déplace cet objet dans la troisième boîte » + geste de pointage sur un objet + trajectoire C allant de l'objet à la boîte*.

Supposons que le module du traitement du geste nous fournisse plusieurs hypothèses :

(a) Une désignation de l'objet (`Pointage[o42]`) suivie de : `trajectoireOuverte[o42,b43]`.

(b) Deux désignations (`Pointage[o42]` et `Pointage[b43]`).

(c) Uniquement la trajectoire (`trajectoireOuverte[o42,b43]`).

Chacune de ces hypothèses quant au geste nous mène au même contenu propositionnel : la trajectoire du déplacement nous impose l'objet concerné et le lieu d'arrivée !

5. Conclusion

Le problème de la fusion des expressions référentielles et des gestes est extrêmement combinatoire, surtout dans un contexte ou les gestes peuvent être ambigus. Dans cet article, nous montrons que la résolution de chaque expression prise individuellement mène à de nombreuses hypothèses, mais que le contenu propositionnel dans son ensemble est souvent le même. La méthode que nous proposons vise donc à obtenir ce contenu propositionnel, les actions du système étant calculées à partir de celui-ci dans un second temps seulement.

Bibliographie

[1] Bellalem, N., *Etude du mode de désignation dans un dialogue homme-machine finalisé à forte composante langagière : analyse structurelle et interprétation*, Thèse de doctorat, Université Henri Poincaré de Nancy, 1995.

[2] Bellik, Y. & Teil, D., L'intégration à base de règles, *Compte rendu des ateliers d'IHM'92, Atelier Interfaces multimodales et architecture logicielle*, Paris, 1992.

[3] Brison, E., *Stratégies de compréhension dans l'interaction multimodale*, Thèse de doctorat, Université Paul Sabatier de Toulouse, 1997.

[4] DeAngeli, A., Gerbino, W., Petrelli, D. & Cassano, G., Visual Display, Pointing and Natural Language: The Power of Multimodal Interaction, *Proceedings of the Visual Conference on Advanced Visual Interfaces – AVI'98*, L'Aquila, 1998.

[5] Gaiffe, B. & Romary, L., Constraints on the Use of Language, Gesture and Speech for Multimodal Dialogues, *ACL / EACL Workshop on Referring Phenomena in a Multimedia Context and their Computational Treatment*, Madrid, 1997.

[6] Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A. & Smith, I., Unification-Based Multimodal Integration, *35th Annual Meeting of the Association for Computational Linguistics*, ACL Press, 1997.

[7] Lopez, P., A LTAG Grammar for Parsing Oral and Incomplete Utterances, *European Conference on Artificial Intelligence – ECAI 98*, Brighton, 1998.

[8] McNeill, D., *Hand and Mind – What Gestures Reveal About Thought*, Chicago Press, 1992.

[9] Nigay, L. & Coutaz, J., Espace problème, fusion et parallélisme dans les interfaces multimodales, *Actes de la conférence InforMatique'93, l'interface des mondes réels et virtuels*, Montpellier, 1993.

[10] Oviatt, S.L., DeAngeli, A. & Kuhn, K., Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction, *Conference on Human Factors in Computing Systems, CHI 97*, New York : ACM Press, 1997.

[11] Reboul, A., Déixis et anaphore, dans : Moeschler, J. & Reboul, A., *Dictionnaire Encyclopédique de Pragmatique*, pages 349–372, Seuil, 1994.

[12] Searle, J.R. & Vanderveken, D., *Foundations of Illocutionary Logic*, Cambridge University Press, 1985.

[13] Thórisson, K.R., Simulated Perceptual Grouping: An Application to Human-Computer Interaction, *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Atlanta, 1994.

[14] Wolff, F., *Analyse contextuelle des gestes de désignation en dialogue homme-machine*, Thèse de doctorat, Université Henri Poincaré de Nancy, 1999.

[15] Wolff, F., DeAngeli, A. & Romary, R., Acting on a Visual World: The Role of Perception in Multimodal HCI, *AAAI Workshop on Multimodal Representation*, Madison, 1998.