Reformuler des expressions multimodales

Elisabeth Godbert

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
Université de la Méditerranée et CNRS
163 Avenue de Luminy - case 901
13288 Marseille Cedex 9 - France
E-mail : godbert@lim.univ-mrs.fr

Résumé – Abstract

Le domaine des "Interfaces Utilisateur Intelligentes" a vu ces dernières années la réalisation de systèmes complexes mettant en oeuvre une interaction multimodale dans laquelle les différentes techniques de communication (textes, gestes, parole, sélection graphique) sont coordonnées en entrée et/ou en sortie. Nous nous intéressons ici aux systèmes qui prennent en entrée des expressions multimodales et en produisent une reformulation en une expression unimodale sémantiquement équivalente. Nous proposons une modélisation du processus de traduction d'expressions multimodales en expressions unimodales, et nous décrivons la mise en oeuvre d'un processus de ce type dans un logiciel d'aide à l'apprentissage du langage.

These last years, "Intelligent User Interfaces" have been developped, in which input and/or output multimodality facilitates human-machine communication (written language, speech, graphic selection, gestures). This paper is concerned with systems in which multimodal input is translated into unimodal expressions semantically equivalent. A model for such a process is proposed. Then, an example is described, with an educational software enabling multimodal communication and implementing this process.

Mots Clés –Keywords

Langage multimodal, coordination des modes, expressions sémantiquement équivalentes.
Multimodal language, coordination of modalities, semantic equivalence.

1 Introduction

Dans de très nombreuses situations l'activité humaine peut être qualifiée de multimodale, en fait dès qu'une personne fait plusieurs choses en même temps : marcher, réfléchir, chanter, parler, manger ... Chercher à définir le traitement automatique d'une activité multimodale, c'est tenter d'élaborer un système capable de réagir de façon appropriée à une telle activité. Dans toute sa généralité, il est infiniment complexe, il requiert en particulier :

- l'analyse en parallèle des différents éléments asynchrones dont est composée cette activité,
- leur interprétation et traduction en une information structurée, qui pourra ensuite être traitée par des processus adaptés aux buts poursuivis, pour réagir de façon adéquate à l'activité prise en entrée.

Pour le cas particulier de l'expression (composition d'un message destiné à un interlocuteur) on utilise très souvent plusieurs modalités, combinées ou non : parole, écrit, gestes, etc.
Depuis une quinzaine d'années, les recherches sur la multimodalité s'intéressent essentiellement aux interfaces qualifiées de multimodales et multimedia, un média étant défini comme un dispositif servant de support à l'information, et une modalité comme une technique d'interaction. L'un des premiers systèmes de ce type est décrit dans (Bolt, 80). Ces systèmes utilisent la multimodalité en entrée et/ou en sortie, et le domaine des "interfaces utilisateur intelligentes" a vu la réalisation de systèmes complexes utilisant le multimédia, et mettant en oeuvre une interaction multimodale dans laquelle les différentes techniques de communication (textes, gestes, parole, vision par ordinateur, images, vidéo,...) sont combinées : on parle alors de système multimédia intelligent (Nigay, Coutaz, 96) (Wahlster et al, 93) (Cohen et al, 98).
Le problème auquel nous nous intéressons ici est la modélisation d'un système qui prend en entrée des expressions multimodales, et en produit une reformulation en une expression unimodale sémantiquement équivalente. Dans un certain nombre de systèmes en effet, il s'avère nécessaire de traduire l'ensemble des entrées en une expression unimodale, celle-ci étant destinée à être ensuite donnée en entrée à un autre module, chargé de poursuivre le traitement. Nous mentionnons dans la partie 2 deux classes d'applications auxquelles nous nous intéressons, et qui requièrent ce type de traitement : les logiciels pour l'apprentissage du langage et les systèmes d'aide à la communication pour les personnes ayant perdu l'usage de la parole. Nous proposons dans la partie 3 une modélisation du processus de traduction d'expressions multimodales en expressions unimodales. Pour finir, nous décrivons dans la partie 4 la mise en oeuvre d'un processus de ce type dans un logiciel d'aide à l'apprentissage du langage.

2 La multimodalité : une aide à l'utilisateur

L'importance pratique de l'aide à l'utilisateur a été soulignée dans de nombreux travaux sur la conception d'interfaces intelligentes, et il est clair que la multimodalité et l'aide à l'utilisateur sont intimement liées : la multimodalité de l'interaction facilite l'expression de l'utilisateur. Les recherches dans ce domaine ont mis en évidence les avantages respectifs des différents modes de communication : langage naturel écrit, parole, pointage graphique, gestes, etc.

2.1 Interfaces multimodales classiques

De façon classique, une interface multimodale est définie pour utiliser au mieux tous les outils habituels des interfaces graphiques (icônes, menus déroulants, boutons, rectangles d'édition, etc.), auxquels s'ajoutent éventuellement un micro et différents capteurs pour permettre une interaction multimodale dans laquelle les différents modes et média sont proposés. L'idéal est que l'interface permette d'utiliser les modalités d'expression non pas séquentiellement mais en parallèle, de façon coordonnée et complémentaire. Par exemple, l'utilisateur doit pouvoir à tout instant choisir et combiner, parmi les modalités <clavier, langage naturel>, <souris, éléments graphiques> et <micro, parole>, celles qu'il préfère pour s'exprimer.
Les réactions de l'utilisateur étant imprévisibles, plus le choix est large pour les modalités d'interaction, plus la conception de ce type de système est complexe. Elle requiert en

particulier la présence d'un analyseur multimodal capable de sélectionner des entrées de différents types (texte, graphique, etc.) et de les fusionner en une expression regroupant le contenu de toutes les entrées.
Si, de façon analogue, on veut mettre en place la multimodalité en sortie, celle-ci passe par la fission de la réaction du système, et par des méthodes de présentation dans laquelle les modes pertinents sont utilisés de façon parallèle et complémentaire. Et la cohérence d'un dialogue multimodal demandera que les modes en sortie soient choisis en fonction des modes utilisés en entrée.

2.2 Interfaces multimodales pour handicapés

Cette vision classique de la multimodalité doit être élargie pour y inclure les systèmes adaptés à des besoins spécifiques des utilisateurs, et dans lesquels on utilise donc d'autres modalités d'interaction. C'est le cas des interfaces dédiées à des personnes souffrant de handicaps. On parle actuellement beaucoup de "droit universel pour l'accès à la société de l'information". De nombreux systèmes ont été développés pour pallier à la perte de facultés de l'utilisateur soit pour la réception d'informations soit pour la composition de messages : systèmes de reconnaissance de gestes, interfaces tactiles, divers types de capteurs, aide à la composition, synthèse vocale...(voir par exemple Jacko et al, 2001). Dans le cadre de la problématique étudiée dans cet article, nous nous intéressons aux systèmes d'aide à la communication pour des personnes ayant perdu l'usage de la parole : si un tel système propose une interface multimodale qui permet la saisie d'entrées provenant de différents capteurs, il est nécessaire d'opérer la fusion de ces entrées en une expression unimodale, qui pourra ensuite être proposée à son destinataire sous forme de langage écrit ou sous forme orale dans le cas où le message écrit est donné en entrée à un synthétiseur de la parole.

2.3 Une nécessité : pouvoir adapter la multimodalité à l'utilisateur

Les systèmes conçus à l'heure actuelle donnent une importance grandissante au "profil utilisateur", dont le rôle premier est de permettre l'adaptation d'un système à son utilisateur. En ce qui concerne la multimodalité des entrées, cette adaptation correspond à la possibilité de choisir la ou les modalités d'expression que l'on préfère.
Cette adaptation à l'utilisateur est particulièrement intéressante dans les systèmes dédiés à des personnes handicapées : selon leur handicap, ces personnes peuvent choisir l'une ou l'autre des modalités offertes. De plus, pour des utilisateurs atteints d'un handicap qui évolue au fil des années, comme dans le cas de maladies dégénératives, cette adaptation a une importance cruciale puisqu'elle permet à l'utilisateur de garder le même système de base, sur lequel on greffe, suivant les besoins, différentes modalités d'interaction.
En ce qui concerne les logiciels éducatifs, les pédagogues insistent souvent sur la nécessité de pouvoir faire varier la difficulté des exercices, et en particulier pouvoir graduer l'aide. Par exemple, il faut pouvoir choisir entre un mode d'expression multimodal ou unimodal, assisté ou non, ou entre divers degrés d'aide au niveau cognitif.
On pourra ainsi définir un *degré de multimodalité*, qui correspondra à la possibilité d'utiliser ou non une ou plusieurs modalités en entrée.

3 Passer d'expressions multimodales à des expressions unimodales

Comme nous l'avons dit, nous voulons définir un système qui prend en entrée des expressions multimodales et les traduit en expressions unimodales sémantiquement équivalentes. Nous proposons dans ce qui suit une modélisation du flux des entrées multimodales, et de leur traduction en expressions unimodales.

Appelons L_{MM} le langage multimodal que l'utilisateur utilise, et L_E le langage unimodal dans lequel doivent être traduites les entrées. Sans rien préjuger de sa nature, L_E est dit "langage de base". On peut penser que, dans de nombreuses applications, L_E est le (ou un sous-ensemble du) langage naturel écrit. Nous en verrons un exemple dans la partie 4.

Les expressions exprimées dans L_{MM} sont saisies de façon asynchrone par différents capteurs supposés indépendants, notés $\{Capt_1, Capt_2,,...\}$, qui peuvent être le clavier, la souris, un micro, ou tout autre type de capteur. Notons que faire varier le degré de multimodalité, pour adapter la multimodalité à l'utilisateur, correspond simplement à l'activation ou la non-activation de chaque capteur.

Appelons $\{m_1, m_2,...\}$ les messages, indicés par leur ordre d'arrivée, que reçoivent les capteurs : à la date t_i, le signal S_i marque la fin de la composition du message m_i et donc son entrée pour le traitement qui doit suivre (on vide alors le buffer du capteur considéré, qui devient prêt à recevoir un éventuel message suivant).

Notons m'_i la traduction dans L_E du message m_i. Nous supposerons que cette traduction est instantanée. Et notons $\{c_0, c_1,...\}$ les expressions unimodales produites par le système au fur et à mesure que les entrées sont saisies, traduites, et fusionnées avec le résultat du traitement des messages précédents. Par définition, l'expression unimodale c_i est équivalente à l'ensemble des messages $\{m_1, ... m_i\}$. Dans ce qui suit, l'opération de fusion est notée "+".

La figure1 illustre le cas le plus simple : la nouvelle expression courante c_i est obtenue par la fusion de l'expression courante c_{i-1} et de la traduction m'_i du message m_i.

Dans ce cas, l'équation de fusion s'écrit : $c_i = c_{i-1} + m'_i$.

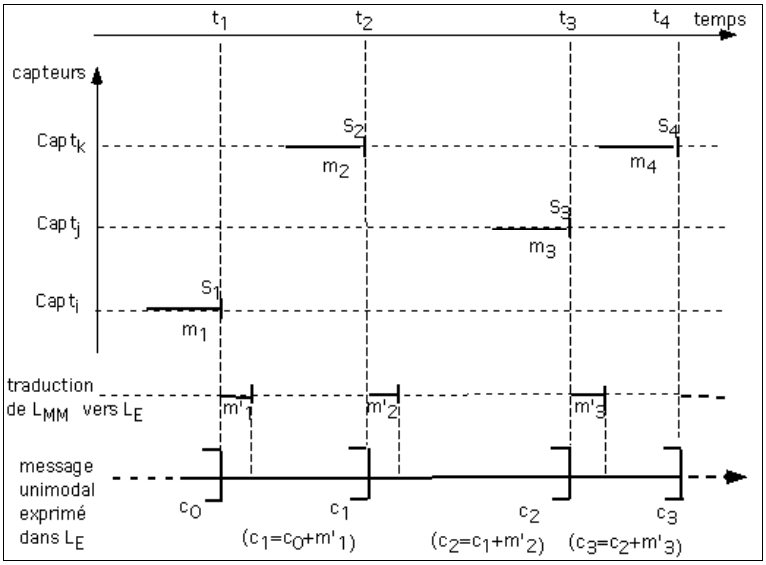

Figure 1 : Passer d'une expression multimodale à une expression unimodale

L'état du système évolue donc de la façon suivante :

- à l'instant t_i, date d'arrivée d'un signal S_i annonçant la saisie d'un message m_i sur un capteur, l'expression unimodale courante est c_{i-1} ;
- à partir de t_i, (et instantanément), le message m_i est traduit en m'_i et fusionné avec l'expression courante, pour donner c_i.

Exemple : dans un système qui combine la saisie de texte au clavier et la sélection graphique de mots, la fusion peut tout simplement être la concaténation de chaînes de caractères.
D'autres situations, moins simples, peuvent se présenter, en particulier lorsque la fusion d'un message m_i avec l'expression c_{i-1} ne peut pas se faire directement pour obtenir une expression unimodale équivalente ; il faut alors attendre la saisie d'un ou plusieurs messages succédant à m_i pour effectuer en même temps la fusion de ces messages avec c_i (un exemple est décrit dans la partie 4). Dans ce cas, l'équation de fusion s'écrit : $c_{i+k} = c_{i-1} + m'_i + m'_{i+1} + ... + m'_{i+k}$

4 La multimodalité dans le système éducatif EREL

Le projet EREL développé ces dernières années, a pour objectif le développement d'un système pour l'éducation et/ou la rééducation du langage et de la cognition chez des enfants présentant des troubles du développement (Godbert, 1998). EREL est un système multimodal qui propose un ensemble d'exercices ludiques de type piagétien conçus pour stimuler et aider l'utilisateur à employer le langage naturel écrit pour s'exprimer autour de différents thèmes illustrés à l'écran. C'est un système réactif dans lequel chaque exercice met en jeu un micro-monde d'objets graphiques, et est organisé pour que l'utilisateur s'exprime par des phrases simples, appelées requêtes, que le système l'aide à composer et auxquelles il réagit (déplacement, création ou suppression d'objets graphiques). Un langage multimodal approprié est associé à chaque jeu pour la composition de ces requêtes.
Suivant le degré de multimodalité choisi, la désignation d'un objet à l'écran peut se faire soit par une simple sélection graphique, soit par du texte, soit par une combinaison des deux modes. Par exemple, l'une des activités proposées par EREL met en jeu des pions de différentes formes et couleurs placés sur un damier. Selon la situation, l'utilisateur peut désigner une case par différentes expressions multimodales sémantiquement équivalentes : (*Pose la croix*) *en E4*, ou *en <clic>*, ou *dans la case <clic>*, ou *à droite du triangle bleu*, etc.
Au fur et à mesure que l'utilisateur compose sa requête, celle-ci apparaît, sous forme écrite, à l'écran. L'opération de fusion d'expressions textuelles et graphiques correspond dans EREL à la génération d'expressions définies pertinentes du langage naturel qui dénotent des éléments qui ont été désignés (en partie ou entièrement) graphiquement. Un certain nombre de choix ont été faits parmi ces expressions discriminantes qu'il est possible de générer dans chaque cas. Ainsi, si l'utilisateur clique dans la case E5 du damier, *en <clic>* sera traduit par *en E5*, et *dans la case <clic>* sera traduit par *dans la case E5* ; *cette <clic> case* sera traduit par *la case E5*, *cette <clic> colonne* sera traduit par *la colonne 5*. Dans le cas où une étoile verte se trouve en E5, nous avons choisi de traduire *cette <clic> étoile*, soit par *l'étoile verte* s'il n'y a qu'une étoile verte dans le jeu, soit par *l'étoile qui se trouve en E5* dans le cas contraire.
On voit ici que pour certaines expressions déictiques, le clic n'est pas interprété directement en langage écrit dès qu'il est produit : il faut attendre le mot suivant (*case, colonne, étoile*, ...) pour générer une expression définie adéquate.

Pour chaque activité proposée par EREL le "langage de base" L_E est un sous-langage du langage naturel écrit. Le langage multimodal L_{MM} est une extension de L_E, c'est l'ensemble des requêtes multimodales que le système sait traiter. L'extension L_{MM} de L_E est définie de la

façon suivante : on ajoute à la grammaire de L_E des règles selon lesquelles certaines catégories syntaxiques peuvent se dériver en des expressions graphiques : des "morceaux de phrases" peuvent donc être exprimés soit par du texte, soit par un pointage graphique, soit par une combinaison des deux. Dans la version actuelle d'EREL, les langages L_{MM} sont suffisamment contraints pour que l'on évite toute ambiguïté ou conflit entre les deux modalités.
Pour le traitement du langage écrit, nous utilisons les outils offerts par le système ILLICO (Pasero et Sabatier, 1999) : en premier lieu des formalismes pour définir des langages par des données lexicales, syntaxiques, sémantiques, conceptuelles et contextuelles ; ensuite, des algorithmes d'analyse/synthèse de phrases ; enfin, des algorithmes pour la recherche des référents et pour le calcul de la représentation logique de la sémantique des phrases.

5 Conclusion

Nous avons proposé un modèle pour la saisie d'expressions multimodales et pour leur traduction en expressions unimodales, pour que celles-ci puissent ensuite être prises en entrée par un autre module de traitement du langage. Nous pensons en effet que cette traduction s'avère nécessaire dans de nombreux systèmes. Nous en avons évoqué deux classes : les systèmes d'aide à l'apprentissage du langage et les systèmes d'aide à la communication pour les personnes handicapées. Pour finir, nous avons décrit comment ce processus de traduction a été mis en oeuvre dans le système éducatif EREL. La multimodalité proposée dans ce système, bien que très modeste, met en évidence des problèmes intéressants relatifs à la génération d'expressions discriminantes dénotant des objets désignés par des expressions multimodales combinant texte et sélection graphique.

Références

Bolt R.A., (1980). "Put-That-There" : Voice and Gesture at the Graphics Interface. *Computer Graphics*, 14(3). Also in *Intelligent User Interfaces*, Maybury M.T. and Wahlster W. (eds), Morgan Kaufmann Publishers, San Francisco, 1998, pp 19-27.
Cohen P.R., Johnson M., McGee D., Oviatt S., Pittman J., Smith I., Chen L., and Clow J., (1998). "Multimodal Interaction for Distributed Interactive Simulation", in *Intelligent User Interfaces*, Maybury M.T. and Wahlster W. (eds), Morgan Kaufmann Publishers, San Francisco, 1998, pp 562-571.
Godbert E., (1998). "EREL : a multimedia CALL system devoted to children with language disorders". In Keith Cameron Ed., *Multimedia CALL: Theory and Practice*, Elm Bank Publications, Exeter, England, 1998, pp 207-216.
Jacko J.A., Vitense H.S., (2001). "A review and reappraisal of information technologies within a conceptual framework for individuals with disabilities". *UAIS Journal*, Springer Verlag 2001, 1, pp 56-76.
Nigay L., Coutaz J., (1996). "Espaces conceptuels pour l'interaction multimédia et multimodale". Techniques et Science Informatiques, vol. 15, n° 9, 1996, pp 1195-1225.
Pasero R., Sabatier P., (1999). "Specifying and Processing Constraints on Formal Representations of Sentences", *Proceedings of the 6th International Conference on Natural Language Understanding and Logic Programming, NLULP 99*, Las Cruces, pp 33-44.
Wahlster W., André E., Finkler W., Profitlich H.J., Rist T., (1993). "Plan-based integration of Natural Language and Graphics Generation". *Artificial Intelligence*, 1993, n°63, pp 387-427.