TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Une caractérisation de la pertinence pour les actions de référence

Frédéric Landragin
LORIA — UMR 7503
Campus scientifique — BP 239
54506 Vandœuvre-lès-Nancy CEDEX
Frederic.Landragin@loria.fr

Mots-clefs – Keywords

Pertinence, référence aux objets, dialogue multimodal, effets contextuels, effort de traitement
Relevance, reference to objects, multimodal dialogue, contextual effects, processing effort

Résumé - Abstract

Que ce soit pour la compréhension ou pour la génération d'expressions référentielles, la Théorie de la Pertinence propose un critère cognitif permettant de comparer les pertinences de plusieurs expressions dans un contexte linguistique. Nous voulons ici aller plus loin dans cette voie en proposant une caractérisation précise de ce critère, ainsi que des pistes pour sa quantification. Nous étendons l'analyse à la communication multimodale, et nous montrons comment la perception visuelle, le langage et le geste ostensif interagissent dans la production d'effets contextuels. Nous nous attachons à décrire l'effort de traitement d'une expression multimodale à l'aide de traits. Nous montrons alors comment des comparaisons entre ces traits permettent d'exploiter efficacement le critère de pertinence en communication homme-machine. Nous soulevons quelques points faibles de notre proposition et nous en tirons des perspectives pour une formalisation de la pertinence.

For automatic comprehension or generation of referring expressions, Relevance Theory proposes a cognitive criterion that allows to compare the relevance of several expressions in a linguistic context. We want here to pursue this work. We propose a more precise characterization of this criterion and foundations for its computation. We extend the analysis to multimodal communication, and we show how visual perception, speech and gesture present multiple interactions between each other in the production of contextual effects. We describe with some features the processing effort of a multimodal expression. Then we show how comparisons between these features lead to an efficient exploitation of the relevance criterion in man-machine dialogue. We raise some weak points of our proposal and we deduce future plans for a formalization of relevance.

1 Introduction

Formaliser la Théorie de la Pertinence (Sperber, Wilson, 1995) à travers la caractérisation et la quantification de la pertinence d'un énoncé est un but ambitieux. Si disposer d'un critère numérique de pertinence s'avérerait utile pour une grande partie des applications du traitement automatique des langues naturelles, l'analyse d'un tel critère pose de nombreux problèmes, liés à la nature et à la multiplicité des paramètres entrant en compte. Ainsi, s'il existe de très nombreux travaux sur la Théorie de la Pertinence, en particulier sur son adaptation à tel ou tel aspect du langage (composants grammaticaux, actes de langages, métaphore, ironie) ou à telle ou telle application (acquisition du langage, dialogue coopératif, analyse littéraire, traduction), il n'existe quasiment aucune proposition de formalisation (voir en particulier le site *Relevance Theory Online Bibliographic Service*, `http://www.ua.es/dfing/rt.htm`). Sperber et Wilson eux-mêmes ne donnent pas de pistes pour un tel travail. Certains auteurs proposent des points de départ, comme (Gazdar, Good, 1982). D'autres tentent des rapprochements entre la Théorie de la Pertinence et une théorie plus formelle. (Akman, Surav, 1995) se focalisent par exemple sur la formalisation du contexte en s'aidant de leur théorie (*Computational Situation Theory*). Ils projettent les notions d'effets contextuels et d'effort de traitement à la base du critère de pertinence sur leur vision particulière du contexte. Récemment, (van Rooy, 2003) montre comment la théorie *Bidirectional Optimality Theory* de (Blutner, 2000) peut être vue comme une formalisation de la Théorie de la Pertinence. Il fait un rapprochement entre deux principes de la première et les notions d'effets et d'effort de la seconde, et aboutit à une quantification des effets, en particulier pour la pertinence d'une réponse partielle à une question. Pour notre part, nous nous focalisons sur les actions de référence, et plus généralement sur la nature du contexte dans la communication multimodale associant le langage et le geste ostensif. A partir de travaux antérieurs (Landragin et al., 2002) montrant l'importance de sous-ensembles contextuels (appelés *domaines de référence*), nous proposons ici une caractérisation des effets et de l'effort dans un cadre computationnel, aussi bien pour la compréhension que pour la génération automatique d'expressions référentielles langagières ou multimodales. Un intérêt de notre caractérisation est de regrouper des considérations syntaxiques, sémantiques et pragmatiques sur le langage et sur le geste, ainsi qu'une prise en compte des particularités du contexte visuel. Nous montrons comment ces paramètres de nature hétérogène interviennent conjointement dans les effets et l'effort, et comment on peut aboutir à une calculabilité de ceux-ci.

2 Pertinence et référence dans le dialogue multimodal

2.1 Le contexte, les effets contextuels et l'effort de traitement

La Théorie de la Pertinence (Sperber, Wilson, 1995) s'intéresse particulièrement à la construction du contexte lors de la communication. Sperber et Wilson définissent le contexte comme un ensemble de propositions (ou hypothèses), ces propositions pouvant être vraies, probablement vraies, plutôt vraies, plutôt fausses, probablement fausses, ou fausses. Ces degrés, appelés forces des propositions, varient au cours de la communication, au fur et à mesure que les énoncés apportent de nouvelles informations. Ainsi, la confrontation de la proposition portée par le nouvel énoncé avec les propositions contenues dans le contexte peut donner lieu à une contextualisation, c'est-à-dire une déduction utilisant ces deux sources d'information comme prémisses. Dans ce cas, on dira que l'interprétation de l'énoncé a donné lieu à des effets contextuels, qui

peuvent être l'effacement de certaines hypothèses du contexte, la modification de la force de certaines hypothèses, ou la dérivation d'implications contextuelles (conclusions nouvelles qui ne seraient pas dérivables de la proposition de l'énoncé seule ou des propositions du contexte seules). Dans le cas contraire, c'est-à-dire si la contextualisation ne fait qu'ajouter au contexte l'information nouvelle, on dira que l'interprétation de l'énoncé n'a donné lieu à aucun effet contextuel. Cette notion d'effet contextuel met en avant l'aspect inférentiel de la communication, qui n'est plus vue comme un simple processus de codage et de décodage.

Les effets contextuels sont le produit de processus mentaux qui demandent un certain effort. L'interprétation nécessite ainsi un effort de traitement qui correspond à la dépense d'énergie des processus mentaux activés. Cet effort de traitement dépend de la longueur de l'énoncé, de la facilité d'accès aux informations encyclopédiques, ou encore du nombre de règles logiques impliquées dans le mécanisme déductif. Avec ces deux notions d'effets et d'effort, (Sperber, Wilson, 1995) définissent la pertinence de manière comparative. Ainsi, un acte de communication est d'autant plus pertinent dans un contexte donné que ses effets contextuels y sont plus importants. D'autre part, un acte de communication est d'autant plus pertinent dans un contexte donné que l'effort nécessaire pour l'y traiter est moindre. La pertinence peut donc être vue comme le rapport des effets contextuels sur l'effort de traitement. Avec nos préoccupations computationnelles, c'est de cette manière que nous l'aborderons et que nous tenterons de l'évaluer. Chaque effet contextuel nécessitant un supplément d'effort, Sperber et Wilson définissent également l'effort impliqué par un effet. Ce supplément d'effort n'est pas coûteux au point d'annuler la contribution de l'effet à la pertinence. Puisqu'il est toujours proportionné aux effets qui le rendent nécessaire, nous pouvons l'ignorer dans l'évaluation de la pertinence.

2.2 Vers une évaluation de la pertinence

La Théorie de la Pertinence ne s'attache pas vraiment à évaluer les effets et l'effort, mais plutôt à décrire comment l'esprit évalue lui-même ses propres résultats et ses propres efforts, et comment il décide en conséquence de poursuivre ses efforts dans la même direction, ou au contraire de les diriger ailleurs (Sperber, Wilson, 1995). Il s'avèrerait pourtant intéressant de pouvoir quantifier la pertinence d'un énoncé. On pourrait comparer les pertinences de plusieurs propositions, ou encore déterminer la forme la plus pertinente d'une intention communicative donnée.

Le problème est large. On ne sait pas quelles sont les opérations élémentaires constitutives des processus intellectuels complexes, et on ne sait donc pas comment évaluer l'effort de traitement d'un énoncé. On sait que certains paramètres comme la durée d'un processus mental ne sont pas de bons indices : du fait de l'impossibilité de détecter une réflexion intense d'une réflexion détendue beaucoup plus lente, la durée d'interprétation n'apporte rien. (Gazdar, Good, 1982) proposent d'évaluer les effets contextuels en comptant le nombre des implications contextuelles, et l'effort de traitement en comptant le nombre des opérations de déduction. Mais (Sperber, Wilson, 1995) montrent que compter chaque opération revient à ajouter une opération de comptage dans l'effort, et que si l'évaluation résultait d'un comptage, les sujets devraient être capables de porter des jugements absolus sur la quantité d'effet obtenu ou d'effort dépensé. Ce n'est pas le cas. Comme l'argumentent Sperber et Wilson, la pertinence s'avère pour le moins difficilement calculable. Quand elle est représentée mentalement, elle l'est sous la forme de jugements comparatifs ou éventuellement de jugements absolus vagues et généraux, et non sous la forme de jugements absolus fins et précis tels que le sont les jugements quantitatifs. Un autre argument œuvrant contre la calculabilité de la pertinence est son aspect subjectif : effets

et effort dépendent de l'individu, de ses dispositions et de ses expériences.

Certains indices semblent cependant intervenir très fortement et permettent d'envisager des pistes fiables pour une évaluation de la pertinence. Par exemple, plus une hypothèse est forte et plus ses effets contextuels risquent d'être importants. En ce qui concerne l'effort, plus le contexte comprend d'informations et plus l'effort de traitement est important. Ce sont ces types d'indices que nous voulons identifier.

2.3 La pertinence pour la référence dans la communication multimodale

Les travaux que nous menons dans le domaine de l'interprétation de la référence aux objets dans le dialogue homme-machine (Landragin et al., 2002) montrent que tout acte de référence passe par l'activation d'un domaine de référence. Cette notion modélise la focalisation dans un espace attentionnel. Elle permet de justifier l'utilisation de descriptions discriminantes dans des ensembles d'objets plus réduits que le contexte complet. Un domaine de référence est un sous-ensemble contextuel, de nature linguistique, visuelle, gestuelle ou encore liée au déroulement de l'interaction ou à la tâche applicative. De manière simplifiée, il s'agit d'un groupe d'objets dans lequel les composants de l'expression référentielle permettent d'une part d'extraire un référent, d'autre part de préparer un support pour l'interprétation d'une future référence. Ainsi, « *le triangle rouge* » contient deux propriétés dont la combinaison doit être discriminante dans un domaine de référence qui doit, pour justifier cette expression et dans la mesure du possible, comprendre un ou plusieurs triangles non rouges. Après l'interprétation de cette expression, le domaine de référence est structuré avec une première distinction entre les triangles et les autres formes, et une deuxième distinction entre le triangle rouge et les autres triangles. Une expression ultérieure telle que « *les autres triangles* » sera interprétée dans le même domaine, dénotant ainsi une continuité dans la chaîne de référence et, d'une manière générale, dans le dialogue. Cette notion de domaine consiste à regrouper dans un cadre unifié les contraintes provenant de toutes les facettes du contexte. Elle nous amène à nous intéresser au point de vue de la Théorie de la Pertinence. Il nous semble en effet que l'interprétation d'une expression référentielle peut exploiter un critère de pertinence pour identifier un domaine de référence pertinent et un référent dans ce domaine. Il nous semble de plus que l'exploitation du même critère de pertinence en génération automatique d'expressions référentielles permettrait de tenir compte efficacement de la notion de domaine lors de la production des réponses du système.

L'application de la Théorie de la Pertinence à la référence dans le dialogue multimodal pose le problème de la nature des informations contextuelles. Le contexte dans la Théorie de la Pertinence se limite à des propositions traduisant des croyances ou des affirmations qui s'expriment par le langage. Le contexte tel que nous le concevons s'avère beaucoup plus large et hétérogène. Il comprend des informations extra-linguistiques pour lesquelles la Théorie de la Pertinence ne dit pas grand chose. Si celle-ci a été élaborée pour un cadre linguistique, les principes qu'elle propose s'avèrent cependant tout à fait valables pour la communication multimodale en général. Ainsi, de même qu'un énoncé oral porte en lui-même une présomption de pertinence quant à son contenu, nous pouvons considérer qu'une trajectoire gestuelle porte en elle-même une présomption de pertinence quant aux objets qu'elle cible (les demonstrata), et qu'une expression référentielle multimodale porte en elle-même une présomption de pertinence quant à la façon dont elle désigne un référent, cette façon intégrant le recours à un domaine de référence. Les effets contextuels et l'effort de traitement se conçoivent aussi bien pour une action de référence multimodale que pour un énoncé oral, et nous allons le détailler maintenant.

3 Une caractérisation de la pertinence

Nous nous plaçons dans le cadre du dialogue homme-machine à support visuel (qui incite l'utilisateur à s'exprimer par la voix et le geste). Nous nous intéressons tout d'abord aux effets contextuels d'un acte de référence. En considérant que le contexte contient des propositions telles que « deux triangles rouges T_1 et T_2 sont visibles dans la scène », « T_1 a une grande taille par rapport à T_2 », « l'objet le plus saillant visuellement est O_7 », « la dernière action de référence s'est effectuée dans le domaine de référence DR_3 », « DR_3 contient T_1 et T_2 » ou encore « aucun objet n'est visuellement focalisé », nous pouvons avancer l'idée que l'expression référentielle « *le grand triangle rouge* » produit les effets suivants :

1. il y a référence ;
2. cette référence porte sur l'objet T_1 dans le domaine de référence DR_3 ;
3. l'objet T_1 est désormais focalisé visuellement et linguistiquement.

Ce dernier effet suffit à montrer qu'en plus des informations nouvelles liées à la référence, le contexte est modifié. En effet, à partir de la proposition « aucun objet n'est visuellement focalisé » et de la résolution de l'expression référentielle, nous dérivons une seule implication contextuelle : « un seul objet est visuellement focalisé, et cet objet est T_1 ». En considérant comme (Gazdar, Good, 1982) que les effets contextuels s'évaluent avec le nombre d'implications contextuelles, nous obtenons pour notre exemple un score de 1. Un changement de domaine aurait constitué une implication contextuelle supplémentaire (« la dernière action de référence s'est effectuée dans le domaine DR_9 ») et donc un score de 2. Nous choisissons pour l'instant de garder cette approche consistant à compter les implications contextuelles : même si les arguments cognitifs de (Sperber, Wilson, 1995) s'y opposent, ce critère semble plausible d'un point de vue logique et facilement calculable par un système. Notre but n'est pas d'élaborer un système en copiant le fonctionnement cognitif humain, mais d'identifier des critères calculables rendant compte de ce fonctionnement. Le critère de (Gazdar, Good, 1982) pour l'évaluation des effets nous semble donc tout à fait acceptable et adaptable à notre conception du contexte. Il s'agit néanmoins d'une première approche qui reste à approfondir. En particulier, il s'avère nécessaire de spécifier précisément les différents types de propositions contenues dans le contexte.

En ce qui concerne l'effort de traitement d'une référence, nous considérons, à la suite de la définition de Sperber et Wilson, qu'il est proportionnel à la complexité des différentes informations intervenant lors du traitement, à leur accessibilité, et à l'importance des interactions entre ces informations. Devant le nombre important de paramètres, nous choisissons une méthode incrémentale pour leur prise en compte : lors d'une première étape, nous ne ferons intervenir que les informations provenant de la perception visuelle, pour intégrer lors d'une deuxième étape les informations apportées par l'expression référentielle verbale. Nous aurons alors une base pour la modélisation de l'interprétation de la référence langagière. Pour celle de la référence multimodale, la troisième étape intégrera les informations apportées par les trois modalités.

3.1 En considérant la perception visuelle

La perception de la scène visuelle influe sur le comportement de l'utilisateur, en particulier sur sa façon de référer aux objets qui y sont visibles. S'il perçoit des objets isolés il va par exemple être tenté de référer individuellement à ces objets, alors que s'il perçoit des groupes il va peut-être désigner des ensembles de référents avant d'en extraire un élément. Nous pro-

posons dans (Landragin et al., 2002) un algorithme de classification automatique des objets pour l'identification de groupes perceptifs en suivant les principes de la Gestalt (Guillaume, 1979). Ainsi, la proximité, la similarité et la continuité dans la disposition des objets permettent de construire une structuration arborescente de la scène en groupes perceptifs. Cette structuration constitue une information à stocker dans le contexte, par l'intermédiaire de propositions telles que « T_1 et T_2 forment un groupe perceptif G_1 sur le critère de proximité spatiale et de similarité de forme » et « G_1 et G_2 sont les deux sous-groupes de G_3 ». Il s'agit d'une représentation cognitive du contexte visuel, permettant la dérivation d'implications contextuelles et par conséquent le calcul des effets contextuels tel que nous l'avons abordé.

Quant à l'effort de traitement, il est lié tout d'abord à la complexité de cette structuration en groupes. Dans les critères permettant de l'évaluer, nous retenons la profondeur de la structure arborescente (c'est-à-dire le nombre de partitions possibles de la scène en groupes perceptifs), et le nombre de nœuds présents dans celle-ci. Nous proposons également de tenir compte du nombre d'objets visibles, ainsi que d'un indice correspondant à la diversité de ces objets (par exemple le nombre de de formes, de couleurs et de tailles différentes).

3.2 En considérant la perception et l'expression verbale

Il s'agit ici d'évaluer la pertinence d'une expression référentielle langagière. Nous avons déjà étudié les effets contextuels, et nous proposons de décomposer l'effort de traitement selon les traits suivants :

1. adéquation de l'expression verbale avec l'intention de référence en contexte applicatif ;
2. adéquation de l'expression avec l'intention de référence en contexte dialogique ;
3. complexité de la perception visuelle ;
4. complexité de l'expression verbale ;
5. effort nécessaire pour isoler le(s) référent(s) à l'aide de la perception, de l'expression verbale et de l'historique de l'interaction.

Le premier trait correspond à l'effort nécessaire pour intégrer l'expression référentielle verbale dans le contexte de tâche. Dans le corpus Magnét'Oz (Wolff, 1999) qui nous a servi pour identifier les phénomènes de référence induits par une tâche très contrainte, l'interaction consiste à ranger des objets selon leur forme. La seule action possible est le rangement, ce qui incite l'utilisateur à produire des commandes telles que « *mets la forme claire dans la deuxième boîte* ». Dans ce contexte, aucune action ne peut être appliquée à un ensemble hétérogène et une expression désignant deux objets de forme différente s'avère incongrue. En raison de cette incongruité, l'effort de traitement de cette expression est élevé.

Le deuxième trait correspond à l'effort nécessaire pour contextualiser l'expression verbale dans le contexte dialogique. Par exemple, « *le rouge* » après « *le triangle vert* » demande un effort particulier pour la résolution de l'ellipse. D'un autre côté, l'expression élidée comporte moins de mots et l'effort lié à sa complexité en diminue d'autant (voir le quatrième trait).

Le troisième trait correspond à l'effort nécessaire pour prendre connaissance de la scène. C'est ce trait que nous avons détaillé lors de l'étape précédente.

Le quatrième trait s'évalue à l'aide de critères syntaxiques tels que le nombre de mots et la complexité de la structure syntaxique, par exemple la profondeur et le nombre de nœuds de l'arbre syntaxique. Les mots ayant plus ou moins d'importance, des pondérations peuvent être

envisagées. Le cas particulier des connecteurs pragmatiques est intéressant : selon (Reboul, Moeschler, 1998), les connecteurs (conjonctions de coordination, « *parce que* ») jouent un rôle au niveau de la facilitation du traitement de l'information. Leur fonction est de minimiser les efforts cognitifs. Bien que Moeschler les analyse dans une proposition et non dans un simple groupe nominal, nous pouvons tenir compte de cette facilitation, par exemple en ne comptant pas le mot « *et* » dans une coordination, ou en le pondérant de manière très faible.

Le cinquième trait s'évalue à l'aide de critères tels que la difficulté à repérer dans la scène les objets vérifiant les propriétés exprimées dans l'expression verbale. Par exemple, « *le triangle rouge* » demandera moins d'effort que « *le petit triangle* » si le référent est le seul objet rouge de la scène alors qu'il n'est pas le seul petit objet (les autres petits objets n'étant pas des triangles). L'historique de l'interaction est pris en compte de la manière suivante : d'une part en ajoutant un effort lié au rappel d'un contexte visuel antérieur, par exemple pour l'interprétation de « *remets le triangle vert* » après « *efface les triangles* » ; d'autre part en ajoutant un effort lié au rappel d'informations sémantiques et pragmatiques utiles pour la résolution de la référence, par exemple pour l'interprétation de « *les autres* ».

3.3 En considérant la perception et l'expression multimodale

Pour une expression multimodale, l'effort de traitement se décompose selon les traits suivants :

1. adéquation de l'expression multimodale avec l'intention de référence en contexte applicatif ;
2. adéquation de l'expression avec l'intention de référence en contexte dialogique ;
3. complexité de la perception visuelle ;
4. complexité de la ou des expressions référentielles verbales ;
5. complexité du ou des gestes ;
6. effort nécessaire pour associer le ou les gestes à la ou aux expressions verbales ;
7. effort nécessaire pour isoler le(s) référent(s) à l'aide de la perception, de l'expression multimodale et de l'historique de l'interaction.

Le premier trait correspond à l'effort nécessaire pour intégrer l'expression référentielle multimodale dans le contexte de tâche. Il se déduit de l'effort correspondant dans l'étape précédente.

Le deuxième trait correspond à l'effort nécessaire pour intégrer l'expression multimodale dans le contexte dialogique. Par exemple, « *celui-là* » après « *celui-ci* » demande un peu moins d'effort que « *celui-ci* » dans la même situation (avec un geste à chaque fois).

Les troisième et quatrième traits ne diffèrent pas de ceux des étapes précédentes.

Le cinquième trait s'évalue à l'aide de critères tels que la longueur et la complexité de la trajectoire gestuelle. La longueur dépend du gabarit des objets, et la complexité du nombre de singularités, une singularité étant selon le modèle de (Bellalem, 1995) une rupture d'homogénéité pour une des propriétés de la trajectoire (les principales propriétés étant la courbure et la vitesse), cette rupture dénotant une intention sémantique.

Le sixième trait s'évalue à l'aide de critères tels que la composition de l'expression multimodale (nombre de gestes, nombre d'expressions verbales), la qualité de la synchronisation temporelle entre geste(s) et expressions(s) verbale(s), la présence d'une ambiguïté dans l'une des modalité (ambiguïté résolue par l'autre modalité au prix d'un certain effort), ainsi que le choix du déterminant et d'éventuels marqueurs déictiques et adjectifs numéraux. Par exemple, un démonstratif s'associe plus naturellement avec un geste qu'un défini, et demande donc moins d'effort.

Le dernier trait s'évalue à l'aide de critères tels que la difficulté à repérer dans la scène les objets désignés par l'expression multimodale. Par exemple, l'expression multimodale composée d'un geste désignant un triangle et associé à « *ce triangle* » demande moins d'effort que l'expression multimodale composée du même geste associé à « *cet objet* », pour laquelle intervient une récupération visuelle de la catégorie. L'historique de l'interaction est pris en compte en ajoutant un effort lié au rappel d'un contexte visuel antérieur ou au rappel d'informations linguistiques utiles pour la résolution de la référence (comme nous l'avons vu lors de l'étape précédente). L'historique des trajectoires gestuelles déjà effectuées intervient également pour diminuer l'effort de traitement d'une trajectoire souvent utilisée.

4 Exploitation en compréhension et en génération

Notre but ici n'est pas de proposer une méthode de quantification absolue, mais de montrer comment des jugements relatifs permettent de comparer la pertinence de plusieurs expressions, trait par trait. Dans la figure 1-A, comparons par exemple les pertinences de quelques expressions désignant le groupe isolé de deux carrés à gauche de la scène. « *Les deux carrés* » s'avère légèrement plus pertinente que « *les carrés de gauche* ». En effet, les contextes applicatif, dialogique et visuel étant invariants, reste à comparer la complexité des expressions et l'effort nécessaire pour isoler les référents. Or la scène visuelle est structurée en deux groupes dont un à gauche comprenant deux carrés. Les deux expressions demandent donc le même effort pour isoler ce groupe. « *Les carrés de gauche* » étant syntaxiquement légèrement plus complexe que « *les deux carrés* », celle-ci se voit attribuer l'effort minimum et par conséquent la meilleure pertinence. Cet exemple illustre l'importance de la structuration de la scène visuelle : *a priori*, « *les deux carrés* » est ambigue puisque cinq carrés sont visibles. Or, comme n'importe quel locuteur peut le constater, cette expression est tout à fait compréhensible et désigne sans ambiguïté le groupe comportant deux carrés, l'emploi de « *deux* » s'appuyant sur la cohésion du groupe. L'exploitation de la pertinence permet au système de retrouver ce mécanisme.

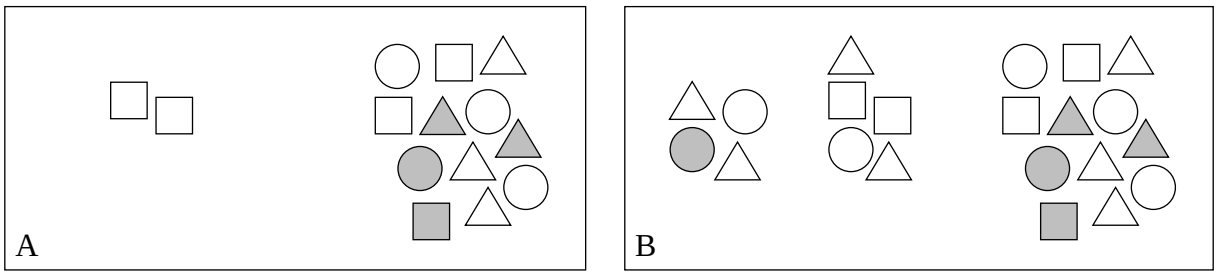

Figure 1: Deux exemples de scènes visuelles

Si nous considérons maintenant l'expression « *les deux carrés à gauche* », nous remarquons immédiatement la plus grande complexité syntaxique et donc la plus faible pertinence. Avec « *les deux carrés blancs* », la complexité augmente, de même que l'effort pour isoler les référents : dans le groupe de droite, le nombre de carrés blancs est également de deux, et ceci peut introduire un doute dans l'interprétation (du fait du sur-emploi d'informations). Considérons d'autre part l'expression référentielle « *ces carrés* » associée à un geste ostensif entourant le groupe de deux carrés à gauche. Si les traits à évaluer sont plus nombreux pour une expression multimodale, l'effort propre à chaque trait s'avère en revanche très faible : la complexité de la perception visuelle est réduite par la focalisation spatiale due au geste, l'expression verbale

est très simple, le geste d'entourage également (car beaucoup de place est disponible), l'effort nécessaire à l'association du geste à l'expression est très faible du fait de l'emploi non ambigu du démonstratif (car aucune anaphore n'est possible dans le contexte dialogique). Quant à l'effort nécessaire pour isoler les référents, il s'avère faible du fait que l'hypothèse consistant à assimiler les référents aux demonstrata est immédiatement vérifiée. « *Ces carrés* » associée à un geste semble donc être l'expression multimodale la plus pertinente. Ce n'est pas le cas dans la scène de la figure 1-B, du fait de la nécessité d'une délimitation précise des demonstrata.

En compréhension automatique, de telles comparaisons de pertinences présentent plusieurs intérêts : premièrement d'aider la résolution de la référence en apportant un point de vue sur les processus intervenant dans la compréhension (la pertinence permet de rejeter certaines hypothèses sur les domaines et sur les référents) ; deuxièmement de remettre en question et comprendre le résultat d'une référence quand l'intention référentielle est clairement identifiée mais que la pertinence est faible (il peut être utile d'identifier le ou les traits pour lesquels un manque de pertinence est flagrant et d'en tirer des conclusions dans la génération de la réponse) ; troisièmement de retrouver la source d'une incompréhension ou d'un malentendu dans le dialogue (en considérant que cette source risque de correspondre à une pertinence faible).

En génération automatique, la pertinence d'une expression référentielle verbale ou multimodale s'évalue par rapport aux différents domaines de référence activés, par rapport à elle-même et à l'intention de référence. Disposer d'une telle évaluation permet d'une part de détecter les éventuelles ambiguïtés entre cette intention et d'autres intentions, entre tel et tel référent, et d'autre part de déterminer parmi les expressions possibles celle qui produit l'effet escompté avec le minimum d'effort pour l'interlocuteur. La pertinence constitue à ce titre un critère idéal pour un choix parmi plusieurs possibilités. Elle diffère de la pertinence en compréhension : dans l'exemple de la figure 1-A, l'expression « *les carrés de gauche* » s'avère plus pertinente en génération que « *les deux carrés* », car aucune ambiguïté n'apparaît sur le domaine de référence et sur les référents. L'information de cardinalité peut s'avérer indispensable, par exemple si le geste désigne un domaine de référence dans lequel les référents doivent être extraits, ou s'il s'agit d'un pointage imprécis vers un groupe perceptif. Dans l'exemple de la figure 1-B, imaginons que l'on veuille désigner les triangles contenus dans le groupe perceptif hétérogène à droite de la scène. L'association d'un geste désignant globalement le groupe, par exemple un entourage, et de l'expression verbale « *les triangles* », peut amener à une extraction incomplète des référents. En les cherchant bien, ceux-ci sont au nombre de cinq, deux d'entre eux étant de couleur grise et les trois autres de couleur blanche. Dans un même contexte, l'expression « *les cinq triangles* » s'avère plus pertinente car elle permet à l'interlocuteur d'être sûr de ne pas oublier un référent. Si nous considérons maintenant le groupe perceptif situé au milieu de la scène, l'association d'un geste désignant globalement le groupe et de l'expression « *les triangles* » s'avère aussi pertinente à générer que le même geste avec « *les deux triangles* ». Le groupe ne comporte en effet que cinq objets, parmi lesquels les deux triangles se distinguent facilement. Or le pluriel « *les* » suffit à déterminer que le résultat de la référence concerne bien les deux triangles possibles. L'adjectif numéral « *deux* » n'apporte aucune information supplémentaire, il ne fait qu'augmenter la complexité de l'expression.

5 Conclusion et perspectives

Les exemples que nous avons détaillés montrent à quel point le moindre indice venant de la perception visuelle ou de l'emploi d'un pluriel ou d'un adjectif numéral intervient dans la com-

préhension et la génération d'une expression référentielle. Face à la multitude d'indices, nous avons proposé une caractérisation en ensembles de traits et une méthode consistant à comparer l'effort de traitement lié à chacun de ces traits. Nous ne présentons pas encore de véritable algorithme, mais quelques pistes solides pour un futur modèle basé sur une quantification de la pertinence pour les actions de référence langagières ou multimodales.

Dans cette première proposition, nous n'avons considéré l'intention référentielle qu'au moment précis de la référence. Or celle-ci peut être effectuée dans un certain but qui n'apparaîtra que plus tard. Évaluer la pertinence sans tenir compte de ce but communicatif s'avère insuffisant. Par exemple, l'emploi de « *l'un* » a un effet linguistique particulier dû à ce qu'on attend ensuite « *l'autre* ». La tâche intervient également dans le sens que si l'utilisateur s'occupe d'un triangle dans un ensemble de deux, le deuxième va sans doute être traité bientôt, du moins si la tâche incite à traiter les objets par catégorie. Enfin, le contexte visuel intervient aussi avec les notions de saillance et de ligne de force dirigeant le regard : une référence à un objet saillant se caractérise par un effort de traitement réduit ; une référence à un objet situé au départ d'une ligne entraîne l'hypothèse d'une référence ultérieure à l'objet suivant dans la direction amorcée.

Avec nos caractérisations, nos exemples et les perspectives citées ci-dessus, nous montrons que la formalisation de la pertinence est encore un objectif à long terme, car faisant intervenir tous les paramètres intervenant dans la communication d'une intention. Nous espérons que les traits proposés, ainsi que les pistes données pour leur quantification, constitue une base intéressante pour de futurs travaux.

Références

Akman, V., Surav, M. (1995), Contexts, Oracles, and Relevance, Proceedings of the *AAAI-95 Fall Symposium on Formalizing Context*, 23-30.

Bellalem, N. (1995), 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.

Blutner, R. (2000), Some Aspects of Optimality in Natural Language Interpretation, *Journal of Semantics*, Vol. 17(3), 189-216.

Gazdar, G., Good, D. (1982), On a Notion of Relevance, In: Smith, N. (Ed.), *Mutual Knowledge*, London, Academic Press, 88-100.

Guillaume, P. (1979), *La psychologie de la forme*, Paris, Flammarion.

Landragin, F., Salmon-Alt, S., Romary, L. (2002), Ancrage référentiel en situation de dialogue, *Traitement Automatique des Langues*, Vol. 43(2), 99-129.

Reboul, A., Moeschler, J. (1998), *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*, Paris, Armand Colin.

Sperber, D., Wilson, D. (1995), *Relevance. Communication and Cognition (2nd edition)*, Oxford UK & Cambridge USA, Blackwell.

van Rooy, R., (2003), Relevance and Bidirectional OT, In: Blutner, R., Zeevat, H. (Eds.), *Pragmatics in Optimality Theory*, Palgrave Macmillan.

Wilson, D. (1992), Reference and Relevance, *UCL Working Papers in Linguistics*, Vol. 4, 165-191.

Wolff, F. (1999), Analyse contextuelle des gestes de désignation en dialogue homme-machine, Thèse de doctorat, Université Henri Poincaré de Nancy.