Relevance and Perceptual Constraints in Multimodal Referring Actions

Antonella de Angeli, Frédéric Wolff, Patrice Lopez, Laurent Romary

Paper 1

Relevance and Perceptual Constraints in Multimodal Referring Actions

Antonella De Angeli*, Frédéric Wolff†◇, Patrice Lopez◇ and Laurent Romary◇

◇LORIA, BP239, 54506 Vandoeuvre-lès-Nancy, France

*Department of Psychology, University of Trieste, Via dell'Universit, 7 I-34123 Trieste, Italy

†Alcatel Business Systems, Corporate Research Center, 1 Route du Docteur Albert Schweitzer, 67400 Illkirch-Graffenstaden, France

{angeli, wolff, lopez, romary}@loria.fr

Abstract. This paper presents a first attempt to score the relevance of multimodal referring expressions within a task oriented environment. It is based upon the application of the ecological approach to multimodal system design, which in particular implies that perception has to play a central role in the understanding of a gestural designation. The results of an experimental work is presented, together with the scores obtained by combining the linguistic characteristic of the referring expression and the properties of the corresponding gestures (if any). Even if the calculus that we present has to be refined it seems to be already suited to validate our approach regarding the importance group salience and access type in the choice of a referring mode.

1 Introduction

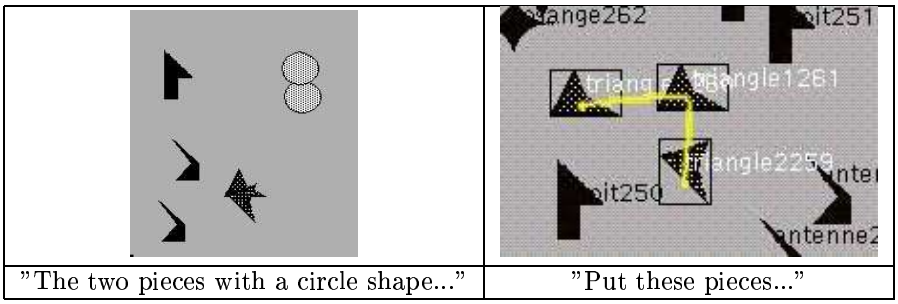
Referring to objects spread on a graphical interface is a typical action in Human-Computer Interaction (HCI). In the direct manipulation paradigm, this action is performed by a simple mouse-mediated pointing as a selection process: interaction is ambiguity-free, but highly restricted. As a new generation of multimodal systems begins to evolve, the number of communicative actions available for indicating visual-targets drastically increases and allows the user to express his intention rather than to perform elementary actions. References can be based on auditory signals (verbal input), motor-visual signals (gestural input), or on a combination of them (multimodal input). Moreover, each type of input can be exploited by a great flexibility of forms. As an example, consider the following multimodal inputs extracted from a corpus collected by a simulation experiment (Wolff (99)). Very different gestures and verbal utterances perform the same communication goal: referring to a group of targets (see Fig. 1.1).

Despite their clear usability advantages, the design of multimodal systems still opens original and challenging problems to HCI researchers. In particular, it requires to develop innovative interaction paradigms to constraint the high variability of natural communication inside computational capabilities. The efficiency of these paradigms strongly depends on their compatibility with cognitive constraints affecting spontaneous behaviour. Indeed, although adaptation is a fundamental human ability, several aspects of communication escape conscious control and involve hard-wired or automatic processes. This is the case for instance of intonation, disfluencies, kinaesthetic motor

Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief at ESSLLI XI
E. André, M. Poesio & H. Rieser (editors).

◇LORIA, BP239, 54506 Vandoeuvre-lès-Nancy, France

*Department of Psychology, University of Trieste, Via dell'Universit, 7 I-34123 Trieste, Italy



"The two pieces with a circle shape..."	"Put these pieces..."

Figure 1.1: Two examples of complex multimodal reference to a group of targets.

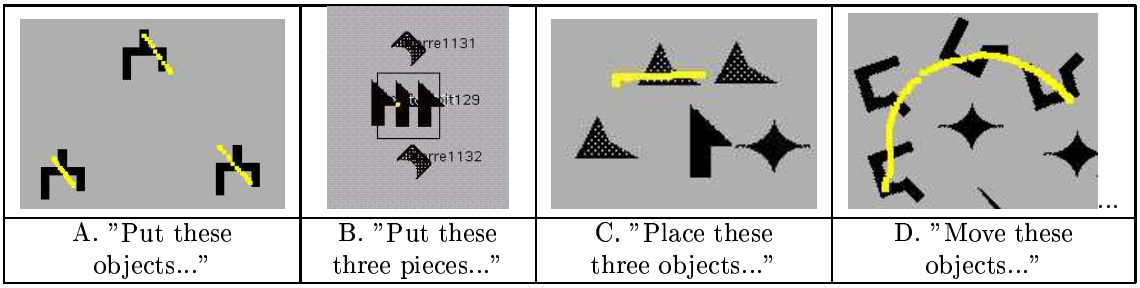
control, cross-modal integration and timing. Automaticity occurs over extensive practice with an activity, when specific routines are built up in the memory. Being performed beyond conscious awareness, automatic processing is effortless and fast, but it requires a strong effort to be modified. Even when people learn new solutions (i.e., set up alternative routines in the memory), as soon as they are involved in a demanding situation, they spontaneously return to their old ways. As a consequence, errors are very likely to occur. Taking into account actual human capabilities and constraints, it is unrealistic to expect that users will be able to adapt all or some parts of their behaviour to suit system limitations. On the contrary, multimodal systems should favour automatic behaviour by allowing the user to express directly intentions. In this way, the effort required to monitor his expression, plan and perform corresponding actions remains minimal. Usable systems require thus a deep understanding of the factors affecting human spontaneous behaviour.

To cope with communication variability, designers need to know the conditions under which specific actions are likely to be produced. Such knowledge can drive the design of effective architecture capable of using all the appropriate cues needed to understand user's communicative intentions. Since HCI is highly different from human-human communication (Jönsson and Dählback (88)), empirical research, expecially in the form of early simulation, is instrumental to the formalisation of a theory of multimodal interaction. Elsewhere (Wolff (99), de Angeli and *al.* (98)) we presented the ecological approach to multimodal system design, an innovative theoretical framework explaining communication variability as a function of cognitive constraint and contextual knowledge. The major goal is linking gesture variability to visual perception. At this aim, the approach proposes to revise gestural communication by introducing it into the perception-action cycle (Neisser (76)), a well established psychological framework describing how action planning and execution is controlled by perception, and how perception is constantly modified by active exploration of the visual field. The basic assumption of our proposal is that gestures are virtual actions (Kita (in press)), re-enactments of real activities in a virtual space. Referring gestures are virtual actions aimed at directing listener's attention towards a target. They do not modify the physical environment where they are produced, as would make grasping the object and moving it in front of the listener. They modify the dialogue context, inducing listeners to shift the focus of their attention towards the target : this corresponds to the semiotic function of gesture. The cyclic nature of cognition was found to provide a powerful conceptual structure for understanding referring gestures (Wolff (99), De Angeli and *al.* (99)). In this paper, we attempt to extend the ecological approach to multimodal system design to cope also with verbal language variability. In particular, following the Relevance Theory we try to understand the link between speech in a discourse context and gesture in a perceptive context. The potentialities of this extension are confirmed by some preliminary results of a simulation study. Exploiting the perceptual constraints people, tend to modify the effect of their utterances.

2 Theoretical framework

The ecological approach is an established psychological theory to perception, cognition and action (Gibson (79)) now adapted to multimodal system design (De Angeli and *al.* (99)). According to ecological psychology, the perception-action cycle is mediated by *affordances*, optic information about objects that convey functional properties. Affordances represent powerful cues of action. They are not properties of the object, but relations derived by the encounter between information coming from the object and the repertoire of physical actions available to the observer. The mutuality of organism-environment relationship is a major theoretical assumption of the ecological approach. Affordances are characteristic of the environment relative to specific individuals. The same physical layout will have different affordances for different individuals, insofar each individual has a different repertory of acts (Gibson (79)). For example, a stone may afford being thrown by an adult, but not by a child. An extension of the concept of affordances to the world of design was initially proposed by Norman (88), but its potentialities in the domain of natural communication is still little understood.

The ecological approach to multimodal system design is aimed at extending the concept of affordances to explain multimodal actions variability. Its basic assumption is that gestures are determined by the mutuality of information provided by the object to be referred to, and the set of movements available to the speaker. The innovative aspect of the approach is the importance attributed to visual perception as a fundamental cue for explaining communicative actions variability. The interplay between perception and gestures has already been proved in de Angeli and *al.* (98), Wolff (99), Wolff and *al.* (98) and De Angeli and *al.* (99). Some examples relative to group identification are reported in Fig. 1.2.



A. "Put these objects..."	B. "Put these three pieces..."	C. "Place these three objects..."	D. "Move these objects..."

Figure 1.2: Examples of group designations.

Form, granularity and size of gestures tend to be adapted to the visual layout of the target. Consider Fig. 1.2 A, where even at the cost of producing very unusual movements, the user adapt her gesture to the group structure. Visual cues are instrumental to resolve ambiguities (i.e., competition of possible candidates to reference identification) often coming from spontaneous gesturing. For example, granularity ambiguities derives from a non 1-to-1 relation between referred area and gesture extent. As shown in Fig. 1.2 B, when the group salience is very high, the gesture can be highly simplified: a small pointing refers to the entire group. The salience of the group is an important predictor of access type. High salient groups are likely to induce group access, (the gesture shows the perimeter or the area of the group); whereas low salient groups afford almost only individual access (a number of gestures indicates the target one by one). An example of intra-group ambiguity is also illustrated in Fig. 1.2 C. Here, the interpretation is complicated by a strong ambiguity in choosing either the individual percept indicated by the gesture or the group indicated by the verbal categorization. Note that only the perceptual context permits to identify the three appropriate targets (object with the same shape). Spontaneous gesture may present also pattern ambiguities (for instance a gesture with a circling form, which is in fact a targeting i.e. a gesture passing through the objects to be referred to), that can be removed by considering the object layout (i.e., the visual context).

Another basic assumption of our model is that human beings are efficient communicators. They

design their utterances to maximise communication relevance. According to Sperber and Wilson (86), relevance is a property of inputs to cognitive processes, embodying the notion of cognitive effect and processing effort. Processing an input yields some cognitive effects when it affects the context of previous assumptions on the topic. However, processing the input involves some mental effort. Applying the Relevance Theory to multimodal communication we expect that users select the most efficient referring strategies, i.e., the one that requires the minimal effort to be understood with the maximal effect. From a general point of view, a referring action produces contextual effects on the system at the price of some efforts to obtain them. The quality of a reference, its relevance, can somehow be viewed as a measure of the ratio effects/efforts, the best referring hypothesis appearing as one having the biggest ratio value associated with. The main problem is the manner of weighting effects and efforts, above all in multimodal communications. Applying the concept of affordances to multimodal communication, we expect that referring actions will be affected by visual characteristics of the target. We assume that, to maximise communication relevance, semantic features are distributed across language, vision and gestures. All together, these modalities supply different and complementary information for meaning composition. The innovative aspect of our approach is the importance given to visual perception in multimodal system design. Traditionally, multimodal interfaces are blind: they do not take into account the visual context on which communicative actions are performed and references are resolved mainly by considering the dialogue context. On the contrary, we assume that information about perceptual field organisation are important to understand referring behaviour.

The main complexity of the multimodal reference is the heterogeneity of the referring expressions, exploiting both advantages of the language and the gesture. We explained that in the context of a multimodal spoken interface, we need to consider the integration of the gesture and the language. We try here to study the variability of referring expressions at the light of the Relevance Theory because this theory concerns a very general purpose and we argue that it can be applied to understand heterogeneous multimodal referring expression. Our ambition is to study :

- first, if our estimation of the relevance of referring expressions suits with the actual data (Is the chosen multimodal referring expression corresponds to the more relevant potential one? Is it the result of a training of the user?) ;
- second to identify the factors and constraints that could help finding dynamically the more relevant referring expression. Such factors would allow the optimisation and the determinization of the reference resolution process in the context of intelligent multimodal interfaces.

Our approach is empirical and this work is more a prospective on-going study than a general model. This paper is aimed at evaluating the effect of perception on the verbal part of multimodal referring actions according to a descriptive scoring methodology.

3 The simulation study

To evince regularities linking perception to multimodal actions, a Wizard of Oz simulation was run Wolff (99), Wolff and *al.* (98). Seven students from the University of Nancy participated in the simulation as volunteers. They were French native speakers. Engaging a dialogue with the simulated system, users were required to move groups of objects into appropriate folders. Interaction was based on speech and gestures, mediated by a microphone and an electronic pen. Targets were abstract-shape figures that inhibit pure linguistic input de Angeli and *al.* (98). They could be targets or distractors. Targets were collections of two or three identically shaped stimuli that have to be moved into the box displaying their figure. Distractors were exclusively used in relation to perceptual field organization and did not have to be moved

Two experimental conditions were tested to study intra-subject variability. Perceptual organization of the visual field was handled according to the principles of the Gestalttheorie Wertheimer (1922), Kanizsa (79), describing the laws underlying spontaneous grouping. The manipulation was based on similarity (objects are grouped on the basis of their salient physical attributes, such

as shape and color); proximity (elements are grouped on the basis of their relative proximity); good continuation (shapes presenting continuous outlines have a better configuration than those with discontinuous ones).

The experiment contrasted High Group-Salience with Low Group-Salience. In the first condition, targets were easily perceived as an homogeneous group, clearly separated from surroundings called distractors. Proximity and good continuation supported similarity. In the low-salience condition, targets were spontaneously perceived as elements of a broader heterogeneous group including distractors. Proximity and good continuation acted in opposition to similarity.

4 Data coding and hypotheses

4.1 Linguistic scoring

Our approach of Relevance Theory modelling applied to multimodal reference is to compute both the cognitive contextual effects and efforts. Effects are obtained by two notions: the number of new data deduced from the referring action and the interest of these data (linked to the intention of reference). In order to study verbalisations occurring with gestural references, we have parsed the transcribed corpora with a Lexicalized Tree Adjoining Grammar (LTAG) and computed a score for each referring expression. We used the definition of derivation of Schabes and Shieber (94) and linguistics principles of Abeillé (93) for LTAG, these choices allow the result of the parsing (the derivation tree) to be equivalent to a semantic dependencies graph. Because each node in the derivation tree represents a predicate, the number of new data is obtained by the number of nodes in this dependency tree resulting from the parsing of the referring expression.

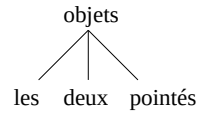
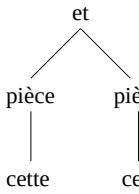
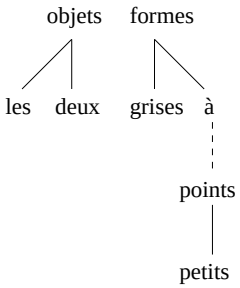
verbal referring expression	Derivation tree	new data scoring	contextual effect scoring
les deux objets pointées (the two pointed objects)		4	2.5
cette pièce et cette pièce (this piece and this piece)		5	3
les deux objets, formes grises à petits points (the two objects, grey shapes with small grey points)		8	5.5

Figure 1.3: Examples of linguistic scoring.

The importance of the predicates for the reference task is taken into account by a weighting:

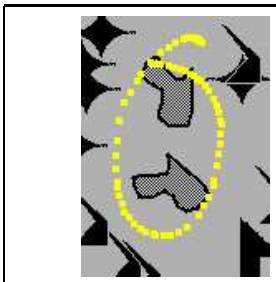
this weighting w is subjective and depends on the application. For example, words such as *object* or *figure* provides no information for the identification of a piece in the visual scene and correspond to a score of 0. We consider that for the semantic head of the linguistic referring expression, we have $w = 0$ for abstract nouns in the application context. For other noun the weight is 1. Considering the simulation application, the list of transcribed abstract nouns is the following one: *objet* (object), *forme* (form), *pièce* (piece) and *figure* (figure). For the weighting of the modifiers we have introduced these general rules depending on how much the object is specified:

- $w = 0$, for an indefinite determiner (a) which provides no information for the reference task ;
- $w = 0.5$, for definite determiner (the) indicated a more precised denomination of the referred object
- $w = 0.5$, for a deictic mark concatenated at the end of the word (in French *objet*-**là**) which determine a link with a referring gesture or a salient object in the dialogue ;
- $w = 1$, for demonstrative article ($this$, $those$) which allow to determine a direct relation between the verbal reference and a referring gesture and indicate that the reference is expressed by the two modalities ;
- $w = 1$, for demonstrative pronouns : $ceux - ci$ (these one), *celui-là* (this one) for similar reasons to the previous item ;
- $w = 1$, for adjectives as they provide information about an object to be referred.

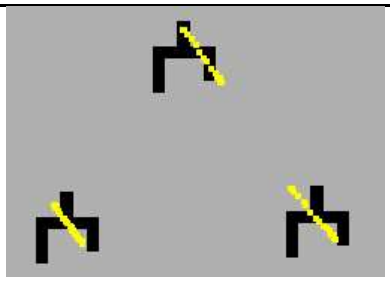
Considering these rules, we have for example the following results : *les objets-là* = *ces objets* = *ceux-là* < *cet objet-là*. For an expression which gives no information about the object to refer as *un objet* (an object) we have $w = 0$ and so a null contextual effect. On Fig. 1.3, we illustrate this scoring with some examples of verbal referring expressions, their resulting derivation trees and the corresponding effect evaluation.

4.2 Gestural scoring

We have realised the same kind of scoring on gestural designation: the gestural part of each multimodal command was tabulated according to the strategy adopted to identify the corresponding group. Two general strategies were found: group-access (reference is achieved by showing the perimeter or area of the group) and individual-access (reference is achieved by indicating each element one by one).



. "ranger ces deux-là..."
(tidy these two one...)



B. "mettre cette piece, cette piece et cette piece..."
(put this piece, this piece and this piece...)

Figure 1.4: Examples of gestural scoring.

In the first case, users intention is rather to point out the group in which to find the referents. In our corpus, gestures are scored as group accesses when more than one object is accessed by only one

gesture (see Fig. 1.4 A). No gestural category was particularly dedicated to such accesses: single circlings, targetings and even pointings were used to refer to groups of objects. This feature leads especially in the cases of targeting and pointing to ambiguities as shown previously. Removing them supposes to take into account implicit information on which the user relies his expression i.e. perceptual groups De Angeli and *al.* (99) . Group accesses could be compared to linguistic plural expression as a way to refer to many objects with one single act. On the contrary, individual accesses were defined when gesture was used to refer to only one object. This strategy could be characterised as a singular expression, even if the user produces sometimes coordination of individual gestures (see Fig. 1.4 B).

4.3 Multimodal scoring

The total effect of a multimodal referring expression is the sum of the effects corresponding to the two modalities. As a general hypothesis we assumed that the role of perception in multimodal actions is twofold. Implicit cues (i.e., objects visual layout) and explicit cues (i.e., intentional gestures) may affect the verbal part of the referring expression. Here, we tested both the effects by: (a) comparing the cognitive effect of verbal expressions produced in the two Group Salience conditions (implicit cues); (b) comparing the cognitive effect of verbal expressions produced together with the two gestural Access Type (explicit cues).

5 Results

A corpus of 98 multimodal commands has been analysed. The average value of the effect is 2.66, with a standard deviation of 1.17. The distribution ranges from 1 to 6 and is affected by a substantial positive skewness, which is difficult to be normalised by transformation. Because of this statistical characteristic, the two experimental hypotheses are separately tested using non-parametric statistic.

As regards the implicit cues hypothesis, a Mann-Whitney U test shows a strong influence of Group Salience on effect, $U = 687(N = 87)$; $p < .05$. In the Low-Salience condition, users produced complex verbalisations; whereas in the High Salience condition, verbalisations were significantly simplified. This result supports our hypothesis that humans spontaneously plan their verbal references having a representation of the visual context in mind. When the group is easily perceivable, they need less complex referring expression to convey their intentions, since they can exploit implicit information. As a corollary, we can deduce that humans naturally attribute to artificial interlocutor their same perceptual capabilities.

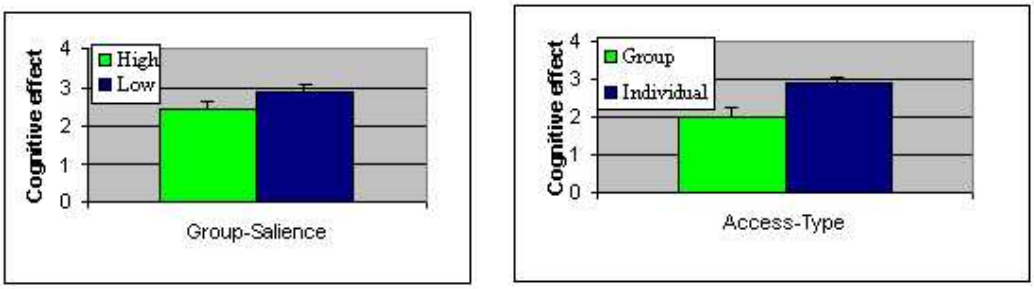


Figure 1.5: Cognitive Linguistic Effect as a function of Group-Salience and Access-Type (Error bars represent mean standard errors).

The influence of explicit visual cues on cognitive effect was straightforward, $U = 324,5(N = 84)$; $p < .001$. The analysis shows that, for multimodal reference, complexity of gestural and verbal expressions are linked. Simple designation (direct group access) comes jointly with simple

verbal production (generally a deictic); complex designation (individual access that indirectly build the group) need complex verbalisations. In the last case, the linguistic part permits to provide temporal links for simplifying the complex spatial designations. The results is consistent with previous findings Wolff and *al.* (98) showing a high correlation between the reference strategies adopted by verbal and gestural languages to identify targets. In particular, the gestural group access was always accompanied by a plural deictic anchor or target descriptions (i.e., "these objects", "the two isolated objects, the two forms"). Moreover, only 1 out of 3 multimodal expression were formed by an individual access accompanied by a plural verbal reference. In such a reference, the linguistics part, and mainly deictics, allows to provide temporal links for the complex spatial designations.

6 Future work

The contextual efforts of a referring action would be modelled by three notions: the number of inference steps to deduce the new data from the referring action, the complexity of this deduction and the date of used information to deduce the new data (memory access). Such a score is of course more complicated to compute than the contextual effect which only requires a descriptive methodology. Evaluation of the contextual effort requires a general theory of reference that takes into account multimodal communication. In order to interpret correctly this kind of natural referring actions, the system has to represent and use jointly for each objects of the application different kind of information: linguistics, pragmatic and perceptive. To conclude considering the proposed contextual effect scoring, we argue that the Mental Representation Theory (Reboul (98)) provides a framework able to compute contextual efforts, to represent such information and use them for reference resolution in accordance with the Relevance Theory. The goal of this general theory of reference, inspired by Sag and Hankamer (84) approach, is to explain the processing of object and event reference resolutions using a specific world representation called Mental Representations. This data structure gathers all information needed to resolve references, including perceptive information. Using a small set of rules, these Mental Representations are updated after each new utterances and allow to solve heterogeneous references as spatial designation.

7 Conclusion

In this paper, we have tried, in the context of a sound theoretical background, to evaluate the precise role of perception in multimodal systems. From the field of psychology, we have extended the ecological approach to show how gestural designation could be seen as the realization of specific affordances induced by Gestalt properties of objects. From a more linguistical point of view, we have extended relevance theory to consider multimodal referring expressions, which has lead us to define a tentative scoring measure of such expressions to evaluate their relevance. Even if the results we presented on our experimental data seems promising, the calculus is still rather coarse and has to be considered within the context of a real modelling of the notion of relevance in multimodal dialogue.

Bibliography

Anne Abeillé, 1993. *Les nouvelles syntaxes. Grammaires d'unification et analyse du français* Armand Colin Editeur, Paris.

A. De Angeli, W. Gerbino, D. Petrelli and G. Cassano, 1998. *Visual display, pointing and natural communication: The power of multimodal interaction.* Proceedings of AVI'98, l'Aquila, Italy.

A. De Angeli, W. Gerbino , L. Romary and F. Wolff, 1999. *The ecological approach to multimodal system design.* Submitted to Proceedings of the 3rd Gesture Workshop, Lecture Notes in Artificial Intelligence, Springer Verlag.

J.J. Gibson, 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

A. Jönsson and N. Dählback .1988. *Talking to a computer is not like talking to your best friend.* Proceedings of the Scandinavian Conference on Artificial Intelligence, pp. 53-68.

G. Kanizsa, 1979. *Organization in vision.* New York, Praeger.

S. Kita. *How representational gestures help speaking.* In: McNeill, D. Language and gesture: Window into language and action, (in press).

U. Neisser, 1976. *Cognition and Reality.* San Francisco: Freeman & Co.

D. Norman, 1988. *The psychology of everyday things.* New York: Basic Books,

A. Reboul, 1998. *Reference, agreement, evolving reference and the theory of mental representations.* In Hommages à Liliane Tasmowski-De Ryjck, Unipress, Padoue.

I. Sag and J. Hankamer, 1984. *Toward a theory of anaphoric processing.*Linguistics and Philosophy 7.

Y. Schabes and Stuart Shieber, 1994. *An alternative conception of tree-adjoining derivation.* Computational Linguistics, vol. 20, pp 91-124.

D. Sperber and D. Wilson, 1986. *Relevance, Communication and Cognition.* Blackwell.

M. Wertheimer, 1922. *Untersuchungen zur Lehre von der Gestalt I.* Psychologische Forschung, 1: 47-58.

F. Wolff, A. De Angeli and L. Romary, 1998. *Acting on a visual world: the role of perception in multimodal HCI.* AAAI, Workshop on multimodal representation, Madison, 1998

F. Wolff, 1999. *Analyse contextuelle des gestes de désignation en dialogue Homme-Machine.* Phd Thesis of Henri Poincaré University, Nancy.