1

Relevance and Perceptual Constraints in Multimodal Referring Actions (Draft version)

Frédéric Landragin[†], Antonella De Angeli[⋆], Frédéric Wolff[†], Patrice Lopez[†] and Laurent Romary[†]

[†]LORIA, BP 239, 54506 Vandœuvre-lès-Nancy, France.

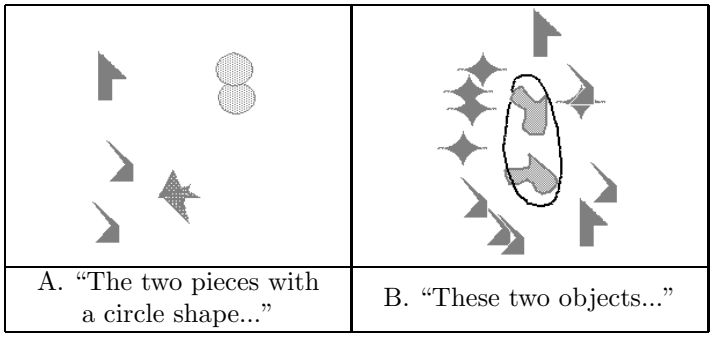
[⋆]NCR Knowledge Lab, 206 Marylebone Road, London, NW1 6LY, UK.

In this chapter we present 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 demonstrative gesture. Experimental results are 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 is limited to contextual effects and has to be refined, it seems to be already suited to validate our approach regarding the importance of group salience and access type in the choice of a referring mode.

1.1 Introduction

1.1.1 Referring actions in multimodal systems

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



A. "The two pieces with a circle shape..." | B. "These two objects..."

FIGURE 1 Examples of complex multimodal reference to a group of targets.

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 thereof (multimodal input). Moreover, each type of input can be exploited through a great flexibility of forms. As an example, consider the following multimodal inputs extracted from a corpus collected by a simulation experiment (Wolff 1999). Very different gestures and verbal utterances perform the same communication goal, i.e., referring to a group of targets (see Figure 1).

Despite their clear usability advantages, the design of multimodal systems still poses original and challenging problems to HCI researchers. In particular, it requires the development of innovative interaction paradigms to constrain 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 (Oviatt and Cohen 2000). 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 for intonation, disfluencies, kinaesthetic motor control, cross-modal integration and timing. Automaticity occurs over extensive practice with an activity, when specific routines are built up in memory. Performed beyond conscious awareness, automatic processing is effortless and fast, but also requires great effort to be modified. Even when people learn new solutions (i.e., set up alternative routines in 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 users to directly express their intentions. In this way, the effort required to monitor their own expressions, as well as to plan and perform corresponding actions remains minimal. Usable systems thus require a deep understanding of the factors affecting human spontaneous behaviour.

1.1.2 The ecological approach and its extension

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 architectures which are 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 1988), empirical research, especially in the form of early simulation, is instrumental to the formalisation of a theory of multimodal interaction (Oviatt 1997, Oviatt et al. 1997, De Angeli et al. 1998). Elsewhere (Wolff et al. 1998, De Angeli et al. 1999a, 1999b) we have 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 ecological approach claims that gesture variability is linked to visual perception. To reach this conclusion, the approach has revised gestural communication by introducing it into the perception-action cycle (Neisser 1976). This is 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. Through the analysis of spontaneous communication, we have demonstrated that the cyclic nature of cognition is a powerful conceptual structure for understanding referring gestures (Wolff et al. 1998, De Angeli et al. 1999b).

The ecological approach to multimodal system design assumes that gestures are virtual actions (Kita 2000), re-enactments of real activities in a virtual space. In particular, referring gestures are considered as virtual actions aimed at directing the listener's attention towards a target. These virtual actions do not modify the physical environment in which they are produced, as would do grasping the target and moving it in front of the listener. They modify the dialogue context, inducing

the listeners to shift the focus of their attention towards the target. This effect corresponds to the semiotic function of gesture.

In this chapter, we attempt to extend the ecological approach to cope also with verbal language variability. In particular, following the Relevance Theory (Sperber and Wilson 1995) we try to understand the link between speech in a discourse context and gesture in a perceptive context. The potential of this extension is confirmed by the results of a simulation study. Exploiting the perceptual constraints, people tend to modify the effect of their utterances.

1.2 Theoretical framework

1.2.1 Affordances and multimodal system design

The ecological approach is an established psychological theory to perception, cognition and action (Gibson 1979) now adapted to multimodal system design (De Angeli et al. 1999b). According to ecological psychology, the perception-action cycle is mediated by *affordances*, that is, 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 1979). 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 (1988), but its potential in the domain of natural communication is still not well understood.

The ecological approach to multimodal system design extends the concept of affordances to explain the variability of multimodal actions. Its basic assumption is that gestures are determined by the mutuality of information provided by the object to be referred to, and by 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 the variability of communicative actions.

1.2.2 Perceptual constraints in multimodal referring actions

Elsewhere, (De Angeli et al. 1998, De Angeli et al. 1999b, Wolff et al. 1998) we have demonstrated that the way a reference action is produced depends on the complexity of extracting the target from the visual context. Gestures are efficient means for coping with the complexity of the

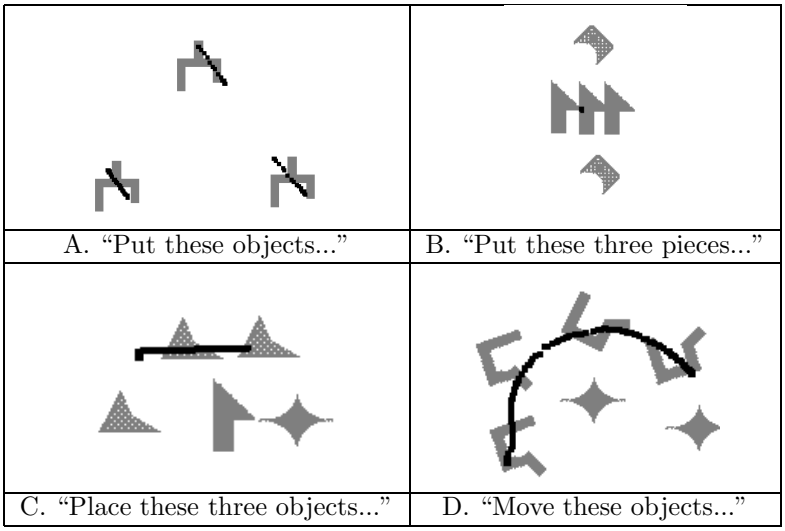


FIGURE 2 Examples of elective gesture for group designation.

visual word. They are deeply connected to visual perception.

In 2D environments where targets are on a same surface, a gesture can pass through the targets to be referred to, or it can define a borderline between the targets to be referred to and the other targets. The first case corresponds to an *elective* gesture (some examples are shown in Figure 2). The second case corresponds to a *separating* gesture (an example was shown in Figure 1-B). When targets are grouped, several gestures can indicate the targets one by one, or one global gesture can be used to indicate the entire group. The first case corresponds to an *individual access*, an example of which is shown in Figure 2-A. The second case corresponds to a *group access*, cf. Figure 2-B. Gestures can take several *forms* like pointing, scribbling, targeting, or circling. Gesture form is not linked to access type. No gestural category was associated with a particular type access. Single circlings, targetings and even pointings were used to refer to groups of objects. Especially in the cases of targeting and pointing, this feature leads to ambiguities (i.e., competition of possible candidates to referents identification). Resolving such ambiguities assumes to take into account implicit information on which depends the referring action, i.e., perceptual groups.

Type, granularity, form and size tend to be adapted to the visual context (size and layout of the objects). Visual cues are features which help solve ambiguities often arising from spontaneous gesturing. Considering

only the gesture and the visual context, Figure 2 shows:

- An example of type access ambiguity (does the small pointing in Figure 2-B indicate a single target or the group of three?).
- An example of scope ambiguity (how many targets are referred to in Figure 2-C: two, or three by including the nearest remaining target of the same shape?).
- And an example of pattern ambiguity (does the gesture in Figure 2-D draw an incomplete circle around the star-shaped targets, or pass through the four others to refer to them by a targeting form?).

An important factor for disambiguation is the visual *salience*. When an object or a group is salient, it is more susceptible to be referred to by a simple gesture. This is the case when it is isolated or when it has visual characteristics which distinguish it from the other objects and attract the user's intention more easily. In Figure 2-B for example, the group salience is very high and induces a group access. In a multimodal context, the verbal referring expression is another factor of disambiguation. The adjective "three" used in Figure 2-B and Figure 2-C is sufficient for the comprehension.

1.2.3 Relevance in multimodal referring actions

Another basic assumption of our model is that human beings are efficient communicators. According to the *cognitive* principle of relevance (Sperber and Wilson 1995), "human cognition tends to be geared to the maximisation of relevance." *Relevance* is a property of inputs to cognitive processes, embodying the notion of *contextual effect* and *processing effort*. Processing an input yields some contextual effects when it affects the context of previous assumptions on the topic. There are three types of effects: *(i)* contextual implication; *(ii)* strengthening or weakening of existing assumptions; *(iii)* contradiction and elimination of existing assumptions. The relation between effects and relevance is the following: the greater the effects, the greater the relevance. However, processing the input involves some mental effort. For the same effects: the smaller the effort needed to achieve them, the greater the relevance. Since the maximal effects can sometimes be achieved only at the price of an enormous effort, Sperber and Wilson introduce a distinction between maximal and optimal relevance. In a communicative situation and for an individual, optimal relevance corresponds to adequate effects for no unjustifiable effort. Assuming that human beings are efficient communicators supposes that they optimise the relevance of their communication acts. This brings us to the *communicative* principle of relevance: "Every act of ostensive communication communicates a presumption of its own

optimal relevance." (Sperber and Wilson 1995)

This principle can be transposed to HCI because of the spontaneous character of the interaction. It applies to multimodal communication because communication relies upon information which is distributed among gesture and language, the whole responding to the communication theory that is relevance. Applying Relevance Theory to multimodal referring actions, we expect that users select the most efficient referring strategy. In this reduced context, we need to specify the notions of relevance, contextual effects and processing effort. Effects correspond to everything that helps to resolve the referring action, i.e., to connect objects to words and gestures. Effects are obtained by two notions: the amount of new data deduced from the referring action and the importance of these data, with respect to the intention of reference. It seems possible to compute an evaluation of effects. Each word or gesture brings a piece of information which helps the resolution. As a consequence, we count them and we take into account the reference task context and the perceptual context by weighting. Processing effort could 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). Scoring the processing effort is a more complex problem which presumes to model cognitive processes. This point is discussed at the end of the chapter. Relevance can somehow be viewed as a measure of the ratio effects/effort. The most relevant referring hypothesis appears to be the one corresponding to the largest ratio value. Optimising the relevance implies maximising the ratio. The main problem in a multimodal context is the manner of weighting effects and effort, considering the particularities of each modality and their integration.

1.2.4 Objective

Applying the concept of affordances to multimodal communication, we expect that referring actions will be affected by the visual characteristics of the target. We assume that semantic features are distributed across language, vision and gesture to optimise communication relevance. All together, these modalities supply different and complementary information for composing meaning. The main complexity of multimodal reference is the heterogeneity of the referring expressions, exploiting advantages of both language and gesture. We mentioned earlier that in the context of a multimodal spoken interface, we need to consider the integration of gesture and language. Here, we wish to study the variability of referring expressions in light of Relevance Theory. Since this theory was developed in general terms, we argue that it can be applied

towards understanding heterogeneous multimodal referring expressions. Our objectives are therefore:

- To study whether our estimation of relevance agrees with the actual data. Can we evaluate the relevance of referring expressions in a multimodal context? Is it a discriminant factor?
- And to identify the factors and constraints which could help find the referring expression with the optimal relevance dynamically. Such factors would allow to determine and to optimize the reference resolution process in the context of intelligent multimodal interfaces.

Our approach is empirical and this study is more a description of prospective ongoing work than the development of a general model. This chapter addresses the evaluation of the effect of perception on multimodal referring actions, according to a descriptive scoring methodology.

1.3 Simulation study and hypothesis

1.3.1 Simulation study

A Wizard of Oz simulation was run to collect a corpus of spontaneous multimodal actions (Wolff 1999, Wolff et al. 1998, De Angeli et al. 1999b). In this technique, a human (the wizard) plays the role of the computer behind the interface in order to test the efficiency of the planned capacities of a dialog system before its implementation. Seven students from the University of Nancy participated in the simulation experiment as volunteers. They were French native speakers. Engaging a dialogue with the simulated system, they were required to move groups of objects into appropriate boxes. The interaction was based on speech and gesture, mediated by a microphone and an electronic pen. The experimental instructions provided to participants were only related to the task and not to the mode of interaction. In order to assure the spontaneous character of the interaction, users were free to use speech and gesture as they wished. This is important because our goal was to collect the largest possible variety of multimodal referring expressions for our corpus. To inhibit unimodal verbal references, the objects to be moved into the boxes were abstract-shaped figures, i.e., having no linguistic term associated with them (De Angeli et al. 1998). The shapes could be targets or distractors. The targets were collections of two or three identically shaped stimuli to be moved into the box displaying their shape. The distractors were exclusively used in relation to perceptual field organization and were not to be moved. The perceptual organization of the visual field was manipulated according to the principles of Gestalt Theory (Wertheimer 1922–1923, Kanizsa 1979) which describe the laws

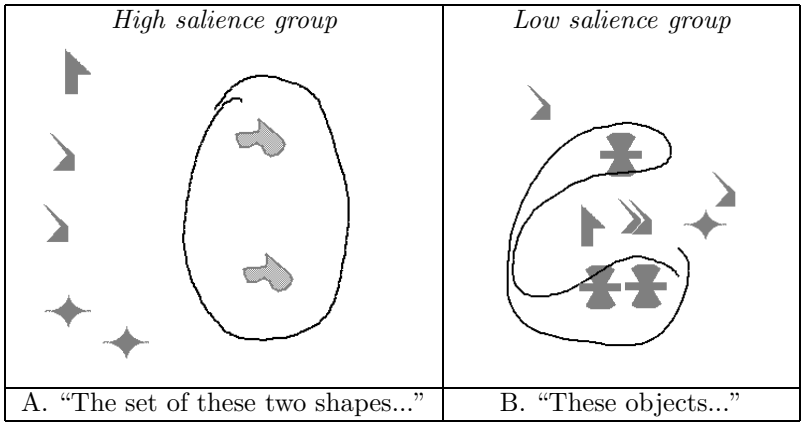


FIGURE 3 Examples of separating gesture for the designation of a high salience group or a low salience group.

underlying spontaneous grouping. The manipulation was based on *similarity* (the objects are grouped on the basis of their salient physical attributes, such as shape and color); *proximity* (the objects are grouped on the basis of their relative proximity); *good continuation* (the groups presenting continuous outlines are more salient than those with discontinuous ones).

The experiment contrasted high salience groups with low salience groups. In the first case, targets were easily perceived as an homogeneous group, clearly separated from surroundings called distractors (see Figure 3-A). In the latter case, targets were spontaneously perceived as elements of a broader heterogeneous group including distractors (see Figure 3-B). The distinction between high salience and low salience is extracted by evaluating the three gestalt grouping criteria and by comparing them. The proximity criterion is evaluated as true when *(i)* there is no distractor between the figures of the group and *(ii)* all distances between each of the figures of the group and each of the other figures are greater than the compactness value of the group, i.e., the maximum distance between two figures of the group. The circling gesture in Figure 3-A points out a high salience group. The two shapes are similar and, even if they are not so close, they are nearer to one another than to any other shapes or distractors. So the similarity and the proximity criteria are both evaluated as true. As we have only two targets, there is no need to evaluate good continuation and therefore the group salience is set to high. The gesture in Figure 3-B points out a low salience group of the three targets with the same shape. Proximity and good continuation,

both evaluated as false, are in opposition to similarity.

1.3.2 Hypothesis

Our main hypothesis is that group salience is an important predictive cue to access type. When group salience is very high, the gesture can be highly simplified. For example, a simple pointing gesture can refer to the entire group unambiguously, as seen in Figure 2-B. (In Figure 3-A, a simple pointing gesture is not conceivable but the circling gesture is also simplified due to the broad spacing. Rapidity and imprecision had no repercussion on the identification). On the contrary, low salient groups afford almost only individual access (if the user chooses group access anyway, the gesture will be very complex as shown in Figure 3-B, which is a singular example).

1.4 Data coding

1.4.1 Linguistic scoring

Our approach of Relevance Theory modelling applied to multimodal referencing supposes to compute both cognitive contextual effects and processing effort. A measurement of contextual effects seems particularly interesting because, when applied to automatic comprehension in multimodal dialogue systems, it could help to resolve the referring action. In order to study verbalisations occurring with gestural references, we have parsed our transcribed corpus with a Lexicalized Tree Adjoining Grammar (LTAG) using the tools described in (Lopez 2000) and computed a score for each referring expression. We used the definition of derivation of Schabes and Shieber (1994) and linguistics principles of Abeillé et al. (1999) for LTAG. These choices allow the result of the parsing (the derivation tree) to be equivalent to a classical semantic dependencies graph. Since each node in the derivation tree represents one and only one predicate, semantically empty words are not present in the derivation tree. Consequently the number of new pieces of data is obtained by the number of nodes in the dependency tree resulting from the parsing of the referring expression.

The importance of the predicates for the reference task is taken into account by weighting. This weighting w is subjective and depends on the application. For example, in our simulation we never have only one kind of object. Distractors are always heterogeneous shapes occurring at each visual scene step. So words such as *object* (contrary to words such as *circle*) provide no information for the identification of a piece in the visual scene and correspond to a score of $w = 0$. We consider that for the semantic head of the linguistic referring expression, we have $w = 0$ for abstract nouns in this application context. Considering the

simulation application, the list of transcribed abstract nouns is as follows: *object, form, piece* and *figure* (in French: "objet", "forme", "pièce" and "figure"). For the weighting of the other nouns and of the modifiers, we have introduced these general rules depending on the level of specification of the object:

- $w = 1$, for non-abstract nouns *(triangle, circle)*.
- $w = 1$, for adjectives *(grey)* as they provide a piece of information about an object to be referred to.
- $w = 1$, for conjunctions *(and)* as they provide a piece of information about how their elements are linked.
- $w = 1$, for prepositions *(with, without)* as they define the role of the objects in the global predicate.
- $w = 0$, for an indefinite determiner *(a)* which provides no information for the reference task.

The main problem concerns the other determiners (definite ones), the demonstrative articles and pronouns, and the deictic marks which in French can be concatenated at the end of a word ("cet objet-**ci** et cet objet-**là**", *this object and that object*). No rule can be defined for the weighting of these words if we do not consider the multimodal context (i.e., is there a gesture or not?). Consequently a score taking only the linguistic part into account cannot be defined, and the complete rules for scoring the linguistic part of multimodal referring expressions will be introduced in section 1.4.3.

1.4.2 Gestural scoring

The experimental conditions concern visual salience, and our hypothesis is based on this salience and on the access type of the gesture. Perceptual constraints are therefore taken into account in the gesture and we base our scoring on its access type.

In our corpus, the gestural part of each multimodal command is tabulated according to the strategy adopted to identify the corresponding group. Gestures are scored as *group accesses* when more than one object is accessed by only one gesture. In this case user's intention is to point out the group in which the referents are to be found. This intention and the corresponding amount of data constitute the effects. Gestures are scored as *individual accesses* when each gesture of the multimodal expression indicates only one object.

1.4.3 Multimodal scoring

Considering that a demonstrative article in the verbal part of a multimodal referring expression produces more effects than a definite one

(because a demonstrative holds a piece of information useful to link gesture and language), we can now introduce the following rules for the weighting of the determiners and articles in a multimodal context:

- $w = 0.5$, for definite determiners ("le", *the*) which indicate a more precised denomination of the referred object.
- $w = 1$, for demonstrative articles ("ce", *this*) 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.
- $w = 0.5$, for deictic marks which determine a link with a referring gesture.
- $w = 1$, for demonstrative pronouns ("celui-ci", *this one*; "celui-là", *that one*; "ceux-là", *those* or *those ones*) for similar reasons to the previous item.

Applying these rules, we get for example the following results: "les objets-là" = "ces objets" = "ceux-là" < "ces objets-là". For an expression which gives no information about the object to refer to like "un objet" *(an object)*, we have $w = 0$ and thus a null contextual effect. On the other hand, for a rich expression like "ce petit triangle" *(this small triangle)*, we have large contextual effects ($w = 3$). Now we have a complete set of rules to score the linguistic part of multimodal referring expressions. Two illustrative examples are shown in Figure 4.

1.5 Results

1.5.1 Corpus analysis

A corpus of 522 multimodal commands has been analysed. All the Figures in this chapter illustrate expressions from this corpus. To test the role of perception on the verbal utterance, we have extracted a sample of 98 significant multimodal referring expressions equally distributed between the two group conditions (De Angeli et al. 1997). In particular, we do not keep the unimodal referring expressions and duplicates. Consequently our sample is not statistically representative of the corpus, but this has no repercussion on our analysis for several reasons. First, our approach focuses on the spontaneous character of the communicative act and thus on its variability. Our goal is to show that relevance is a criterion to make sense of this variability. Consequently the sample only needs to be representative of the variability and we do not address the fact that someone used the odd expression more or less frequently. This approach justifies why we think we can make due with seven subjects. Considering the possible types of referring expressions, we were of

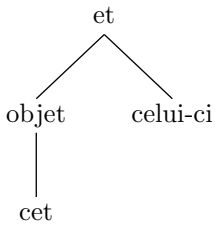
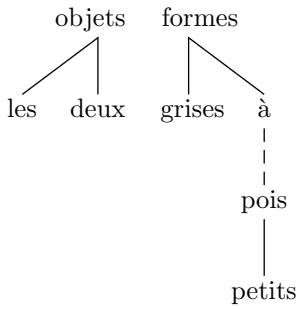
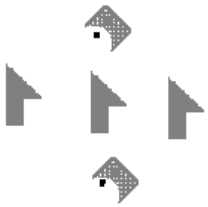
verbal referring expression	"cet objet et celui-ci" *(this object and this one)*	"les deux objets, formes grises à petits pois" *(the two objects, grey shapes with small points)*
derivation tree		
new data scoring	4	8
linguistic scoring	$1+0+1+1=$ **3**	$0.5+1+0$ $+0+1+1+1+1=$ **5.5**
gesture(s)		
gestural scoring	group access	individual access
	A. *Example of a coordination in the referring expression, with one long gesture (low group-salience).*	B. *Example of a precision in the referring expression, with two pointing gestures (low group-salience).*

FIGURE 4 Examples of multimodal scoring.

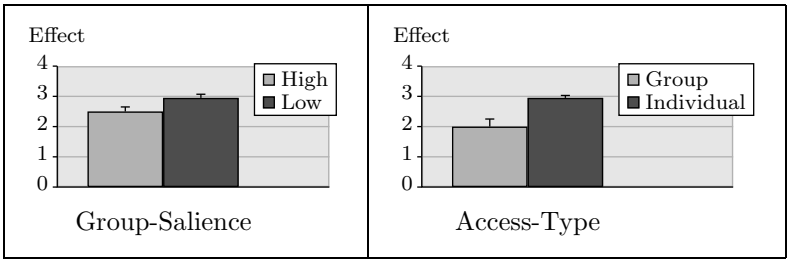


FIGURE 5 Contextual effects as a function of Group-Salience and Access-Type (Error bars represent mean standard errors).

the opinion that after seven simulation sessions a satisfactory variability was reached, i.e., increasing the size of the corpus would not lead to an increase in the number of different multimodal referring expressions. Moreover, we do not want statistical results on the use frequency of each type of referring expression. A person-oriented dialogue system must understand the most complex and significant expressions, and not only the most likely.

Our hypothesis concerns the group salience which was supposed to be a predictive cue for gesture access type. As a preliminary result which must still be verified, we found that perception has an effect not only on the gesture, but also on the verbalization. This effect is twofold: *implicit* cues and *explicit* cues may affect the verbal part of the referring expression. It is implicit when the verbal expression is affected by the visual layout of the object to be referred to (in high salience conditions, verbalization is simplified). It is explicit when the verbal expression is affected by the gesture. We can observe that a simple group gesture is accompanied by a simple verbalization, and a complex gesture requires a complex verbalization. Here we have tested the twofold effect by:

- Comparing the effects of verbal expressions produced under the two group salience conditions (implicit cues).
- Comparing the effects of verbal expressions produced together with both gestural access types (explicit cues).

The average value of the contextual effects in our sample is 2.66, with a standard deviation of 1.17. The distribution ranges from 1 to 6 and is affected by a substantial positive skewedness, which is difficult to normalise by transformation. Because of this statistical characteristic, the two effects were tested separately using non-parametric statistics (see Figure 5). This procedure allows us to process low-quality data, with respect to variables for which the parameters of the population are

unknown (hence, the name non-parametric). To evaluate our hypothesis, the Mann-Whitney U test (Blalock 1979) was applied. It allows testing differences between groups, verifying whether two sampled populations are from the same population. The observations from both groups are combined and ranked (in the case of ties an average rank is assigned). Then, the number of times a score from group 1 precedes a score from group 2 and vice versa are calculated. The logic behind the test is trivial. If two groups do not differ, then the first observation should come from group 1, the second from 2, the third from 1, the fourth from 2 and so on. Instead, if the two samples come from different populations, their rank orders should differ. The Mann-Whitney statistical value U is obtained by counting the number of times a value from the group with the smaller sample precedes a value from the group with the larger sample. The probability value (p) associated with the U statistic depends on the total number of observations (N). The p value is the basis for deciding if a statistical result like the one observed would occur by chance. Saying that a correlation is significant at the 0.05 level means that you can be 95% confident that the results are due to a real difference in the population and not to chance factors.

For the implicit cues hypothesis, the Mann-Whitney U test shows a strong influence of group salience on contextual effects, $U = 687$ ($N = 87$); $p < .05$. Under the low salience conditions, users produced complex verbalizations; whereas under high salience conditions, verbalizations 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 more easily perceptible, they need less complex referring expressions to convey their intentions, since they can exploit implicit information. As a corollary, we can deduce that humans naturally attribute their own perceptual capabilities to an artificial interlocutor.

The influence of explicit visual cues on contextual effects was straightforward, $U = 324,5$ ($N = 84$); $p < .001$. The analysis shows that, for multimodal reference, complexity of gestural and verbal expressions are linked. A simple referring gesture (direct group access) comes jointly with a simple verbalization, generally a deictic. A complex gesture (individual access that indirectly build the group) needs a complex verbalization. In the latter case, the linguistic part provides temporal links for simplifying the complex spatial designations. The results are consistent with previous findings (Wolff et al. 1998, De Angeli et al. 1999b) showing a high correlation between the reference strategies adopted by verbal and gestural languages to identify targets. In particular, gestural group access was always accompanied by a plural deictic anchor or target de-

scriptions (i.e., *these objects, the two isolated objects, the two forms*). Moreover, only one out of three multimodal expressions were formed by an individual access accompanied by a plural verbal reference.

For the implementation of dialogue systems, these results are important because they point out the significance of taking perceptual constraints into account. Even if the task of our study leads to referring actions on sets of similar objects (and this could have favoured the perception of salient groups), it appears that visual perception gives some hints to interpret gestures and verbal expressions. This could lead to more comprehensive systems.

1.5.2 Effects and effort and their use in multimodal systems

We argue first that evaluating contextual effects and processing effort can aid comprehension. Sperber and Wilson (1995) consider that an evaluation of effects and effort with counting operations is not representative of human abilities. Humans are not able to pass judgment on the quantity of obtained effects or effort. Nevertheless, if it is possible to provide such capacities to a dialogue system, the automatic comprehension process will be improved because:

1. It could identify heuristics to prune the referring search space. Given several candidates for a multimodal referring expression, a post-evaluation of their relevance can show preferential procedures and their significant parameters for the reference resolution.
2. It could question the result of a reference resolution when its relevance is low.
3. It could detect the ill-chosen utterances that reveal a particular behaviour of the user.
4. It could be helpful to go back in the dialogue history to find the source of a misunderstanding.

To sum up, managing scores is a way of representing complex processes in communication.

Secondly, in automatic generation, an evaluation of relevance appears to be the ideal criterion for choosing among the large range of available referring expressions. More precisely, it is a criterion to minimize the number of backtracks in the generation search space. The main problem in generation is the impossibility to generate all of the possible expressions before chosing one of them. Moreover, the possible referring expressions in a multimodal context explode, considering the great number of gesture forms and the great number of verbal expressions. Exploratory choices must be done with the possibility to look back and question these choices. By taking perceptual constraints into account

and using relevance evaluation, we argue that the most efficient choices will be done from the very beginning of the process.

1.6 Future work

To evaluate contextual effects, we have proposed scoring rules for the linguistic part of multimodal referring expressions. This is enough to point out the link between perception and verbalization, but not to present a real metric for multimodal input. With this intention we also need to extend the score to the contextual effects of gestures. As seen in section 1.2.3, effects are obtained by two notions: the amount of new data deduced from the referring action and the importance of these data, with respect to the intention of reference. We can evaluate the number of new pieces of data deduced from a gesture by enumerating each particularity that brings a piece of information helpful for the referent identification. Considering the possible types of gestures we collected in our corpus, it is the case of a stop in the gesture or an abrupt change of direction. The notion of *singularity* (Bellalem and Romary 1995) includes these particularities and provides a framework to analyse them (a semantic feature is associated with each singularity). Then we need to evaluate the importance of these data. A weighting could be introduced for this purpose taking into account the perceptual context (position of the targets, salience, group). Then the scores may be confronted and the correlation between perception, gesture and language may be tested.

The processing effort of a referring action is of course more complicated to compute than the contextual effects which only require a descriptive methodology. Evaluation of the processing effort requires a general theory of reference that takes into account multimodal communication. In order to correctly interpret this kind of natural referring actions, the system has to represent and uses jointly for each objects of the application different kind of information: linguistics, pragmatic and perceptive. To conclude considering the proposed contextual effects scoring, we argue that the Mental Representation Theory (Reboul 1998) could provide a framework able to compute the processing effort, to represent such information and to use it for reference resolution in accordance with Relevance Theory. The goal of this general theory of reference, inspired by Sag and Hankamer (1984), 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 utterance and allow to resolve heterogeneous references

as spatial designation. As seen in section 1.2.3, processing effort includes 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. Mental Representation Theory could lead to an evaluation of these notions. Inference has an important role in the operations over mental representations, and the accessibility of information is modelled by the notion of *domain of reference*. A domain of reference is a subset of the set of mental representations for a given individual at a given time (it is very much like the notion of context in Relevance Theory). Considering this notion, we plan to develop a general methodology for our scoring.

The implementation of dialogue systems integrating Mental Representations, scores of contextual effects and processing effort, is one of our main ongoing projects.

1.7 Conclusion

In this chapter, 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 demonstrative gestures could be seen as the realization of specific affordances induced by the gestalt properties of objects. From a more linguistic point of view, we have extended Relevance Theory to consider multimodal referring expressions, which has led us to define a tentative scoring measure of such expressions to evaluate their relevance. Although the results we presented seem 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.

References

Abeillé 1999. A. Abeillé, M.-H. Candito and A. Kinyon: FTAG: Current Status and Parsing Scheme. In *Proc. of Venezia per il Trattamento Automatico delle Lingue* (VEXTAL), Venice.

Bellalem and Romary 1995. N. Bellalem and L. Romary: Reference Interpretation in a Multimodal Environment Combining Speech and Gesture. In *Proc. of First International Workshop on Intelligence and Multimodality in Multimedia Interfaces*, Edinburgh.

Blalock 1979. H.M. Blalock: *Social Statistics*. McGraw-Hill, New York.

De Angeli et al. 1998. A. De Angeli, W. Gerbino, D. Petrelli and G. Cassano: Visual Display, Pointing and Natural Communication: The Power of Multimodal Interaction. In *Proc. of International Working Conference on Advanced Visual Interfaces* (AVI98), l'Aquila.

De Angeli et al. 1999a. A. De Angeli, W. Gerbino, L. Romary and F. Wolff:

The Ecological Approach to Multimodal System Design. In A. Braffort et al. (Eds.): *Gesture-Based Communication in Human-Computer Interaction, Lecture Notes in Artificial Intelligence 1739*, Springer-Verlag, Berlin.

De Angeli et al. 1999b. A. De Angeli, L. Romary and F. Wolff: Ecological Interfaces: Extending the Pointing Paradigm by Visual Context. In P. Bouquet et al. (Eds.): *Modeling and Using Context, Lecture Notes in Artificial Intelligence 1688*, Springer-Verlag, Berlin.

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

Jönsson and Dählback 1988. A. Jönsson and N. Dählback: Talking to a Computer is not like Talking to your Best Friend. In *Proc. of the Scandinavian Conference on Artificial Intelligence*, Tromsø.

Kanizsa 1979. G. Kanizsa: *Organization in Vision*. Praeger, New York.

Kita 2000. S. Kita: How Representational Gestures Help Speaking. In D. McNeill (Ed.): *Language and Gesture*, Cambridge University Press, New York.

Lopez 2000. P. Lopez: LTAG Workbench: A General Framework for LTAG. In *Proc. of 5th International Workshop on Tree Adjoining Grammars and Related Formalisms* (TAG+5), Paris.

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

Norman 1988. D. Norman: *The Psychology of Everyday Things*. Basic Books, New York.

Oviatt 1997. S. Oviatt: Multimodal Interactive Maps: Designing for Human Performance. Human-Computer Interaction 12, pp.93-129

Oviatt et al. 1997. S. Oviatt, A. De Angeli and K. Kuhn: Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proc. of Conference on Human Factors in Computing Systems* (CHI'97), ACM Press, New York. Also in E. André (Ed.): *Proc. of the ACL Workshop on Referring Phenomena in a Multimedia Context and their Computational Treatment* (ACL/EACL-97), Madrid.

Oviatt and Cohen 2000. S. Oviatt and P. Cohen: Multimodal Interfaces that Process what Comes Naturally. Communication of the ACM, vol.43, no.3, pp.45-53

Reboul 1998. A. Reboul: A Relevance Theoretic Approach to Reference. In *Proc. of Relevance Theory Workshop*, Luton.

Sag and Hankamer 1984. I. Sag and J. Hankamer: Toward a Theory of Anaphoric Processing. Linguistics and Philosophy 7, pp.325-345

Schabes and Shieber 1994. Y. Schabes and S. Shieber: An Alternative Conception of Tree-Adjoining Derivation. Computational Linguistics 20,

pp.91-124

Sperber and Wilson 1995. D. Sperber and D. Wilson: *Relevance. Communication and Cognition* (2nd edition). Blackwell, Oxford.

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

Wertheimer 1923. M. Wertheimer: Untersuchungen zur Lehre von der Gestalt II. Psychologische Forschung 4, pp.301-350

Wolff et al. 1998. F. Wolff, A. De Angeli and L. Romary: Acting on a Visual World: The Role of Perception in Multimodal HCI. In *Proc. of AAAI Workshop on Multimodal Representation*, Madison.

Wolff 1999. F. Wolff: *Analyse contextuelle des gestes de désignation en dialogue homme-machine*. Ph.D. Thesis of Henri Poincaré University, Nancy.