Interpretation of Multimodal Designation with Imprecise Gesture

Ali Choumane, Jacques Siroux

IRISA/Cordial, University of Rennes 1
6 rue de Kerampont, BP80518, 22305 Lannion, France
ali.choumane@irisa.fr, jacques.siroux@univ-rennes1.fr

Keywords: Multimodal human computer communication, reference, gesture, salience

Abstract

We are interested in multimodal systems that use the following modes and modalities: speech (and natural language) as input as well as output, gesture as input and visual as output using screen displays. The user exchanges with the system by gesture and/or oral statements in natural language. This exchange, encoded in the different modalities, carries the goal of the user and also the designation of objects (referents) needed to achieve this goal. The system must identify in a precise and non-ambiguous way the objects designated by the user. In this paper, our main concern is the multimodal designations, with possibly imprecise gesture, of objects in the visual context. In order to identify such a designation, we propose a solution which uses probabilities, knowledge about manipulated objects, and perceptive aspects (degree of salience) associated with these objects.

1 Introduction

We stand the present work[1] within the general framework of human-machine multimodal dialog systems [12]. The aim of such systems is to allow users to obtain the realization of services. For example, nowadays, multimodal systems are conceived to provide information on schedules for air flights, to elaborate on itineraries and to help produce models and plans.

The interaction between human users and systems to supply services requires to attain consensus about the user's goal. This consensus concerns a mutual comprehension of intentions which may happen (and has to be satisfied) during the interaction. It also concerns a shared line of-sight to all manipulated entities (e.g. parameters, objects, etc.) needed to accomplish the task.

The user designates these entities by acting on available modes and modalities: oral, language, gesture, etc. we call these designations "referential activities". An important role of a system is to recognize and understand these referential activities.

[1]This work is partially financed by the grant 2116B2-9/ARED 1800 of the regional council of Brittany, France.

This task is arduous because the system is confronted to many difficulties. The designation activity of the user is not reliable. Indeed, ambiguities, errors, and hesitations lead to "noise" or misunderstandings which are likely to be made worse by the hardware devices and the programs of the system. Finally, although multimodality is normally used to improve communication and to decrease the number of ambiguities, the joint use of many modes increases the number of technical problems and may degrade user's performance.

In this paper, we are interested in input to human-machine multimodal communication systems. A system, in order to understand the goal of the user, must correctly merge inputs which come in from different modes. A critical point of this merge is the resolution of referential expressions (REs). We propose an algorithm to resolve multimodal designations, with possibly imprecise gesture, to objects in the visual context. In addition, we consider cases of designations with or without information coming from the natural language modality (e.g. "I would like hotels here + gesture", "I would like a campsite along this river + gesture"). This algorithm includes two main strategies. On the one hand, it is based on the probability of designation of an object by a given gesture. On the other hand it is based on the salience of objects in the visual context. If the first strategy does not succeed in determining the referent, we think that the attention of the user could be influenced by the presentation of objects in the visual context. Thus we take into account the notion of salience in the reference resolution process in the second strategy.

After a section in which we more precisely present our context of work and the problem addressed, we explain the main elements of the proposed solution. We start by analyzing various possible cases to take into account, then we show the role of probability [2] and of salience [7] in RE resolution. Finally, we detail various steps in the algorithm for designations identification.

2 Main Problems

Our framework is the Georal tactile system [16] which has been implemented on the multiagent platform DORIS [9]. Georal is a multimodal system principally used to provide information of a touristic and geographical nature. Users can ask for information about the location of places of interest (beach, campsite, château, church, etc.) by specifying a place, a zone (particular geographical or

cartographical element: river, road, city, etc.) (figure 1); they can also request the distance and an itinerary between two localities.

Georal offers the user the following modes and modalities:

- Oral input as well as output to the system. Users can formulate their requests and responses to the system by voice and in natural language (NL) in a spontaneous manner (no particular instructions of elocution). Some system output is given in speech synthesis to the user.
- Visual mode: the system displays a map of a region on the screen; this map contains the usual geographical and touristic information: cities, roads, rivers, etc. Zooming effects, highlighting, and flashing allow the system to focus the user's attention.
- Gesture mode by the intermediary of a touch screen: the user can designate elements displayed on the screen by various types of gesture (point, zone, line, etc.).

Figure 1: A part of the map displayed on the Georal screen

A dialog with Georal consists of one or several exchanges. An exchange consists of communication turns of the user and the system [1]. The user's communication turn consists of an oral utterance and/or a gesture input. The system's communication turn consists of an oral output (by voice synthesis) and display on the screen. For example, a simple exchange contains two communication turns: a turn for the user (question) and a turn for the system (response). Notice that an exchange can contain others nested exchanges (e.g. in the case of clarification questions). In a user communication turn, the problem for the system is to resolve REs, i.e. find the referent of a symbol in one modality using information present either in the same or in other modalities.

Within this framework, we have proposed a new definition of a general model for RE resolution [3]. This model is based on two fundamental principles. On the one hand, it associates both a well defined language to each modality (NL, gesture, visual) and a mediator one. On the other hand, functions link objects from each modality to another, and allow reasoning and referents identification. The languages allow us to represent, for each communication turn, objects resulting from modalities. Moreover, the current context and the interaction histories are memorized using those languages. Treatments will be associated to each modality (e.g. anaphora treatment for NL). Specific treatments will be setup to determine the referents named on several modalities.

The purpose of this paper is to present a part of the processing associated to the model for RE resolution. The aim is to find an object designated in a multimodal manner with gesture, produced in a visual context (i.e. on the screen). This visual context is common to both the user and the system. The designated object represents the referent of RE. The proposed method relies upon three of the languages from the general model: the language which encodes the oral mode, the language which encodes the gesture mode, and the language which encodes the visual mode.

There are several types of REs: those which refer to entities in the NL history as the anaphora [11], those which do not have linguistic antecedents because there are employed in first mention [17], [10] and/or which refer to objects in another modality which correspond, for example, to the visual context in the Georal system. This last type of REs is produced:

- Jointly with a gesture. In this case, there are deictic REs in which the referent is the object designated by the gesture.
- Without gesture.

In this paper we are interested in REs produced jointly with gesture. The critical point in this processing is the imprecision of the gesture which may cause an ambiguous designation. We use the term "ambiguous gesture" for gestures that can have several possible interpretations. In addition, we take into account the problem of possible incoherence between modalities. This problem could arise when a gesture is accompanied with an oral utterance carrying information about the searched referent (the type of the referent for example). Note that this problem can also be caused by errors made by the oral recognition system when the type of referent that is searched for, for example, is not recognized correctly. We discuss problems of ambiguous gesture and incoherence between modalities as well as a resolution method.

3 Related Work

Previous work on multimodal reference resolution includes the use of linguistic approaches with spatial aspects and contextual factors [5, 15], the use of focus to disambiguate gestures to objects on a graphical display [18], etc. Pineda and Garza [13] propose a theory of representation and interpretation for multimodal messages, and a model for multimodal reference resolution. In this model, the notion of modality is captured in terms of a formal language and its interpreter.

A study by Landragin [6, 7] is based on visual salience. An object is salient when it attracts a user's visual attention more than others. This salience can be useful in input interpretation, for example, for multimodal reference resolution. Chai and al. [2] have proposed a graph-matching algorithm for reference resolution. Information gathered from multiple input modalities and the context is represented as attributed relational graphs. This approach identifies the most probable referents by optimizing the satisfaction of semantic, temporal, and contextual constraints on the gesture and oral. For example, in this system, a probability is assigned to each object that is likely to be selected by a gesture. This probability is a function of the distance from the gesture to the object and the radius of a circular region which is centered at the coordinates of the selected point gesture (in the case of point gestures). This system performs mutual disambiguation, where each modality helps to correct errors in the others. However Eisenstein and Christoudias [4] assert that this approach restricts users to a predefined grammar and lexicon, and relies heavily on having a complete formal ontology of the domain. More recently, Qu and al. [14] propose to use a notion of salience driven language models and gesture to improve the natural language understanding.

Our approach is inspired by both visual salience and designation probability. We propose a twofold strategy-based algorithm in which, on the one hand, we calculate the designation probabilities depending on the object and gesture types (cf. section 4.3), and on the other hand, we consider a wider concept of salience than visual salience which takes into account the context of the interaction (cf. section 4.4).

4 Analysis and Proposed Solution

We deal more precisely with the multimodal designation using gesture, towards an object in the common visual context (the screen in the Georal system). We propose a solution for inputs which can be represented by the regular expression "I would like X (here | by there | along this Y | on the left of this Y | ...)", accompanied by a gesture on the screen, where $X \in$ {campsites, hotels, etc. } and $Y \in$ {river, road, etc.}. We also take into account purely gesture inputs as responses from the user to the system question for example (cf. section 4.1).

4.1 Various possible cases

There are several possibilities for referencing illustrated by examples 1, 2, and 3 (U: User, S: System).

Example 1
U: *I would like campsites here* + designation gesture on the screen.

Example 2
U: *I would like campsites along this river* + designation gesture on the screen.

Example 3
U_1: *I would like campsites*.
S: *there are several localities which answer your request* + displaying the names of these localities on the screen. *In which place should I search?*
U_2: a gesture to one locality among those displayed.

In example 1, the oral input accompanied by a gesture (complementarity use) in which the referent of *here* is the object designated by the gesture. We detect and resolve a possible ambiguity in the gesture by following the algorithm proposed below (cf section 4.5).

Example 2 contains the RE *along this river* produced with a gesture. This RE contains information about the referent that should be taken into account in the resolution process. To do so, we consider three possible cases:

1. There is not a river in the visual context and the gesture designates an object whose type is different to river (there is no resolution). In this case the response to the user is a dialog decision. An information message and a clarification question are sufficient.

2. There is not an incoherence problem between modalities: the gesture designates several objects that are of the type river. In this case, we have a gesture with an intra-type ambiguity problem. There is ambiguity between objects of the same type (river type in this case).

3. There is incoherence between modalities: gesture with an inter-type ambiguity problem. There is ambiguity between objects of different types. In example 2, this designation ambiguity could be between a river and a city, etc. We resolve this incoherence problem by a filter based on object types (cf. section 4.3) and then the problem become that described in the previous case (intra-type ambiguity). In example 2, we select only objects of river types.

In example 3, U_2 is an input which contains only a gesture as a response from the user to the system. This is a particular case of example 1 in which the possible problem of ambiguity of gesture could be between objects proposed by the system (details follow).

It can be observed, that the previous three cases lead to two categories of multimodal inputs:

1. We don't have (or have little) information, which comes from the NL modality (examples 1 and 3), required to determine the referent. The resolution consists of determining the object designated by the gesture. The critical point in which the system is confronted is the ambiguity of gesture.

2. We have information, which comes from the NL modality, required to determine the referent (as the type of referent in example 2). However, another problem is added to the possible problem of gesture ambiguity, that of incoherence between modalities.

4.2 Notations

We use the following notations when formalizing our solution:
CVC_c is the current common visual context between the user and the system.
e is a given exchange between the user and the system.
t is a given communication turn. We recall that t of the user usually consists of an oral utterance and/or a gesture input.
R = {r_k, $1 \leq k \leq K$ / ER(r_k)}, the RE(s) produced by the user in the communication turn t.
T = {g_i, $1 \leq$ i $\leq$ I}, the gesture(s) carried out by the user in the communication turn t.
O_i = {o_j, $1 \leq j \leq J$}, the set of candidate objects referents designated by the gesture g_i.
$S_c(o_j)$ is the salience (cf. section 4.4) of the object o_j in the current visual context c (CVC_c).

We assume that in a given turn t of a user, $|R| \in \{0, 1\}$ and $|T| = 1$ (K=1). i.e. t contains only zero or one RE and only one gesture which designates one or several objects in the CVC_c. This restriction is to avoid oral/gesture alignment problems. If |T| > 1 and/or |R| > 1, temporal addressing is added to align oral and gesture [4]. We don't take into account this case in this paper. Consequently:

- g_1 is the only gesture in communication turn t. In example 1, we dispose of one gesture g_1 which accompanies one RE r_1 which is the word "here". The problem of resolving the RE r_1 is then reduced to determine objects designated by the gesture g_1.

- O_1 contains candidate objects referents designated by the gesture g_1 (and thus by r_1 if |R|=1).

4.3 Determining the Set of Candidate Referents

The set of candidate objects referents designated by a gesture g_1, noted O_1, contains all objects "probably indicated" by this gesture. i.e., an object o_j will be selected as a candidate in O_1 if $p(o_j/g_1) \geq \epsilon_1$ and the type of o_j is equal to Y (if Y exist). $p(o_j/g_1)$ is the conditional probability of o_j given the gesture g_1, $\sum_{j=1}^{J} p(o_j/g_1) = 1$, and $\epsilon_1 > 0$. We choose only objects in which the probability of designation by g_1 is equal to or higher than a threshold ϵ_1 (non null). This avoids the inclusion of objects "not concerned" with the gesture in O_1.

The probability $p(o_j/g_1)$ is formulated by:

$$p(o_j/g_1) = \frac{p(g_1/o_j)p(o_j)}{p(g_1)}$$

in which the only factor to compute is $p(g_1/o_j)$. We consider three types of gestures for this probability:

- In the case of line or point gestures, $p(g_1/o_j)$ is function of the distance from the gesture g_1 to the object o_j, where

$$p(g_1/o_j) = \frac{e^{-\bar{d}(g_1,o_j)}}{\sum_{m=1}^{J} e^{-\bar{d}(g_1,o_m)}}$$

$$and\ \ \bar{d}(g_1, o_j) = \frac{1}{N} \sum_{n=1}^{N} \min_{m \in [1,M]} d(x_n, y_m)$$

$$\forall x_n \in g_1\ and\ y_n \in o_j$$

- In the case of zone gestures, we propose another way to calculate $p(g_1/o_j)$. In the framework of the Georal system, several cases are taken into account (these cases are related to the types of displayed objects on the screen, cf. section 2):

  – For point objects, such as cities, which belong to a zone gesture, the probability $p(g_1/o_j)$ = 1.
  – For objects which occupy a surface on the map, such as forests, the probability is:

$$p(g_1/o_j) = \frac{\frac{area(g_1 \bigcap o_j)}{area(o_j)}}{\sum_{m=1}^{J} p(g_1/o_m)}$$

  – For objects of polyline type, such as roads and rivers, the probability is:

$$p(g_1/o_j) = \frac{\frac{length(g_1 \bigcap o_j)}{length(o_j)}}{\sum_{m=1}^{J} p(g_1/o_m)}$$

The determination of set O_1 is part of the referent search algorithm detailed below (cf. section 4.5).

4.4 Salience and its Purpose

Salience interferes strongly during the interpretation of an utterance in a dialog situation or during the comprehension of a text: by pointing at an element, salience draws attention to this element and makes its taking into account prior in the reference and coreference resolution process [6]. We find in the literature two types of salience: linguistic and visual salience.

Linguistic salience, which depends only on the NL modality, provides, for example, help in anaphora resolution [8]. In every human-machine communication using the visual mode, visual salience constitutes an identification criterion

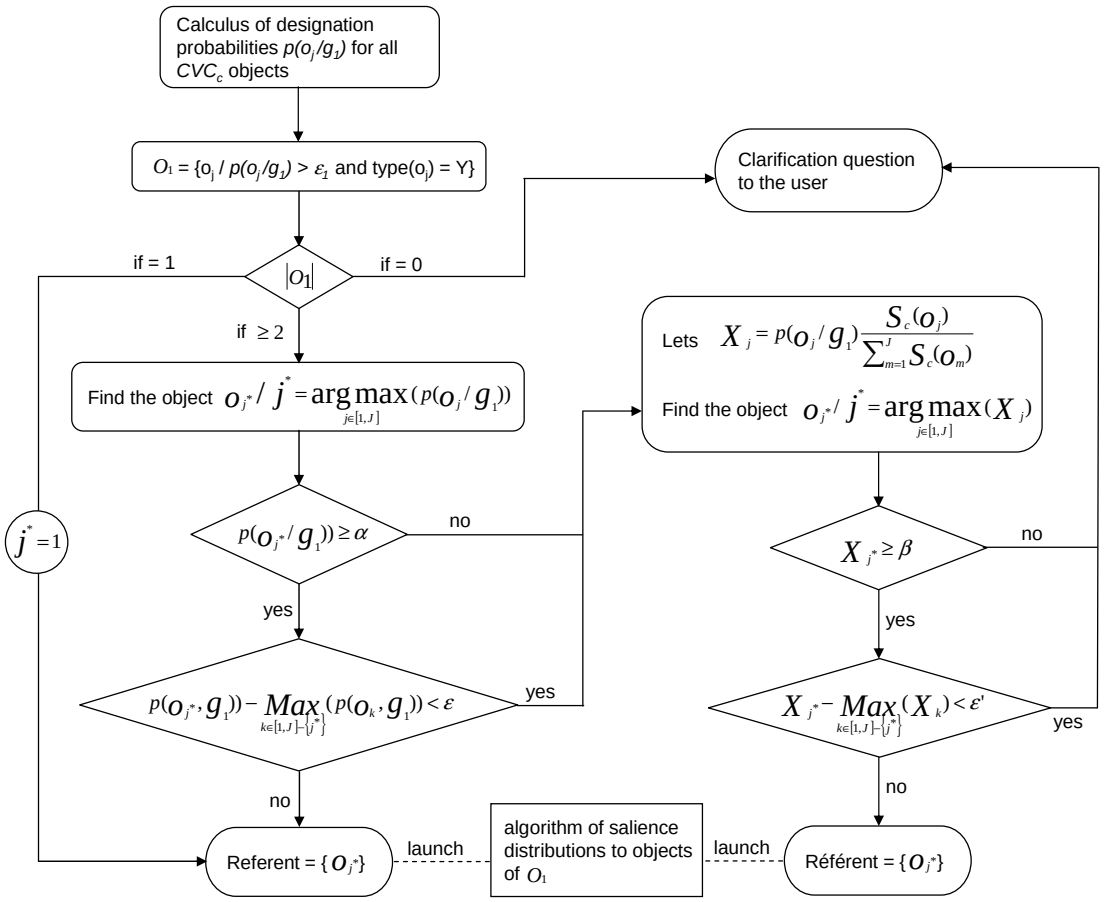

Figure 2: Algorithm for referents search designated by a gesture in the visual context

for the object designated and caught in a priority way [6].

Our approach is based on what we call "contextual salience". We thus aim at wider concept then that of the visual salience. The contextual salience of an object will be changed during interaction depending on if the object is designated by the user. We also take into account visual characteristics of objects which might capture the user's attention. Hereafter, we refer to contextual salience by only the word "salience". We will go on to show, how this salience allows the resolution of ambiguous gestures if the probability-based method fails.

We distinguish between two times of use of salience:

- At the beginning of the dialog we provide default values to CVC objects. These default values are related to the application. The determination of a numerical computing method of the salience [6] is beyond the scope of this paper. Let us note simply the existence of several factors which contribute to making an object salient and interfering in the quantification of the salience of this object. These factors include among others the color, size, and complexity of an object. This initialization step will be taken at the beginning of each dialog (at least an exchange).
- During the interaction, we modify the salience of objects in O_1 at the end of each interpretation of user input. Thus the salience of the referent(s) increase and the salience of the other objects in O_1 decrease. We recall that the set O_1 in a user's communication turn contains the candidate objects referents of g_1. Lets $S_c(o_j)$ be the salience of the object o_j in the CVC_c. At the end of the interpretation of user's input, and after using the saliences of CVC_c, we modify saliences of the object(s) in O_1. The interpretation by the system of every user input will take into account all contextual information (visual, linguistics, etc.).

Here is the simplified algorithm of salience distributions (a and b are two constants to adjust, with $a > 0$ and $b \leq 0$):

if beginning of dialog **then**
  for all $o_j \in CVC_c$ **do**
    $S_c(o_j) \leftarrow S_0(o_j)$ (saliences initialization)
  end for
else {this is the end of the interpretation of the user's input and we dispose of set O_1. We will modify the saliences of the objects in O_1 for the next CVC_c}
  for all $o_j \in O_1$ **do**
    if o_j is a referent **then**
      $S_c(o_j) \leftarrow S_c(o_j) + a$
    else {i.e. this is an object to penalize}
      $S_c(o_j) \leftarrow S_c(o_j) + b$

end if
end for
end if

This algorithm is called by the referent(s) search algorithm (cf. section 4.5). If the salience distributions algorithm have set O_1 in the input, then the saliences of these objects o_j will be modified by the values a and b depending on if the object is a referent or not. The constants a and b will be adjusted by later experiments to carry out.

4.5 Referent Search Algorithm

The algorithm we propose (flowchart in figure 2) consists of several steps:

- The first one is the step of calculating the designation probabilities $p(o_j/g_1)$ of objects in CVC_c, given a gesture g_1. These probabilities are calculated depending on the gesture type (point, line, zone, etc.) and object type in CVC_c (cf. section 4.3).

- The second step consists of determining set O_1 of the candidate objects referents. An object o_j is selected as candidate in O_1 if:

  1. Its probability of designation by g_1 is "relatively high" (i.e. $p(o_j, g_1) > \epsilon_1$), depending on the choice of ϵ_1.
  2. And its type is equal to Y (type(o_j)=Y), if Y exists.

- Then, we test the number of element in O_1. Three cases are taken into account:

  1. If $|O_1| = 0$, then the information message will be shown and the clarification question will be asked.
  2. If $|O_1| = 1$, then the referent of g_1 is the object represented by o_1.
  3. If $|O_1| \geq 2$, two strategies are setup:

     (a) The first strategy consists of selecting the object in O_1 which has the highest probability. This is the object o_{j^*} such as:

     $$j^* = \arg\max_j p(o_j/g_1)$$

     with

     $$p(o_j/g_1) \geq \alpha \qquad (1)$$

     and

     $$(p(o_{j^*}/g_1) - \max_{k \in [1,J]-\{j^*\}} p(o_k/g_1)) \geq \epsilon_2 \quad (2)$$

     α is the confidence threshold. It is required to avoid the choice of an improbable object. We intend to refine the calculus of α in later experiments.
     ϵ_2 is a real number high enough to say that g_1 designates only the object with the highest probability o_{j^*}. This is to detect ambiguous cases.

     If the above conditions (1) and (2) are satisfied, then the referent designated by g_1 is found (it is the object o_{j^*}). We modify the saliences of the objects in O_1 by calling the salience distribution algorithm shown above (cf. section 4.4).

     (b) The second strategy is applied when condition (1) or (2) is not satisfied. Thus, we think that the attention of the user could be influenced by object presentation in the visual context. We normalize probabilities $p(o_j/g_1)$ by X_j, with:

     $$X_j = \frac{S_c(o_j)}{\sum_{m=1}^{J} S_c(o_m)}$$

     and we search for the object o_{j^*} such as:

     $$j^* = \arg\max_j X_j$$

     with

     $$X_j \geq \beta \qquad (3)$$

     and

     $$(X_j - \max_{k \in [1,J]-\{j^*\}} X_k) \geq \epsilon_3 \qquad (4)$$

     If the above conditions (3) and (4) are satisfied, then the designated referent by g_1 is found (it is the object o_{j^*}). We modify the saliences of the objects in O_1 (idem the first strategy).

  If condition (3) or (4) is not satisfied, then the system produces a question to ask to the user to choose one of the objects in O_1. If such objects are displayed in a special graphical manner (zoom, flushing, etc) their saliences will be modified. Notice that in this case, there will be nested exchange.

The salience of objects in CVC_c will be reinitialized starting with the first exchange in the next dialog.

5 Conclusion

We have proposed a solution that addresses the multimodal designation of objects in a common visual context between the user and the system. More precisely, we have dealt with both cases of ambiguous gestures and the incoherence problem between modalities. This solution is based on probabilities, knowledge about manipulated objects, and perceptive aspects (degree of salience) associated with these objects. It implements an algorithm with two main strategies: it is based, on the one hand, on the designation probability of an object by a given gesture, on the other hand, on the salience of objects in the visual context. If the first strategy fails to determine the referent (because the probability of designation is less than the confidence threshold or because we have detected an ambiguous designation),

we think that the attention of the user could be influenced by object presentation in the visual context. Thus we take into account the salience concept in the reference resolution process in the second strategy. The completion of the resolution process causes, whatever the result, the modification of the saliences for the next communication turn. The proposed algorithm is under development. We intend to lead experiments evaluating it and refining some parameters.

References

[1] E. Bilange. *Dialogue personne-machine. Modélisation et réalisation informatique*. 2-86601-324-7. Hermes, 1992.

[2] J. Y. Chai, P. Hong, and M. X. Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th international conference on Intelligent user interface*, pages 70–77, New York, NY, USA, 2004. ACM Press.

[3] A. Choumane and J. Siroux. Toward a generic model including knowledge and treatments for multimodal reference resolution. In Vicente P. Guerrero-Bote, editor, *Proceedings Inscit2006*, volume 2, pages 298 – 302, Mérida - Spain, October, 25th-28th 2006.

[4] J. Eisenstein and C. M. Christoudias. A salience-based approach to gesture-speech alignment. In *HLT-NAACL*, pages 25–32, 2004.

[5] D. Hernández. *Qualitative Representation of Spatial Knowledge*. Springer Berlin / Heidelberg, 1994.

[6] F. Landragin. Traitement automatique de la saillance. In *Douzième conférence sur le traitement automatique des langues*, pages 263 – 272, 2005.

[7] F. Landragin, N. Bellalem, and L. Romary. Visual salience and perceptual grouping in multimodal interactivity. In *First International Workshop on Information Presentation and Natural Multimodal Dialogue*, pages 151–155, Verona, Italy, 2001.

[8] S. Lappin and H. J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535 – 561, 1994.

[9] J. L'Hour, O. Boëffard, J. Siroux, L. Miclet, F. Charpentier, and T. Moudenc. Doris, a multiagent/ip platform for multimodal dialogue applications. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 3049–3052, Jeju Island, Korea, 2004.

[10] H. Manuélian. *Descriptions définies et démonstratives: analyses de corpus pour la génération de textes*. PhD thesis, Université Nancy 2, novembre 2003.

[11] R. Mitkov. *Anaphora Resolution*. 0-582-32505-6. Pearson Education, 2002.

[12] S. Oviatt. Ten myths of multimodal interaction. *Commun. ACM*, 42(11):74–81, 1999.

[13] L. Pineda and G. Garza. A model for multimodel reference resolution. *Computational Linguistics*, 26 (2):139–193, 2000.

[14] S. Qu and J. Y. Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *ICMI '06: Proceedings of the 8th International Conference on Multimodal Interfaces*, pages 193–200, New York, NY, USA, 2006. ACM Press.

[15] D. Schang. *Représentation et interprétation de connaissances spatiales dans un système de dialogue homme-machine*. PhD thesis, Université Henri Poincaré, Nancy 1, 1997.

[16] J. Siroux, M. Guyomard, F. Multon, and C. Rémondeau. Multimodal references in georal tactile. In *Workshop Referring Phenomena In a multimedia Context And Their Computational Treatment, 35th Meeting Of the ACL*, Madrid - Spain, 1997.

[17] R. Vieira and M. Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26 (4):539–593, 2000.

[18] W. Wahlster. An intelligent multimodal interface. In North-Holland Publishers, editor, *In Z.W Raz and L.Saitta (eds.), Methodologies for Intelligent Systems*, New York, 1988.