Referring to Objects Through Sub-Contexts in Multimodal Human-Computer Interaction

Frédéric Landragin and **Laurent Romary**
LORIA, campus scientifique – BP 239
F-54506 Vandœuvre-lès-Nancy CEDEX – FRANCE
{Frederic.Landragin,Laurent.Romary@loria.fr}

Abstract

There is no one-to-one relation between referential terms and types of access to the referents (referring modes) in multimodal human-computer interaction. We propose a classification of referring modes implying sub-contexts. We describe in detail the nature of these contextual subsets and we show their importance for each type of referring action. Then we define a relation between terms and modes, and we deduce a list of disambiguation principles for the computation of referential terms in order to identify the correct referring mode, the correct sub-context, and the correct referent.

1 Introduction

Many linguistic, pragmatic, and philosophical works deal with referring phenomena. For example, see (Karmiloff-Smith, 1979) or (Corblin, 1987) for a systematic linguistic approach to French referring phenomena, (Sperber and Wilson, 1995) or (Reboul and Moeschler, 1998) for a pragmatic approach, and (Récanati, 1993) for a philosophical work. Two points of interest are the linguistic form of the referring expression (the 'referential term') and the interpretation process (the 'referring mode'). Proper names, definite descriptions like "the red triangle," indexicals like "that," complex demonstratives like "this red triangle," are examples of referential terms. Direct and indirect access to the referent, specific and generic interpretation ("a triangle has three sides") are examples of referring modes.

From a computational point of view, no definitive link can be established between the referential term and the referring mode. The same referential term can be interpreted in several ways, the ambiguity lying in the multiple possible choices of referring mode (Reboul and Moeschler, 1998). In order to resolve the reference, a dialogue system has to be aware of the possible ambiguities (Beun and Cremers, 1998) and has to choose the relevant hypothesis.

In this paper we focus on the nature of these ambiguities, taking as a guide the notion of 'reference domain.' A reference domain is a structured sub-context in which the reference occurs. For example, "this red triangle" associated to a pointing gesture refers to a particular triangle in a domain including several red triangles. The contrast conveyed by the demonstrative determiner is then justified by the presence in the domain of at least another, not-focused, red triangle. The notion of reference domain is useful, first to exploit all the components of the referential term (determiner, category, modifiers, spatial information), second to take implicit attentional phenomena into account. For example, if the user refers to "these triangles" when pointing out two triangles in a salient group of three similar triangles (see the Figure 1), reference domains may be useful to identify the ambiguity between referring to two or three triangles. As the visual scene includes two strong perceptual groups, one on the left and one on the right, the contrast conveyed by the demonstrative can apply between these two groups, or between the pointed triangles and the third one. The reference domain corresponds to the whole visual scene in the first case, and to the left group in the second case. Then, if the user refers to "the two circles" without any pointing gesture, the reference domain corresponding to the left group will be useful to understand the referring intention, that is "the two circles in the same attentional space."

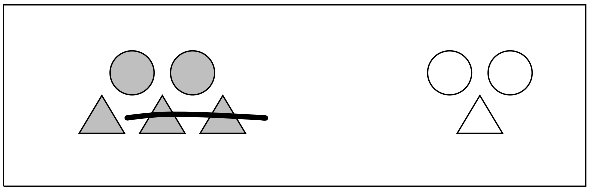

Figure 1: The use of sub-contexts.

As opposed to approaches like the one of (Kehler, 2000), we consider that simple algorithms are not sufficient for a multimodal system to identify the referents. As we see with our example, the gesture does not always give the referents, and the components of the verbal expression are not sufficient to distinguish them. But the combination of these ostensive clues with infered contextual considerations does. We focus here on the access to the referents through reference domains. Using a multimodal corpus (Wolff et al., 1998), we explore the possibilities of referring modes through reference domains. Based on this empirical data, we describe the foundations for a computational model of reference resolution, ie., a model of identification of the correct mode, domain, and referent. Since we used the corpus to collect referring phenomena and to build our model, we cannot validate it using the same corpus. Thus, we present a multimodal dialogue system which is currently under construction, and we show how such a system can validate our approach of referring.

2 Referential Terms and Referring Modes

To a referential term corresponds a preferential use. An indefinite noun phrase is generally used to introduce a new referent; a definite is an indicator to the necessity of identifying a particular referent (Corblin, 1987). A pure demonstrative is generally used together with a pointing gesture. But all of these referential terms have other uses. Reboul in (Moeschler and Reboul, 1994) proposes the following referring modes: direct reference, indirect reference, demonstrative reference, deictic reference and anaphorical reference. Proper names constitute the preferential direct referring mode, and demonstratives the preferential demonstrative referring mode. The problem is that a same noun phrase can be used for several referring modes. For example, the demonstrative noun phrase "this object" can be used for a demonstrative reference that implies an ostensive gesture, or for an anaphorical reference, the antecedent being a previous noun phrase like "the blue triangle." Indeed, demonstratives just as definites can be used for anaphorical purposes.

Thus, there is no one-to-one relation between referential terms and uses. To interpret them, we need to consider the context, which includes perceptual information, previous referring actions (and their results), and various world knowledge. With this linguistic, gestural, and visual information, the context is heterogeneous. Moreover, its scope can be enormous.

To face these problems, we have to consider sub-contexts. Our hypothesis is that each referring action occurs in a reference domain. This contextual subset is generally implicit and has to be identified by the system. It corresponds to a model of user's attention, user's memory, and conversation's area. The utterance's components allow to extract the referents from this subset and to prepare the interpretation of a future reference. For example, the referential term "the red triangle" includes two properties that must be discriminative in a reference domain that must include one or more "not-red triangles." A further referential term like "the other triangles" may be interpreted in the same domain, denoting a continuity in the reference sequence.

Reference domains can come from visual perception, language or gesture, or can be linked to the dialogue history or the task's constraints. Visual domains may come from perceptual grouping, for example to model focus spaces (see (Beun and Cremers, 1998)). When a particular colour is

<table>
<tr><th>Mode</th><th colspan="2">Mechanism details including the nature–linguistics or multimodal–of the referential expression</th><th>Examples with singular, quantifier, plural, numeral adjective</th></tr>
<tr><td>new-ref</td><td colspan="2">The referential expression does not refer but can be the antecedent of a future anaphor. No coreferent pointing gesture.</td><td>“Create a square,” “some squares,” “squares,” “two squares” (all possibilities).</td></tr>
<tr><td rowspan="3">ext-any-ref</td><td rowspan="3">The activated reference domain must be more reduced than the whole ontological class of objects.

It can be:</td><td>delimited by a coreferent pointing gesture</td><td>“Delete a square,” “some squares,” “squares,” “two squares” with a gesture delimiting a set of squares (all possibilities).</td></tr>
<tr><td>delimited by a previous referential term</td><td>“Select the squares and the triangles” followed by “delete a square,” “some squares” (interpreted as “delete some of the selected squares”), “squares,” “two squares” (all possibilities).</td></tr>
<tr><td>unprecised (in which case we consider the whole visual context)</td><td>“Delete two squares” interpreted as “delete two of the visible squares,” and eventually as “delete two of the visually salient squares” (all possibilities).</td></tr>
<tr><td>ind-par-ref</td><td colspan="2">It is the pointing gesture that forces the choice of the referent. This case constitutes a deviance from classical theories like the one of (Corblin, 1987). Nevertheless, we found it in the corpus of (Wolff et al., 1998).</td><td>“Delete a square” with a gesture pointing out a particular square, “some squares,” “squares,” “two squares” with gestures pointing out particular squares (all possibilities).</td></tr>
<tr><td>gen-ref</td><td colspan="2">Reference to a class of objects. No coreferent pointing gesture.</td><td>“A square has four sides,” “squares have four sides,” “two triangles with a common side make a quadrilateral” (quantifiers are impossible).</td></tr>
</table>

Figure 2: Indefinite noun phrases.

focused (ie., activated in the short-term memory, for example after “the red triangle” and “the red circle”), a reference domain based on the similarity of objects considering their colour is built on. Some domains may come from the user’s gesture (see (Landragin, 2002)), others from the task’s constraints. All of them are structured in the same way. They include a grouping factor (‘being in the same referring expression,’ ‘being in the same perceptual group’), and one or more partitions of elements. A partition gives information about possible decompositions of the domain (see (Salmon-Alt, 2001)). Each partition is characterized by a differentiation criterion, which represents a particular point of view on the domain and therefore predicts a particular referential access to its elements (‘red’ compared to ‘not-red,’ ‘focused’ compared to ‘not-focused’).

This unified framework allows to confront the various contextual information, and to model the implicit whatever its origin between perception, speech and gesture. Considering reference domains, a referring action can:

1. **‘new-ref’ mode:** introduce a new referent in a linguistic manner;
2. **‘ext-any-ref’ mode:** extract any element from an activated reference domain;
3. **‘ext-par-ref’ mode:** extract a particular element from an activated reference domain;
4. **‘ind-par-ref’ mode:** indicate a particular referent that is focused elsewhere;
5. **‘ind-par-dom’ mode:** indicate a particular reference domain whose one element is focused;
6. **‘gen-ref’ mode:** refer to a generic entity, that is not a set of particular referents nor a reference domain.

In this list, the introduction of a new referent by a multimodal referring action is seen as the linguistic mention of a referent that is focused by the coreferent ostensive gesture (‘ind-par-ref’ mode). One important point is that we consider that referring directly to a particular object is impossible without an activated domain. Consequently, the direct reference mode of Reboul corresponds here to ‘ext-par-ref’ mode. Mentional expressions (see (Corblin, 1999)) with “first,” “second” or “last”

Mode	Mechanism details		**Examples** with singular, plural, numeral adjective
ext-par-ref	The activated domain can be:	delimited by a coreferent pointing gesture	"**The triangle**" with a gesture delimiting a group of geometrical forms including one triangle (all possibilities).
		delimited by a previous referential term	"Select the blue triangle and the green square" followed by "delete **the triangle**," "select the triangles" followed by "**the two red triangles**," "the red triangle, the green one and the blue one" followed by "**the first**," "the group" followed by "**the triangle**" (all possibilities).
		delimited by a previous focussing on a visual space	After some references to objects at the left of the visual scene, "**the triangle**" can refer to "the triangle on the left" (all possibilities).
		delimited by a precision in the referential term	"**The triangle on the left of the scene**" (all possibilities).
		unprecised (in which case we consider a salient focus space)	"**The triangle**" interpreted as "the salient triangle" (all possibilities).
ind-par-ref	The referent can be:	given by a coreferent pointing gesture	"**The triangles**" with a gesture pointing a group of triangles (all possibilities)
		given by a previous referential term	"Select a red triangle" followed by "**the triangle**," "the triangle and the square" followed by "**the two forms**" (all possibilities).
ind-par-dom	The focused element can be:	given by a coreferent pointing gesture	"**The triangles**" with a gesture pointing out one triangle (in this particular example, a pointed object is extended to a group of similar objects, so only the plural is relevant).
		given by a previous referential term	"The square with circles around" followed by "**the group**" (this is also a particular case).
gen-ref	No coreferent pointing gesture.		"**The triangle** is a simple geometrical form", "**the triangles** have three sides" (numeral adjectives are impossible).

Figure 3: Definite noun phrases.

also correspond to 'ext-par-ref' mode, the differentiation criterion for the referents identification being the rank. Words like "other" and "next" have particular mechanisms. They refer to not-focused elements of a domain that has just been used, and are then included in 'ext-par-ref' mode.

3 The Modes Linked to a Referential Term

The possible modes considering the type of determiner are grouped in the following tables: Figure 2 for indefinite noun phrases (including headless ones), Figure 3 for definite noun phrases, Figure 4 for demonstrative noun phrases, Figure 5 for personal pronouns, and Figure 6 for demonstrative pronouns. With all these possibilities, we show how complex the relation between terms and modes is. A system that has to interpret spontaneous multimodal expressions must know all this information.

4 Reference Resolution in Theory

In this section we present the foundations for a model of reference resolution using reference domains. From the components of the verbal utterance and from the possible ostensive gesture, we deduce a list of clues that the system may exploit to identify the correct referring mode, the correct reference domain and the correct referent. We start with the determiner and then we detail the role of the predicate and of the other linguistic components. Following (Corblin, 1987), we make the hypothesis that the propositional context (and not only the determiner in the referential term) will favour the specific or the generic interpretation. Indeed, we consider that generic references are not usual in human-computer interaction. Thus, we ignore here the 'gen-ref' mode.

For an indefinite noun phrase, the system may choose between 'new-ref,' 'ext-any-ref,' and 'ind-par-ref' referring modes. The presence of a point-

Mode	**Mechanism details**		**Examples** with singular, plural, numeral adjective
ext-par-ref	A coreferent pointing gesture is impossible. Focussing is necessarily due to a previous referential term.		“Select the blue triangle and the green square” followed by “delete **this square**” (all possibilities).
ind-par-ref	The referent can be:	given by a coreferent pointing gesture	“**This triangle**” with a gesture pointing out one triangle. This is the most common multimodal referring expression (all possibilities).
		given by a previous referential term	“Select the blue triangle” followed by “**this triangle**,” “the triangle” followed by “**this form**” (all possibilities).
ind-par-dom	The focused element can be:	given by a coreferent pointing gesture	“**These triangles**” with a gesture pointing out one triangle with a particular aspect (extension to a group of similar objects, so only with a plural).
		given by a previous referential term	“The square with circles around” followed by “**this group**” (the same particular example than in definites).
gen-ref	Three referring modes can be distinguished:	transition from a gestural antecedent to a generic interpretation	“**These forms**” with a gesture pointing out one triangle (numeral adjectives are impossible).
		transition from a linguistic antecedent to a generic interpretation	“This strange form” followed by “**these forms**” (only plural with no numeral adjective).
		direct multimodal generic interpretation	“**This form**” with a gesture pointing out one triangle, that can be interpreted as “this type of form,” and is by consequence ambiguous with a specific interpretation (only singular).

Figure 4: Demonstrative noun phrases.

Mode	**Mechanism details**	**Examples**
ind-par-ref	A coreferent pointing gesture is impossible. The referent can be given by an obvious intention (it depends on the situation) or by a previous referential term.	“Sélectionne le triangle bleu”/ “select the blue triangle” followed by “supprime-**le**”/ “delete **it**”. One other case is when the pronoun refers to another specimen of the referent linked to the antecedent: “j’ai supprimé le triangle mais **il** est revenu”/ “I deleted the triangle but **it** appears again” (singular and plural are possible).
ind-par-dom	A coreferent pointing gesture is impossible. The focused element is given by a previous referential term.	“Ajoute un triangle vert”/ “add a green triangle” followed by “supprime-**les**”/ “delete **them**” (the plural form is necessary to build on the domain).
gen-ref	A coreferent pointing gesture is impossible. This case corresponds to the transition from a linguistic antecedent to a generic interpretation.	“J’ai ajouté un triangle rouge parce qu’**ils** attirent le regard”/ “I added a red triangle because **they** are eye-catching” (the plural form is necessary).

Figure 5: Personal pronouns.

ing gesture may help the system: if the gesture delimits a set of objects not reduced to the referent(s), the only possible mode is ‘ext-any-ref;’ if the gesture is pointing out the referent(s), the only possible mode is ‘ind-par-ref.’ If no coreferent gesture is produced, there is an ambiguity between ‘new-ref’ and ‘ext-any-ref’ modes. The presence of an activated linguistic domain will favour the ‘ext-any-ref’ mode. In the other case, the predicate will disambiguate: a verb that denotes the introduction of new referent(s) like “add” and “create” will force the ‘new-ref’ mode. The ‘ext-any-ref’ mode will be chosen otherwise.

In the ‘new-ref’ interpretation, the system has to add the new object(s) in the visual domain corresponding to the scene. In this domain, a new partition is created, with a differentiation criterion linked to the predicate. The chosen referent(s) are focused in this partition. In the ‘ext-any-ref’ interpretation, the considered domain is the activated

Mode	Mechanism details	Examples
ext-par-ref	A coreferent pointing gesture is impossible. The focussing is necessarily due to a previous referential term.	"Le triangle, le carré et le rond"/ "the triangle, the square and the circle" followed by the mentional reference "**celui-ci**"/ "**this one**" (singular and plural are possible).
ind-par-ref	Demonstrative pronouns combine a demonstrative reference and an anaphor. They are associated to a pointing gesture to refer to a new object with the characteristics of a previous referent. The focussing is then necessarily due to a coreferent gesture.	In "sélectionne ce triangle bleu"/ "select this blue triangle" followed by "supprime **celui-ci**"/ "delete **this one**," "celui-ci"/ "this one" together with a coreferent gesture refers to another blue triangle (singular and plural are possible).
gen-ref	See gen-ref mode for personal pronouns.	"J'ai ajouté un rond vert et un triangle rouge. **Ceux-ci** attirent le regard"/ "I added a green circle and a red triangle. **These ones** are eye-catching" (the plural form is necessary).

Figure 6: Demonstrative pronouns.

one and the process is the same. In the 'ind-par-ref' interpretation, the process is the same, except that the choice of referents is not free but constrained by the gestural interpretation.

For a definite noun phrase, the system may choose between 'ext-par-ref,' 'ind-par-ref,' and 'ind-par-dom' modes. A fine analysis of the referential term and the possible pointing gesture is not sufficient to disambiguate. All hypotheses have then to be kept. In the 'ext-par-ref' interpretation, the system has to extract and to focus the referent from the activated domain. This referent has to be isolated with the category and its modifiers. For the 'ind-par-ref' and 'ind-par-dom' modes, the system has to build a new domain around the focused referent. The differentiation criterion of the new partition in this domain is the referent category.

For a demonstrative, the process is nearly the same than for definites, except that for 'ind-par-ref' and 'ind-par-dom' modes, the differentiation criterion of the new partition is given by the predicate or by the intervention of a pointing gesture. That shows the main difference between definites and demonstratives: the contrast between the referent and the other elements of the reference domain is due to category (and modifiers) for definites, and to focussing for demonstratives.

For a personal pronoun, the system may choose between 'ind-par-ref' and 'ind-par-dom' modes. The clue to disambiguate is a change in the use of singular or plural forms: if a transition occurs from a singular to a plural form, then the 'ind-par-dom' is identified. In this case, the system has to build a new domain around the focused element, the differentiation criterion of the new partition being the category. In the other case, the focussing nature does not change and then no new domain has to be built.

For a demonstrative pronoun, the presence of a pointing gesture forces the 'ind-par-ref' interpretation ('ext-par-ref' interpretation otherwise). In this case, the system has to extract and to focus the referent from the activated domain, the differentiation criterion being the order of mention. For the 'ext-par-ref' interpretation, the system has to build a new domain around the focused element, the new differentiation criterion being the gestural intervention.

Now that we have these interpretation rules, we can precise what a complete reference resolution model may do. The main point is the creation of a new reference domain, especially for definites and demonstratives. The linguistic and contextual clues are sometimes not sufficient for the delimitation of such a domain. For this reason, we propose to manage underdetermined reference domains, as it is done in (Salmon-Alt, 2001) and then in (Landragin et al., 2002). The linguistic and gestural information allow to build an underdetermined domain that groups all constraints. Then, the reference resolution process consists in the unification of this underdetermined domain with the domains that appear in the context. The one with the best unification result is kept for the referent identification.

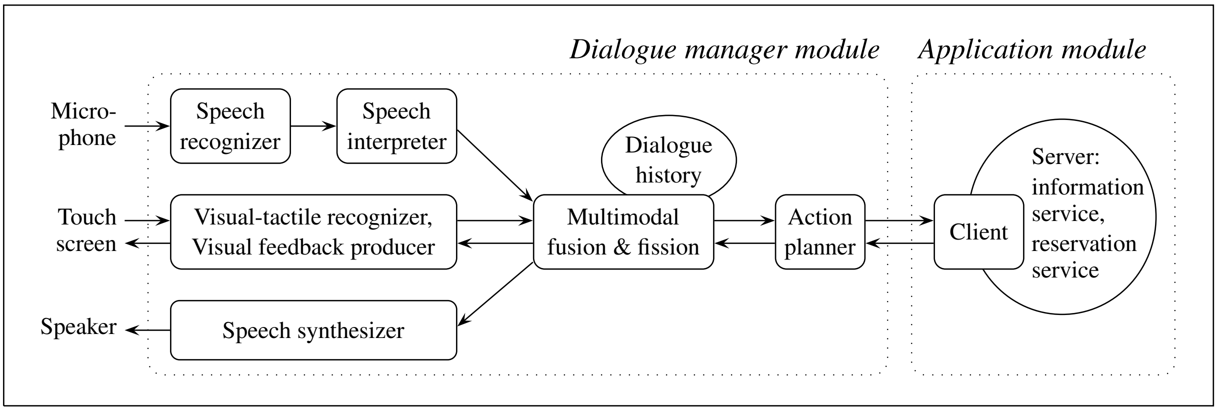

Figure 7: Architecture of the INRIA demonstrator for the OZONE European project.

5 Reference Resolution in Practice

One of the purposes of the OZONE European project is to design a dialogue system with two clearly separated modules, the first devoted to the application and the second (the dialogue manager) to the communicative abilities. Thus, several applications like finding a theater or a cinema, or like reserving a a train ticket or a room in a hotel, can be plugged to the same dialogue manager. A demonstrator, which is currently under construction, will implement that, including a multimodal interaction implying the microphone and the touch screen of a Tablet PC.

Considering a transport information and reservation application, the visual context is a map displayed on the screen. Gestures pointing out streets or train stations can be done. Here is a sample of a dialogue between the user and the system:

User: "How can I get to Paris?"

System: "You can either take a bus, a taxi or a train" (displaying some ways on the map).

User: "How long does it take to go from here to there?" (with an imprecise gesture pointing out a way between two train stations).

System: "Forty minutes..." This answer shows that the system is able to use the dialogue history for managing reference domains linked to the possible transport means (bus, taxi and train), and to exploit the gesture to build on a reference domain that groups the visible train stations.

The input and output are both multimodal, so fusion and fission algorithms are required, as it is showed in Figure 7. The resultant 'multimodal fusion & fission' module has to resolve multimodal referring expressions. Reference domains are managed in this module. The dialogue history is needed to resolve anaphorical expressions. This module also has to translate the whole utterance in a logical form that will be treated by the 'action planner', in order to choose an answering strategy.

Even if the objects are not triangles nor circles but streets and train stations, all that we have explored above has an importance in this demonstrator. With these objects, with spontaneous speech and gesture on a touch screen, all the described referring phenomena and ambiguities are possible. For example: "what are the two stations on the left of the map?," "show me a way to go to Paris," "I want to go to Versailles-Chantiers" followed by "this station is far away from Paris," etc. Thus, the resolution of multimodal referring expressions is an important part of the dialogue management. Moreover, sub-contexts exist and have to be taken into account using reference domains. Our approach and our model of reference seem to be relevant for an implementation like the one of the OZONE project.

6 Conclusion and Future Work

Reference to objects can take several forms which are not linked to particular mechanisms of identification. The choice of a determiner, of the singular or plural form, of a coreferent pointing gesture, lead to clues that specify some aspects of the interpretation process. In this paper we investigate multimodal human-computer interaction involving simple objects. We explore the multiple possibilies of referring modes. We show that even in such a limited context many ambiguities can occur. We propose a list of disambiguation principles based on the notion of reference domain and of the concrete examples we found in the corpus of (Wolff et al., 1998) and in linguistic classical works like (Moeschler and Reboul, 1994), (Reboul and Moeschler, 1998) or (Sperber and Wilson, 1995). The examples we investigate illustrate a number of reference possibilities in terms of anaphor, transition from specific to generic interpretation, associations of referential terms and pointing gestures, etc.

Our proposition is focused on the identification of the referring mode and has to be complemented by an algorithm of reference resolution. The global method presented in (Landragin et al., 2002) fits well this concern, and the implementation of the OZONE demonstrator follows this method. As future research, we plan to make more precise the details of the interpretation process under the light of the classification we present here. One problem, given our focussing on complex phenomena (for example when the pointed objects are not the referents), is the lack of multimodal corpora suitable for evaluating such a classification. Nevertheless, the phenomena can easily be found in human-human communication; we need algorithms for a system to understand these phenomena, even if their evaluation is difficult for the moment.

Acknowledgements

This work was supported by the IST 2000-30026 OZONE EC project (http://www.extra.research.philips.com/euprojects/ozone/).

References

Robbert-Jan Beun and Anita Cremers. 1998. Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition*, 6(1/2):121–152.

Xavier Briffault. 1991. Cognitive, Semantic and Linguistic Aspects of Space. *Proceedings of the 9th IASTED International Symposium on Applied Informatics*, Innsbruck.

Francis Corblin. 1987. *Indéfini, défini et démonstratif.* Droz, Genève.

Francis Corblin. 1999. Mentional References and Familiarity Break. *Hommages à Liliane Tasmowski-De Ryck*, Unipress, Padoue.

Annette Herskovitz. 1986. *Language and Spatial Cognition*. Cambridge University Press, Cambridge.

Annette Karmiloff-Smith. 1979. *A Functional Approach to Child Language. A Study of Determiners and Reference*. Cambridge University Press, Cambridge.

Andrew Kehler. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin.

John Kelleher. 2003. A Perceptually Based Computational Framework for the Interpretation of Spatial Language. Ph.D. Thesis, Dublin City University.

Frédéric Landragin. 2002. The Role of Gesture in Multimodal Referring Actions. *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces*, Pittsburgh.

Frédéric Landragin, Susanne Salmon-Alt, and Laurent Romary. 2002. Ancrage référentiel en situation de dialogue. *Traitement Automatique des Langues*, 43(2):99–129.

Jacques Moeschler and Anne Reboul. 1994. *Dictionnaire encyclopédique de pragmatique*. Seuil, Paris.

Anne Reboul and Jacques Moeschler. 1998. *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*. Armand Colin, Paris.

François Récanati. 1993. *Direct Reference: From Language to Thought*. Blackwell, Oxford.

Susanne Salmon-Alt. 2001. Reference Resolution within the Framework of Cognitive Grammar. *Proceedings of the Seventh International Colloquium on Cognitive Science*, San Sebastian, Spain.

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

Frédéric Wolff, Antonella De Angeli, and Laurent Romary. 1998. Acting on a Visual World: The Role of Perception in Multimodal HCI. *Proceedings of AAAI Workshop on Multimodal Representation*, Madison.