Spatial language, deixis and illustration

Luis Pineda
luis@leibniz.iimas.unam.mx

Abstract

In this paper it is argued that the purpose of graphical illustration is not only to facilitate comprehension of texts and to support effective communication, but also to provide a context through which the referents of textual descriptions in multimodal documents can be fixed. A discussion on how Jackendoff's program of conceptual semantics and Jackendoff and Landau's insights about the relation of spatial language and spatial cognition can inform the conceptualization and design of multimodal generation systems is also presented.

1 Two current approaches to multimodal generation

Intuitively, illustration facilitates comprehension of texts and supports effective communication. However, how this effect is achieved needs to be understood in a principled way. A common assumption in multimodal generation systems is that a multimodal document can be understood as a sequence of acts whose purpose is to achieve a communicative goal. According to this, theories about the structure of text, such as Rhetorical Structure Theory (RST) (Man and Thompson, 1988), in which a text is structured as a hierarchy of rhetorical relations consisting of a *nucleus* and a number of *satellites* which state the essential and contingent parts the message, can be extended or extrapolated to incorporate information expressed through non-textual modalities. Examples of relations in RST are *motivation*, *elaboration*, *enablement*, etc. An operational version of RST for text and multimodal generation has been developed by (Moore, 1995). Another interesting case of study in this direction is the WIP system (Wahlster et al., 1993) in which a text illustrated with pictures is thought of as hierarchical structure in which some of the rhetorical relations are expressed textually, as in RST, but some others through graphical means. This hierarchy is the product of an incremental planning process that aims to achieve a given communicative goal. In a typical example, the instructions for filling the water container of a coffee machine are expressed by a rhetorical structure in which the main act and one satellite or subsidiary act are expressed textually (i.e., the *request* act *remove the cover* and the *motivation* act *to fill the container* respectively) but a subsidiary act providing information to *enable* the task is expressed graphically as shown in Figure 1. Graphical rhetorical relations can also be partitioned in main and subsidiary graphical acts, and pictures can be composed dynamically at the time the hierarchy is produced through the planning process. One additional advantage of this approach is that multimodal documents generated out of a main rhetorical act are coherent, as they can be read as structured objects where the context and the referents for subordinates relations lay within the boundaries set off by the superordinated nodes of the structure.

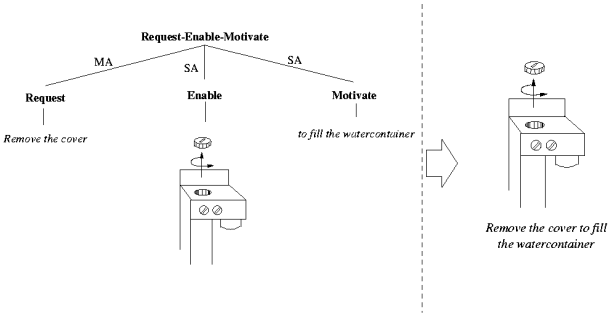

Figure 1: Multimodal rhetorical structure

Promising as this approach might seem, its wide application is limited due to the complexity of the planning task, to the limitations on the kind of pictures that lend the compositional analysis, and to the difficulty of using these kinds of models by human-users, as complex logical specifications, comprehensible only to the specialist, need to be employed to specify presentations. Another problem for this and related multimodal generation techniques is how to allocate information to specific modalities. Here, a large number of criteria can be found in the literature (for instance, (Feiner and McKeown, 1993; Arens et al., 1993)) and an intuitive agreement ex-

ists in that graphics are more effective to express concrete information while text is best, and some times indispensable, for expressing information with an abstract character; however, no exhaustive classification and general agreement about optimal domain independent media allocation rules is available.

A practical alternative for the construction of multimodal presentation systems in which predefined "canned" pictures can be used for illustration has been suggested by van Deemter (Deemter, 1998). In this latter approach pictures for illustrating actions are annotated with a logical representation of their meanings; through an interface, users can select a text to be illustrated with a set of predefined pictures, and a semantic method for retrieving an appropriate picture from memory is presented. In this method, the semantic representation of a reference text is used as the "index" for a picture in the picture's database. In the search for a sound rule of similarity, van Deemter suggests that the picture selection process can be thought of as a valid deductive inference and explores two possible implementation strategies. In the first, the picture selected by a text is the most general (weakest) picture whose representation implies the representation of the reference text (so-called *Rule A*); in the second, the picture selected is the most specific (strongest) picture whose representation is implied by the representation of the text (*Rule B*). These two rules have a fixed referent which is the representation of the text, the index, and selects the picture with the closest meaning. Let T and P be the sets of models satisfying the text and the picture, respectively; note that while Rule A selects the least informative picture such that T includes P, Rule B selects the most informative picture such that P includes T. Although at first sight approaching the meaning of a text by the meaning of a picture *from within* or *from without* seem equally plausible strategies, van Deemter argues that, at least for the purpose of illustration of textual descriptions of actions, Rule B rather than Rule A should be employed. The argument is that pictures are usually under-specified and express less information than the text they are supposed to illustrate, and as long as a picture does not expresses information not contained in the text which could mislead the reader, prompting a false implicature in the Grecian sense, it can indeed be used for illustration. Additionally, pictures might be ambiguous, and if there are interpretations of a picture which are not interpretations of the text then such a picture would not imply the text, preventing the use of rule A. The inclusion relations corresponding to both of the rules are illustrated in Figure 2.

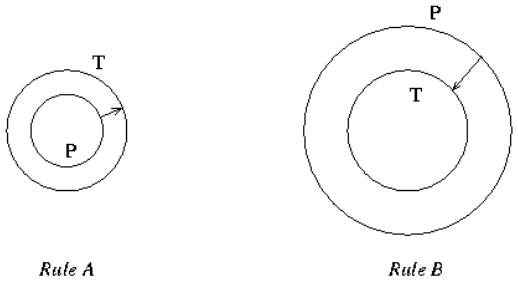

Figure 2: Picture selection rules

2 How illustration helps

After this brief summary of these two approaches to multimodal generation one question that comes to mind is what is the purpose of graphical illustration in general. In the example of WIP mentioned above, for instance, the relation of *enabling* that was expressed graphically could also be expressed textually, as it is done in natural language generation system based on RST. In van Deemter's approach, on the other hand, pictures do not convey information which is not expressed by the text, making graphical information redundant. So, why to illustrate at all? We suggest that an explicit answer to this question could help to understand better current approaches to the problems of content selection and modality allocation in multimodal generation systems and open new lines of research. With this question on mind, next we address the relation of linguistics and graphical communication. First, we discuss how visual recognition of objects and scenes supports effective communication; then we review how the nature of the language that is used to talk about spatial objects and relations illustrates the organization of spatial representations, and finally we discuss the relation between illustration and indexicallity.

2.1 Visual recognition of objects and scenes

One possible reason for illustration is that visual information can enhance the effectiveness of the interpretation process. In this regard it is interesting to mention that according to Biederman's Recognition by Components (*RBC*) theory of object recognition in high-level computer vision (Biederman, 1990), there is a large number of terms in the mental lexicon that name familiar concrete objects which share a characteristic shape (e.g., a chair, a giraffe or a mushroom); following (Jolicoeur et al, 1984) these terms are called *entry-levels*. The figure of these reported for English is approximately 3000, and similar figures are likely to hold for other human languages. *RBC* suggests that entry-levels index spatial representations of objects. These representations are compositions of 3-D spatial primitives which can be recognized in terms a small number of

invariant viewpoint properties. The building blocks of these compositions are called *geons* and are produced by a generating axis and a cross section, a small set of specific "generalized cones" along the lines of Marr's theory of vision (Marr, 1982). Consequently, the concept of a thing that has a generic shape can be activated either through visual perception or through the linguistic modality, or both; consider also that recognition time for visual objects is very fast, in the order of 100-milliseconds, suggesting that the concept referred to by a picture of a thing can be activated on the mind of the human-interpreter long before the corresponding linguistic sign is read or heard in the input string. Activation of the lexical concepts through the visual channel can also help to rule out potential lexical ambiguities, facilitating incremental linguistic interpretation. Another interesting experimental result is that the time required for the recognition of familiar scenes is in the same order of magnitude than the recognition time for individual objects (Biederman, 1990) and, as a consequence, the concepts of a number of graphical objects and relations can be activated by a simple glance, facilitating greatly the interpretation process. So, using pictures to illustrate, even if they convey redundant information, provides for effective presentations. Furthermore, if only the recognition of objects and familiar scenes is required text information might be redundant.

2.2 Linguistic descriptions for spatial objects and relations

Now, we look at the relation between linguistic and graphical information from the point of view of language. An insightful source for this is Jackendoff's program of conceptual semantics (Jackendoff, 1992), specially in relation to spatial language and spatial cognition (Jackendoff and Landau, 1992) (from here on J&L). This program has the purpose to answer the questions of what is the relation between language and spatial cognition such that it allows people to talk about visual perception, on the one hand, and whether spatial language provides a window on the nature of spatial cognition, on the other. The basic assumption is that any aspect of spatial understanding that can be expressed in language must also be expressed in the underlying modules of spatial cognition, where the knowledge required for object recognition, search, location and navegation is represented.

From a critical review and extension of Lieberman's theory of object recognition, J&L suggest the so-called *design of language hypothesis (DLH)*. They notice that large amounts of information about the world codified in spatial representations of objects is filtered out in their corresponding linguistic descriptions, and significant aspects of spatial cognition cannot be expressed linguistically: "one can recognize with great accuracy complicated contours and surface patterns but they are very hard to describe to someone else" (ibid., pp. 120). Additionally, metric information used for motor behavior, for instance, cannot be expressed without using a culturally stipulated system of measurements; consequently, the linguistic description of objects is highly restricted in relation to the spatial counterpart and qualitative spatial information is best expressed through graphics.

J&L also notice that while objects that are being *named* can be differentiated in relatively complex geometrical terms, objects that are *located,* and also the regions in which they are located, receive very schematic geometrical descriptions. Although *RBC* provides for spatial descriptions of objects, orientational features, like *top* and *bottom*, or *front* and *back*, cannot be easily captured in this theory. Parts like tops and bottoms are different in quality to proper parts like doors or wings, as the former terms name parts in relation to the inherent orientation of objects. In order to capture descriptions using such a kind of terms, J&L suggest to augment the expressive power of *RBC* and introduce, in addition of the generating axis of a geon, two orientational axes orthogonal to the generating axis and to each other; axes, in turn, can be marked as directed or symmetric. With these distinctions on hand spatial-linguistic properties of objects like *ends, front* and *back* can be defined. If an object has a horizontal generating axis that is relatively larger than the orientating axes, that is, it if is relatively long and narrow, for instance, it can be said to have *ends*; if an object has a horizontal directed axis which normally faces the observer or the direction of motion, the region determined by that end of the axis is the *front* of that object. J&L also suggest to generalize the theory of volumetric primitives to the 2-dimensional case, to be able to define *surface* objects like disks, lakes, rugs and tabletops. They also suggest extensions to name objects with wholes, containers and related objects. Finally, they conclude that with the help of the additional expressive machinery relatively rich descriptions of objects can be produced, and these descriptions can be used for object identification.

However, the language used for expressing spatial relations has rather different properties. Consider first that spatial relations are mainly expressed through prepositions, and that the number of these is rather small, a hundred at the most, while the number of names for spatial objects is in the thousands. This suggests that there is a limit on the spatial relations that can be expressed through language, and on the amount of information that can be expressed about the objects standing in such relations. Spatial prepositions normally denote a re-

lation between a figural object and a spatial region which in turn is demarcated by a reference object. In the expression *unscrew the cap and squeeze a small amount of ointment, about the size of a match-head, on your little finger*, for instance, the small amount of ointment is the figural object for the preposition *on* and *your little finger* is the reference object demarcating the spatial region were ointment is to be applied. The observation is that figural and reference objects receive always very coarsed linguistic descriptions. J&L argue that "most of the detail in the system's shapes descriptions is concentrated on the reference object –which defines the space in which the figure is located– and even that is highly restricted. The geometry of the figure goes beyond 'thing there' only in the small class of cases in which the issue is its orientation (*along*, *across*, *around*) or its distribution through a region (*all over*, *throughout*)" (ibid., pp. 123).

The contrast about amount of information contained in descriptions used to identify objects and the ones used to locate objects in spatial relations suggests that the information expressed through spatial language reflects the amount of information hold in the spatial representation for the corresponding spatial inferential task. On the basis of this observation J&L proposed the so-called *Design of Spatial Representation Hypothesis* (*DSRH*) according to which the difference on the kind of descriptions used for object identification and object location reflects a very important property of spatial cognition: there is a very strict demarcation of the *what* and *where* information for the representation of spatial information. The *where* system is largely schematic and contains only the information that is essential for object search and for locational and navegational purposes; the *what* system, on the other hand, provides a considerable amount of concrete detail about object's shape.

The suggestion is that the *what* and *where* systems are different modules of spatial cognition, from which linguistic descriptions can be articulated. To represent two cars on a map, for instance, a mark to state the position for each of these objects would be enough for locational and navegational purposes; however, for making sense of the representation fully, to know which mark stands for which car is also required. Consequently, the marks representing the positions of the cars in the locational module of spatial cognition must be bounded to their corresponding entries in the conceptual structure. For visual recognition, on the other hand, the detailed shapes of the cars, say in a photograph, must be linked to the corresponding entry in the conceptual structure for identification purposes. The advantage of this architecture of spatial representations is that tasks involving location and navegation can be accomplished without carrying the heavy informational weight of the shape descriptions of these objects; on the other hand, to identify an object can be done regardless is location or its relations to other objects. Also, while shape descriptions in the *what* system stand for general concepts, token instances in the where system stand normally for concrete individuals: the *what* system is recognition module in which a visual lexicon plays an essential role while the *where* system is a working memory for representing concrete situations in such a way that the mind is in direct contact with the world through perception, or perhaps with some internal reality through imagery. Interestingly enough (Ungerleider, 1982) have found that damage to the inferior temporal cortex of monkeys produces deficits in pattern and shape recognitions, the *what* system, whereas damage to the posterior parietal cortex impairs following routes, reaching for objects and using landmarks to locate objects. Similar results have been found in people with brain damage to the *what* system but leaving that the *where* intact (Fara et al., 1988).

2.3 Graphical context and indexicallity

Multimodal information is not always redundant. In many presentations the purpose of illustration is not only to enhance effectiveness but also to convey information not present in the text (Nielson and Lee, 1994). In WIP's multimodal explanation in Figure 1, for instance, the graphical illustration introduces referents for objects and actions mentioned in the text. In this and similar examples, the interpretation of the multimodal message depends on the ability of the interpreter to correlate graphical and linguistic information in a correct and a efficient fashion. To appreciate this better, consider the following text from [Bransford and Johnson, 1972]:

> If the balloon popped the sound wouldn'ttt be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could not be accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.

For most people, this text is unintelligible unless it is supplied with the illustration in Figure 3.

From a quick glance at the picture, it can be appreciated that the referents for many terms and expressions denoting individuals, properties, actions, locations and paths, etc., are immediately available from the graphical illustration. It follows that to understand a multimodal message in which symbols and expressions of both of the modalities contribute with informative content, correference relations between graphical and linguistic terms and expressions must established. Note also that the lack of proper anaphoric antecedents for the terms and expressions that make the text so obscure, in conjunction with the facilitation provided by a non-linguistic context, which happens to be graphical, suggests that the interpretation of such terms and expressions is indexical rather than anaphoric.

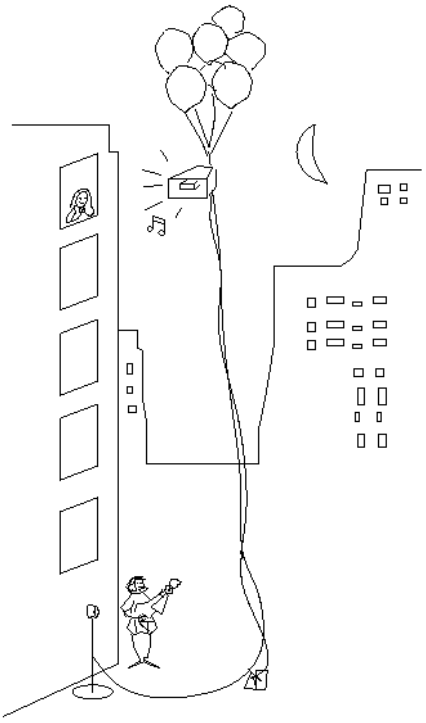

Figure 3: Graphical context for indexical interpretation

For our current purpose, it is interesting to note that a consequence of the distinction between the *what* and *where* systems of spatial representation is that while the links between entry-levels in the mental lexicon and the spatial description of their corresponding shapes in the *what* system are static information, useful to activate the concept of a thing either through the linguistic or the visual channel, the links between individual concepts of things (and also of places, paths, and even events, actions, properties and amounts, according to the types of conceptual entities posed by conceptual semantics(Jackendoff, 1992)), and their schematic representations in the *where* system must be established dynamically for every instance of a spatial cognition problem. Similarly, the links between the figure and reference object and region of a prepositional phrase and their corresponding graphical illustrations, the situation depicted by the prepositional phrase in the *where* system, must also be established dynamically. The implication of this for Bransford and Johnson's example is that as the picture describes a situation in a very coarse and schematic fashion, presumably in the *where* system, and also provides a context for the interpretation of the referential expression in the text, the links between the referents introduced by the text and their corresponding graphical illustrations have to be established dynamically. Consequently, as these textual referents are not interpreted in relation to the meanings of lexical or sentencial concepts, nor through anaphoric inferences (i.e., there are not linguistics antecedents available), the interpretation of text illustrated by pictures is mainly an indexical phenomena.

3 Some thoughts for multimodal generation

Visual object recognition, the definition of different modules of spatial cognition and the static versus dynamic definition of links between signs standing in correference relations have implications for multimodal generation. One direct consequence is that if the purpose of illustration is to activate the concept of a thing through its graphical representation what matters is the particulars of object shape, without regard to its location or relation to other objects. If the intention of the interpreter is to illustrate a location, a path or a spatial relation in a concrete spatial situation, on the other hand, the shape properties can be abstracted away, as spatial or geometrical relations is what matters.

A second important implication for multimodal coherence is that the meaning of a generic picture, in the same way that the meaning of a word, is linguistic knowledge: the link between a sign and the concept it represents is wired up in the mind of the interpreter when such linguistic knowledge is acquired. Words in the lexicon and pictures in the visual lexicon have a meaning, but these signs out of a context do not refer. Correference relation between signs of one or several modalities, on the other hand, have to be established dynamically. These referents are not meant to activate a concept in the mind and as a consequence they do not have meanings; their purpose is rather to pick up individual object and relation in the world, and consequently have referents. This reflection evokes suggestively Donnellan's (Donnellan,1966) distinction between the attributive versus the referential use of descriptions. In the attributive

use, the individual that is described is assimilated into a conceptual class in the interpreter's mind, and then the purpose of the description is to establish a permanent link in a conceptual structure; in the referential use, on the other hand, the purpose of the expression is to pick up a particular individual in a particular situation establishing a contingent link between a term and an object in the interpretation context. Kaplan's DTHAT operator (Dthat,1978) is the formal device to achieve this latter goal. In general, the process of establishing these kinds of coreference relations has been labeled the problem of multimodal reference resolution, and a theory and an algorithm for resolving these kinds of reference is advanced in (Pineda and Garza, 2000), where an extensive case to argue that the phenomenon is indexical is also presented.

These observations have implications for both of approaches to multimodal generation discussed in the introduction of this paper. In relation to van Deemter's picture selection rules it is not the meaning of the picture that has to be used for retrieving a picture but rather its purpose regarding the *what* versus *where* distinction. If an action is depicted as the relation between shapes, either in a single or several graphical states, pictures can be quite underspecified, as the shapes of the objects involved would be accessed through a different channel, namely, the linguistic one. With this reflection on mind an alternative strategy for selecting pictures would be defined as follows Rule B' as follows: used the weakest picture whose representation is implied by the text (a variant of Rule B) such that figural and reference object and region in spatial relations can be bound to schematic representations of spatial objects and places in the picture. This rule would allow to define a database of very schematic pictures that can be quite underspecified but applicable to many situations, as the reference text would fix its referent for particular illustrations.

Another consequence of the distinction of *what* and *where* system is that valid deduction for selecting pictures is a somehow restricted principle. Consider that a very schematic picture can be very ambiguous, specially if shape information is abstracted away, and a relaxation of Rule A such that if there is a model for the picture included in the set of models for the text, that picture is acceptable for illustration, even if there are some models for the picture that are not models for the text. In this situation, what matters is that the geometrical relations that are relevant for the illustration are present, even if there are interpretations for the graphical symbols that are irrelevant for the context. This is so because these would not be considered by the interpreter as the identification of objects would depend again of linguistic information. Rules A' and B' are

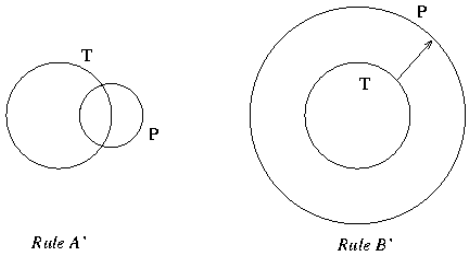

Figure 4: Relaxation of picture selection rules

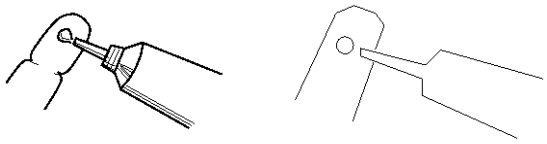

Figure 5: The use of shematic pictures

illustrated in Figure 4.

To appreciate the difference between the original and relaxed pictures selection rules consider that if the user selects the text *spread the ointment on your little finger*, Rule B would chose the illustration Figure 5a, as it is the most informative picture implied by the text; however, if the purpose is to illustrate the relation between finger, ointment and tube, the illustration in Figure 5b would be appropriate also, despite the little informative content that it carries. Figure 5a illustrates information in both the *what* and *where* system, but Figure 5b illustrates only the information in the where system. The *what* information in this latter case is supplied by the text.

Now, we conclude this paper on a reflection of how these observations affect the underlying strategy of systems like WIP. Texts in multimodal documents contain references to individual objects, properties, locations, paths, events and states. However, if these referents are introduced through descriptions used in the referential sense they must be bound either to a referent introduced before in the multimodal communication flow or to a referent introduced through a different modality. If proper anaphoric antecedents are not available, as in WIP or in Bransford and Johnson's text, one can tell what the text means but not what it refers to. Here, the function of graphics is to introduce referents of the appropriate semantic type for all names and descriptions used referentially.

The present discussion suggest a new strategy for illustration in multimodal generation, as follows: identify the types of referents for the expression that need to be illustrated. Then, select a proper picture according either Rule A' or B'.

4 Acknowledgments

The author gratefully acknowledges the support from the Institute for Applied Mathematics and Systems (IIMAS) at the National University of México (UNAM) and Conacyt grant 400316-5-27948-A.

References

Yigal Arens, Eduard Hovy, and Mira Vossers (1993). On the knowledge Underlying Multimedia Presentations. In Intelligent Multimedia Interfaces, edited by Mark T. Maybury, pp. 280-306. AAAI Press / The MIT Press.

Irving Biederman (1990). Higher-Level Vision. In Visual Cognition and Action: An Invitation to Cognitive Science, Volume 2. edited by Daniel N. Osherson, Stephen M. Kosslyn, and John M. Hollerbach. pp. 41-72. Cambridge, Mass.: MIT Press.

Bransford, J. D. & Johnson, M. K., (1972). Contextual Pre-requisites for understanding: some investigations of comprehension and recall. Journal of Verbal Learning and Verbal Behaviour, 11, 717-726.

Van Deemter (1998), "Retrieving Pictures for Document Generation".In Proc. of Fourteenth Workshop on Language Technology, University of Twente, The Netherlands, pp.117-128.

Donnellan, S. K., (1966). Reference and Definite Descriptions. The Philosophical Review, 75, 298.

David Kaplan. 1978. DTHAT. Syntax and Semantics, Vol. 9.

Farah, M. K., Hammond, D. Levine, and R. Calvanio, (1988), Visual and Spatial Mental Imagery: Dissociable Systems of Representation, Cognitive Psichology 20, 439-462.

Steven K. Feiner and Kathleen R. McKeown (1993). Automating the Generation of Coordinated Multimedia Explanations. In Intelligent Multimedia Interfaces, edited by Mark T. Maybury, pp. 117-138. AAAI Press / The MIT Press.

Ray Jackendoff, (1992), What is a concept, That a Person May Grasp It?, in Languages of the Mind: Essays on Mental Representation. pp. 21-52. The MIT Press.

Ray Jackendoff and Barbara Landau, (1992), Spatial Language and Spatial Cognition, in Languages of the Mind: Essays on Mental Representation. pp. 99-124. The MIT Press.

Jolicoeur, P., M. A. Gluck, and S. M. Kosslyn (1984). Picture and Names: Making the connection. Cognitive Psychology 16, pp. 243-275. MIT Press.

Mann, W. C. & Thompson, S. A. (1988), "Rhetorical Structure Theory: Towards a functional theory of text organization", Text 8(3), 243-281.

Marr, D. (1982). Vision. San Francisco: Freeman.

Johanna Moore. 1995. Participating in Explanatory Dialogues: interpreting and responding to questions. A Bradford Book, The MIT Press, Cambridge.

Irene Nielson and John Lee, (1994), Conversation with graphics: implications for the design of natural language/graphics interfaces. International Journal on Human-Computer Studies, Vol 40. pp 509-541.

Luis Pineda and Gabriela Garza. A Model for Multimodal Reference Resolution. Computational Linguistics 26(2), pp: 139-193.

Ungerleider, L. G., and M. Mishkin (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield eds., Analysis of Visual Behavior, 549-586. Cambridge, Mass.: MIT Press.

Wolfgang Wahlster, Elisabeth André, Wolfgang Finkler. Hans-Jürgen and Thomas Rist. 1993. Plan-based integration of natural language and graphics generation, Artificial Intelligence 63: 387-427, Elsevier.