EUROGRAPHICS 2003 / P. Brunet and D. Fellner
(Guest Editors)

Volume 22 (2003), Number 3

Multimodal generation, spatial language and illustration

Luis A. Pineda

Abstract

In this paper the role of graphical illustration in multimodal presentation systems is discussed. First, a review of two current approaches to multimodal generation systems is presented; from this, the question how illustration supports the interpretation process is addressed from both a perceptual and a cognitive perspective. Following Jackendoff's semantics, it is described how representations and computational processes involving the identification of objects, on the one hand, and those involving the interpretation of spatial relations, like position, size and orientation, on the other, have very different cognitive and computational properties. The relation between the meaning of an expression and its interpretation in relation to a multimodal context is also discussed. Finally, on the basis of the properties of the language that is used to talk about space, and the contextual factors of the interpretation of multimodal expressions, a new strategy for the design of multimodal presentation systems is presented.

1. Two 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)[17], 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 of 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[19]. Another interesting case of study in this direction is the WIP system[26] in which a text illustrated with pictures is thought of as a 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

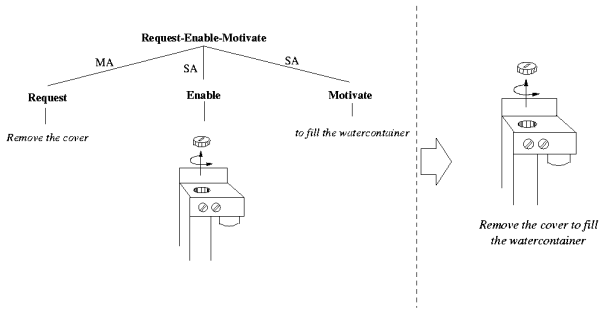

Figure 1: *Multimodal rhetorical structure*

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 demarcated by the superordinated nodes of the 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 compo-

sitional 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[11], and an intuitive agreement exists 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.

For all these reasons, it is interesting to investigate practical alternatives for the construction of intelligent authoring systems. In this regard, van Deemter suggests that an ideal multimodal authoring system should have facilities for: (1) easy determination of content, (2) easy determination of style and layout, (3) easy allocation of media, (4) easy annotation of non-generated presentations and (5) easy post-editing[7]. In this context, Van Deemter presents the system PILLs, which attempts to meet points (1) and (2), and sets the ground for addressing points (3) and (4) too.

For the knowledge acquisition process, point (1) in the list, PILLs uses a knowledge editing method called WYSIWYM[23]. This method provides an interactive interface to a KL-ONE-type knowledge-base (KB)[2, 3, 4], where the knowledge or information content of the application domain is stored. In this family of KBs two kinds of knowledge are distinguished: terminological and assertive. Terminological knowledge is related to the meaning of words and sentences that are needed to describe a knowledge domain, and it is represented through expressions of the so-called terminological box or T-BOX. This kind of knowledge consists of the concepts that one needs to know beforhand to be able to engage in a conversation or a problem-solving task in a given knowledge or application domain. Expressions in the T-BOX are similar to dictionary entry definitions, in that when one needs to know the meaning of a word, one can look up its definition in the dictionary or in an encyclopedia. However, one can know the meaning of a word, or the concept expressed by a word or a sentence, without having a particular object or situation in mind; expressions in the T-BOX are intended to capture only what the symbols in the representational language mean, and for this reason these expressions have no referential content nor assertive import. To relate words with particular objects, on the other hand, a specific situation in the world with a number of individuals with their properties and relations is required: a context. Names are used to identify objects in specific contexts, and prepositions, for instance, are used to identify the spatial or temporal relations between the objetcs in the context. For this reason, knowledge related to a situation is called assertive, and in KL-ONE and related formalisms it is represented through expressions of the so-called assertional box or A-BOX; expressions in this latter structure have a full referential content and assertional import. Also, pecific problem-solving situations represented in the A-BOX hold only during the time the context is present, and changes in the world or progress in the course of a problem-solving task can change the state of the A-BOX.

In PILLs, the T-BOX is codified in advance by a specialist in the application domain, and final users are allowed to describe particular problem situations in the A-BOX through WYSIWYM. The method allows users to create entries in the KB through an incremental refining process; it also provides a natural language generation facility for producing template-like textual descriptions of the content of the KB as a feedback through which users can verify that the input to the KB is what they mean; hence, the name WYSIWYM: *What You See Is What You Meant*. Once the content of the document has been provided by the user through this knowledge acquisition facility, PILLs can generate coherent natural language expressions of the information contained in the KB through a second, more sophisticated, natural language generation facility. These latter descriptions correspond to the output text of the authoring tool proper.

PILLs meets also point (2) of van Deemter list through the WYSIWYM editing method also, but using instead a second KB where knowledge about presentation styles and layout, which is used during the document generation process, is stored.

In order to meet points (3) and (4), PILLs has been extended with ILLUSTRATE. Through this new facility, users can select a stretch of the feedback text with the mouse during the knowledge acquisition process with the intention to illustrate graphically the content of its underlying description in the A-BOX. For this purpose, a predefined library of pictures is used. Here, a very interesting feature is introduced: through WYSIWYM, users are enabled to build descriptions for pictures in the picture's library[†]. With this strategy, the format of the descriptions in the KB representing the meaning of pictures is the same as the format of the domain knowledge descriptions stored in the KB.

ILLUSTRATE uses the underlying representation of the text in the A-BOX and the descriptions of all the pictures in the library. The question is then, what picture should be

[†] Whether these representations capture the full meaning of pictures is not addressed in Van Deemter's system. In this regard, a very interesting question for computer graphics and computer vision is how to induce such kind of representations through a computational process; however, this problem is very hard, specially if it is considered that pictures are not photographic but figurative images of objects in the world. To address this problem, a theory of semantics of graphics in which graphical symbols have a conventional interpretation needs to be defined. For a proposal in this direction see[21].

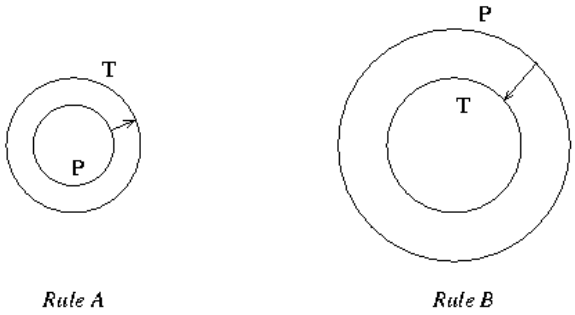

Figure 2: *Picture selection rules*

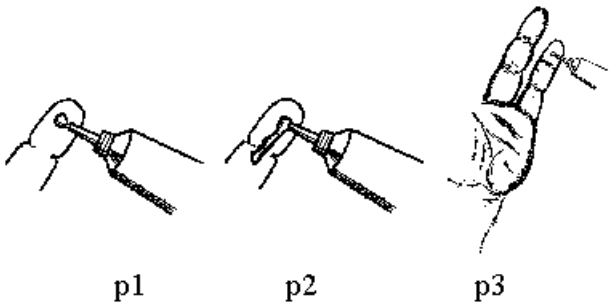

Figure 3: *A picture's library*

chosen for a particular illustration process. To answer this, van Deemter advances a semantic method in which the picture chosen for illustration is the one whose meaning, or rather, the meaning of its corresponding description in the A-BOX, is the closest one to the meaning of the description underlying the feedback text to be illustrated [6]. To implement this method, a formal notion of similarity is required; for this purpose, van Deemter suggests that the picture selection process can be thought of as a valid deductive inference. He 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 use the logical representation of a text in description logic, created by the authors through WYSIWYM, as a fixed reference (i.e., a kind of semantic index), and select the picture with the closest meaning, expressed through the same formalism. 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*. The inclusion relations corresponding to both of the rules are illustrated in Figure 1.

Although at first sight approaching the meaning of a text by the meaning of a picture *from within* or *from outside* seem equally plausible strategies, van Deemter argues that *Rule B* rather than *Rule A* should be employed for the picture selection process. The argument is that if the representation of a picture implies the representation of a text (i.e., *Rule A*), all the information contained in the text is also contained in the picture, as this latter object is more specific, but there might be additional information in the picture that could mislead the reader, prompting a false implicature in the Gricean[12] sense; subsequently, *Rule A* should not be considered for illustration.

Next we illustrate the picture selection process using Van Deemter's example. In general, descriptions in the KB are of the following form:

$\mathrm{TYPE}_0(e)$ &

$\mathrm{ROLE}_1(e) = x_1$ & ... & $\mathrm{ROLE}_n(e) = x_n$ &

$\mathrm{TYPE}_1(x_1)$ & ... & $\mathrm{TYPE}_n(x_n)$

where e, x_1,..., x_n are variables or constants standing for instances of types; each instance has one property, called its type, and can have a number of attributes with their corresponding values (the roles). The values are, in turn, instances of other types[6]. The expression *squeeze a small amount of oinment on your little finger*, for instance, can be represented by the description **t** through WYSIWYM, as follows:

t:

Squeeze(e) &

Actor(e) = Reader &

Actee(e) = z &

Ointment(z) &

Quant(z) = Small &

Source(e) = t &

Tube(t) &

Target(e) = u &

LittleFinger(u) &

Owner(u) = Reader

Informally, **t** states that there is a squeezing event *e* such that the actor of *e* is the reader, the actee of *e* (the squeezed thing) is *z* and *z* is of type oinment; also, the quantity of this substance is small, the source of *e* is the tube *t*, the target of the squeezing event is the little finger *u* and the owner of which is the reader. Here, the verb *squeeze* is defined as a basic constant in the terminological box of the KB and represents a concept that requires a number of roles whose values are of certain types.

Suppose now that there are three pictures available in a picture data-base as shown in Figure 1.

Let us assume the representations of pictures **p1**, **p2** and **p3** are as follows:

p1:

Squeeze(e') &

Actee(e') = z' &

OintmentOrCream(z') &

Quant(z') = Small &

Target(e') = 'u &

Finger(u')

p2:

Squeeze(e') &

Actee(e') = z' &

OintmentOrCream(z') &

Target(e') = u' &

Finger(u')

p3:

Squeeze(e') &

Actee(e') = z' &

OintmentOrCream(z') &

Target(e') = u' &

Finger(u') &

LittleFinger(u')

As was mentioned, these descriptions are produced through WYSIWYM. In **p1**, the actee and target roles of the squeeze event have been specified, but the agent role has not. The representations **p2** and **p3** can be interpreted along similar lines.

Now, we turn to the picture selection process: using the description **t** as the index, select the picture whose description is the closest in meaning to **t**. Clearly, *Rule A* cannot be used to select **p1**, **p2** or **p3** because **t** is not logically implied by the representation of any of these pictures (e.g., non of these pictures show the actor of the squeezing act and whose finger is it). On the other hand, no such problem arises if *Rule B* is used. In this latter case **t** implies all three pictures. Now the criteria of the amount of information carried out by the descriptions must be used to select the most specific, or strongest, picture. For this, consider that the larger the number of properties that an object have, the more specific it is; then, a criteria to state the specificity of a description is simply the number of properties or attribute/value pairs that it has. Using this measure, the specificity of **p1** and **p3** is 6, and the specificity of **p2** is just 5. Accordingly, **p1** and **p3** are equally informative and both are more informative than **p2**. While **p1** states that the quantity of the squeezed thing is small, **p3** states instead that the finger of the target is the little finger. Then, according to van Deemter's strategy, either **p1** and **p3** can be chosen randomly for this particular picture selection process.

After this summary of those two approaches to multimodal generation we can appreciate both of the systems take different approaches to the questions of what and how to illustrate. In WIP content is stated when the goals of the communicative process are defined by the user, and media allocation is determined in terms of media allocation rules, and the media-related rhetorical acts; ILLUSTRATE, on the other hand, avoids the use of modality selection rules and the decision of what to illustrate is made directly by the user. These systems also diverge in how the illustration process is carried out; in the case of WIP, a pragmatic strategy based on the definition and execution of a plan with the purpose to achieve a communicative goal is employed, while in the case of ILLUSTRATE, a syntactic and semantic strategy for assessing similarity of meanings is employed instead. As can be seen, WIP's strategy seems to be more general, but ILLUSTRATE seems to be more realistic for the implementation of practical applications.

2. Visual recognition of objects and scenes

Visual information enhances the effectiveness of the interpretation process: concepts of concrete visual objects are or can be accessed faster through the visual than through the linguistic modality. A computational explanation of this fact is suggested by Biederman's Recognition By Components theory of object recognition (RBC) in high-level computer vision[1]; according to RBC 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[16] these terms are called *entry-levels*. The figure of these reported for English is approximately 3000, and a similar number 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 of 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[18]. 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; furthermore, RBC predicts that, due to the links established at entry levels, the activation of a visual lexical entry will not only activate the corresponding concept, but it can also associate the lexical entry in the linguistic modality and *vice versa*, strenghtening the activation of the associated concept; furthermore, recognition time for visual objects is in the order of 100-milliseconds, much faster than syllables, which are recognized in the order of a third of a second, suggesting that the concept referred to by a picture of a thing can be ac-

tivated on the mind of the human-interpreter long before the corresponding linguistic sign is heard[1].

An alternative view of visual object recognition purports that object representations are based on multiple image-based views that are matched to input shapes through normalization processes. Nevertheless, if an object is held in the mind through a number of view-dependent representations, all these representations must be related, perhaps through an intermediate binding structure, if they are indeed representations of the same object. In this latter view entry-levels would have to index not the 3-D representation of an object, as in RBC, but rather the set of view-dependent representations constituting the representation of the object[24].

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[1] 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 of a text dealing with those objects. Furthermore, if only the recognition of objects and familiar scenes is required text information might be redundant. So, using pictures to illustrate, even if they convey redundant information, provides for effective presentations. However, text can also enhance presentations (i.e., with captions or even full paragraphs that annotate pictures), and indeed, if pictures are taken to be the main modality of a message, text can help or even be indispensable to "illustrate" pictures.

3. Linguistic descriptions of spatial objects and relations

Let us now 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[14], specially in relation to spatial language and spatial cognition[15] (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 navigation is represented.

From a critical review and extension of Biederman's theory of object recognition, J&L 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. In this regard, they suggest that objects that can be identified through visual perception can be described in detail because these linguistic descriptions report the spatial structure of visual entries in the visual lexicon, which is a very rich source of information. The language used for expressing spatial relations, on the other hand, has rather different properties. Spatial relations are mainly expressed through spatial prepositions, and the number of these is rather small, a hundred at the most, while the number of names of types of 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 relation between a figural object and a spatial region which in turn is demarcated by a reference object. This reflects the relation between figure and ground of visual representations. In the expression *squeeze a small amount of ointment on your little finger*, for instance, the small amount of ointment is the figural object (the figure) of the preposition *on* and *your little finger* is the reference object demarcating the spatial region were ointment is to be applied (the ground). The observation is that while prepositions impose very few constraints on the shape the reference object, if any, and subsequently, on the form of the spatial region demarcated by the reference object, the descriptive load of the expression is concentrated on the figure and the verb; *squeeze*, for instance, describes the whole of the surface where the ointment is to be applied; in *the book is {standing, lying, leaning, resting} on the table*, the verb encodes object-internal information (i.e., the intrinsic orientation axis of the 3-D representation of the book)[13]. In addition, spatial prepositions are also rather vague about the possible spatial configurations in which the figure is standing in relation to the ground.

This underspecification results in that there are many possible spatial configurations that can satisfy the relation denoted by the preposition and its arguments, and conversely, specific spatial configurations can be referred to through different expressions employing different prepositions; furthermore, the limits between the situations satisfying the relation, and those that do not, are rather fuzzy[10].

This kind of underspecification has been explored in the context of the VITRA (Visual Translator) project and the SOCCER program in which verbal descriptions are generated out of the graphical representation of an idealized soccer game[22]. In this system, the positions and velocities of dots representing the players and the ball at a given time are translated into a number of spatial expressions, related to the goals of the game, that refer to the visual relations in different ways (i.e., using different prepositions). For this translation a number of complex heuristics, based on the use of a function which describes a "potential field" for each preposition, which is centered on a reference object, are employed. The definitions of these functions (i.e., the form of their po-

tential fields), depend on the ideal geometrical meanings of the prepositions, which are used to chose the more appropriate preposition for each particular visual situation.

The inverse problem of producing the image described by a spatial expression is also studied in VITRA; for this, the system ANTLIMA[22] employs the same kind of heuristics to generate images out of spatial linguistic descriptions that are produced by the visual translator itself. This latter images permit to verify whether the images that can be produced out of the interpretation of the linguistic descriptions are similar to the images originally input to the system; these synthetic images are then used as models of the images that the users can grasp out of VITRA's linguistic output. In this latter translation process, the underspecification of spatial linguistic descriptions is compensated through the functions associated to the ideal meaning of prepositions and their associated potential fields. In the context of this project, it is argued that VITRA implements Herskovits theory of semantics and pragmatics of locative expressions[10].

According to this latter theory, the meaning of a spatial preposition is a function of an ideal geometrical meaning; in each particular situation, this function can be modified by pragmatic factors like relevance, salience, tolerance and typicality of the figural and reference objects of the locative expression. In addition, there is a level of geometric conceptualization that mediates the spatial situation referred to by the locative expression and language; it is at this level where geometrical meanings of prepositions, their geometrical transformations (i.e., functions from geometrical descriptions to geometrical descriptions), and the geometrical descriptions of the objects related in a locative expression, are defined. This theory postulates also an addition level of representation in which conventional aspects of locative meaning are captured in terms of a set of "use types", which state idiosyncratic uses of spatial prepositions directly in the lexicon.

VITRA provides a specific implementation to solve the problem of underspecification of spatial language that refers to spatial relations, and Herskovits provides a comprehensive account of the semantics and pragmatics of locative expressions; however, in these approaches the question of why locative expressions carry so little information content at the time descriptions of objects can be informatively rich is not adressed, neither the question of what is the relation between descriptions of spatial objects and descriptions of spatial relations.

In this regard, J&L suggest that the information expressed through spatial language reflects the amount of information held in the spatial representation for the corresponding spatial inferential task and propose the so-called *Design of Spatial Representation Hypothesis* (DSRH) [15]; according to this hypothesis, 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. While the *what* system provides a considerable amount of concrete detail about object's shape, the *where* system is largely schematic and contains only the information that is essential for object search and for locational and navigational purposes. The advantage of this architecture of spatial representation is that tasks involving location and navigation can be accomplished without carrying the heavy informational weight of the shape descriptions of these objects; object identification, on the other hand, can be accomplished regardless of the location of objects or the spatial relations between objects in a spatial scene. The suggestion is that the descriptions of objects and the description of relations are articulated out of different modules of spatial cognition: while a very rich description of a car can be made out of its photograph, for instance, only a very coarse description of the relation of two cars can be produced out of a schematic map where the cars are represented by simple marks.

Furthermore, there is neurological evidence that the *what* and *where* systems belong, at least in part, to independent functional modules of the brain; it has been found that damage to the inferior temporal brain 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[25]. Similar results have been found in people with brain damage to the *what* system but leaving the *where* system intact[9].

4. Meaning, reference and indexicallity

The meaning of a composite expression depends on the meaning of its parts and their mode of grammatical composition, as stated by Frege's principle of compositionality[8]. However, the reference of a composite expressions is not necessarily a function of the references of its constituent parts. Knowing the meaning of an expression is not always enough to fully understand the expression. This can be illustrated with the following well-known example taken from[5]:

> If the balloon popped the sound wouldn't 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 despite that the meaning of the lexical and sentential units in the text can be grasped. To appreciate this better, consider that terms like *the correct floor* or *the fellow* have meaning and can be understood, but in a rather poor way, as there is no way to tell what the definite descriptions are about. These textual images, the shapes of these words, evoke their corresponding concepts directly, but there is no context to extract a fully coherent content, and there is a feelling that something is missing in the interpretation.

However, if the interpreter is provided with the drawing in Figure 4, the whole thing makes sense. From a quick glance at the picture, it can be appreciated that the referents for many terms and expressions denoting individuals (i.e., *the wire, the fellow*), locations (*on the instrument*), paths (*from the correct floor*), amounts (*flow of electricity*) events (*the balloon popped, prevent the sound for carrying*) and states (*to be well insulated, a situation involving less distance*) are available from the graphical illustration.

To appreciate this better, we refer back to the discussion of KL-ONE in Section 1. Suppose that the lexical and sentential concepts required to understand the meaning of the balloon text are properly codified in the T-BOX of a knowledge representation component of a natural language understanding system. Now consider two interpretations scenarios for the sentences in the balloon example: in the first the knowledge in the T-BOX is the only knowledge available for the interpretation process, but in the second, in addition to the T-BOX representation, the situation in Figure 4 is properly represented in the A-BOX, and this representation is available for the interpretation process. As the interpretation of the linguistic message in the first scenario can only use meanings, the interpretation consists also of meanings only, and cannot be related to the world: such a system would understand the message but not what the message is about. However, in the second scenario, there is a specific interpretation context provided by the graphical information, with referents for the concepts alluded to in the textual message, and the resulting interpretation provides the specific information required to give a fully coherent content to the message. In this latter setting, linguistic and graphical symbols correfer (i.e. have common referents in the actual situation of the world that the multimodal message conveys). Also, in the second scenario it is necessary to bind or correlate graphical and linguistic symbols to make sense of the message. As these symbols are input from different modalities, the bindings must be established dynamically, and indeed, when an interpreter is able to establish such bindings successfully, and a stable and coherent content is accessible, the feeling of understanding is much more accomplished.

Note also that Bransford and Johnson's text contains no pronouns or descriptions that refer back to other terms introduced previously in the same text; in the expression *The fellow was singing loud, he was very excited*, for instance, it is possible to suppose that the reference of *he* is the fellow; here, *he* is an anaphoric pronoun, and the inference by which we come to know that this pronoun refers to the fellow is know as an anaphoric inference process. However, the text in the example was carefully designed to omit linguistic antecedents for the interpretation of anaphoric terms and expressions, closing this possible source of correference relations, which would need to be established dynamically too. However, if a text were presented instead of the picture, and anaphoric antecedents were provided, a context would also be available and the text could be interpreted successfully by establishing appropriate bindings between linguistic terms. On the other hand, if the picture had been presented without the textual description, the referential content would also be absent. The graphical symbols and relations could be understood through their meanings, but a large amount of the information conveyed by the text would also be missing. We can think of the shapes as entries in a visual lexicon which are bound to their corresponding concepts, and we can grasp the meaning of the picture, but again, to establish a full coherent content, a context and a binding process to relate linguistic and graphical symbols is required. We refer to this process as *multimodal reference resolution*, and a theory and an algorithm for resolving these kinds of reference in simple situations (e.g., for interpreting captions of pictures) is advanced in[21].

More generally, a context is a set of individuals, properties and relations that are present in every particular interpretation situation, and this set varies according to the time and place in which a message is interpreted, and also in relation to the speaker and hearer involved in the situation. These dimensions *index* the context of use, and the symbols of a message take their referents from this set for every context or index, and a given symbol or description can refer to different objects in different contexts. For this reason, a context must be introduced somehow in the interpretation process. The information describing a context has to be expressed also through a message, but the relation between a message and its interpretation context is relative: the message is interpreted in relation to the context. So, expressions can take the role of messages in some interpretation situations, but the role of context in others. Accordingly, illustration can be thought of as a relation between figure and ground of a multimodal message. In the same way that an image is inserted and contrasted with its background in a painting or a photograph, a linguistic expression, either textual or spoken, can be thought of as a *figure* which makes full sense in relation to a context, the ground, which might be fully or in part graphical. But from an alternative perspective, the figure can be graphical and the ground textual. Symbols in multimodal messages can have context independent meanings, but to fully refer they must be bound to a context, which can be provided linguistically, graphically or in several modalities.

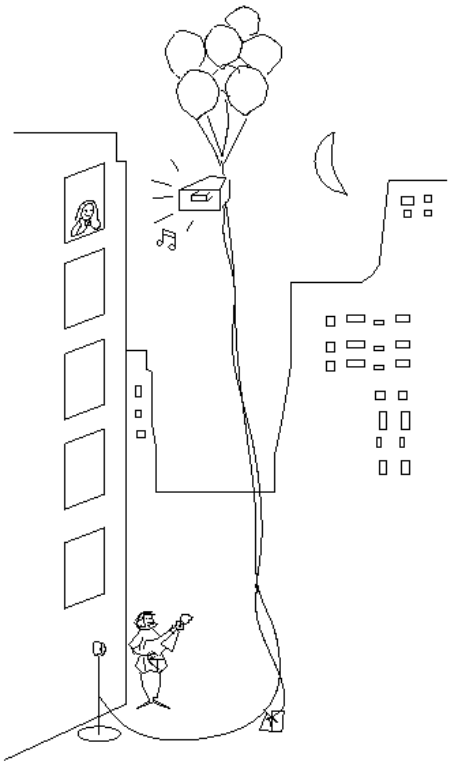

Figure 4: *Graphical context for indexical interpretation*

5. A new strategy for illustration

From the present discussion we make the following observations:

- Multimodal presentations support effective interpretation both by accessing faster lexical information in the *what* system, and by providing direct interpretation of spatial information in the *where* system of human cognition.
- Multimodal presentations including symbols out of context enrich messages by providing meanings of symbols (i.e., providing dictionary definitions through the graphical modality). These presentations take advantage of the *what* system. Graphical configurations can be taken as lexical units (i.e., scenes describing an action).
- Multimodal presentations including symbols of different modalities bound to each other enrich, in addition, the referential content of messages. This kind of presentations take advantage of the *where* system, in addition to the *what* system, of mental cognition.

In multimodal presentation systems, these observations could be used as design guide-lines. If what is needed is to identify an object or a prototypical relation that is described textually, a picture illustrating the object, state or action, ought to be highlighted in as much detail as possible. Multimodal presentations based on this strategy are focused on providing dictionary definitions for textual expressions whose meaning may not be familiar to users. ILLUSTRATE exemplifies this strategy. The picture selection rules employed in this system select a picture in terms of its meaning, and in a context independent fashion. In this regard, note that the information in the textual description **t** in Section 1 that cannot be inferred from the representations of pictures **p1**, **p2** and **p3**, blocking the use of *Rule A*, is precisely the indexical information conveyed by the pronoun *your* in *squeeze a small amount of ointment on your little finger*; this pronoun establishes that the agent of the squeezing action in the interpretation time and place is whoever is the reader. For this reason, this expression has factual import (imperative in this case) and referential content.

If, on the other hand, what is intended is to take advantage of the expressive properties of different media (e.g., providing abstract information through language but concrete information through graphics, yet in a related and unified fashion), the interpretation of one modality should be made relative to the context established by the other. In this latter case, a presentation system should help users to establish proper bindings between parts of the text and their corresponding parts in the pictures. A system with this latter orientation is WIP; it is focused on the construction of an interpretation context, and great emphasis is placed on providing enough information for effective reference resolution. The *enable* relation expressed graphically in Figure 1, for instance, not only provides a graphical background against which the *request* and *motivation* relation can be interpreted, but also permits the user to bind linguistic and graphical symbols, as the three rhetorical structures are produced with a common underlying context and the same planning act. These bindings, however, are not established dynamically by the system, as textual and graphical symbols are realizations on the same underlying variable, assigned to one or the other modality according to the system's medium selection rules.

One question that comes to mind is whether it is possible to design multimodal presentation systems with the simplicity of ILLUSTRATE, and yet with the referential power of WIP. For this, we first notice that using valid deduction for selecting pictures is a somehow restricted strategy. Consider that if there is a model for the picture included in the set of models for the text, it can be used 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 such interpretations would not be considered by the interpreter as the identification of objects would depend on linguistic information also. To implement this strategy we relax *Rule A* and propose *Rule A'* as follows: chose the picture whose representation has the largest intersection with the representation of the index text (in terms of the number of literals, either explicit or implicit, in both of the representations) such

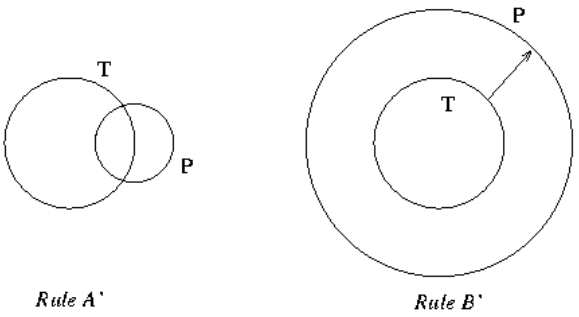

Figure 5: *Relaxation of picture selection rules*

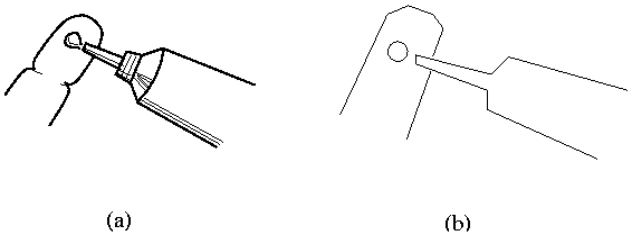

Figure 6: *The use of schematic pictures*

that figural and reference objects of locative expressions in the text can be bound to specific objects in the picture. This strategy would profit from the *what* system because it would choose the picture with the largest conceptual content that satisfies the textual part of the message, and to certain extent of the *where* system, because it would only consider pictures with appropriate figural and reference objects for locative expressions.

However, we can have an additional profit from the *where* system. Pictures expressing rich conceptual information can only be used to illustrate very specific situations; but if conceptual information about objects is relaxed, and schematic pictures about generic relations between objects are included in the picture's database, similar pictures can be used to illustrate different textual descriptions. In this situation, the spatial relations between these objects could be illustrated through schematic graphics and the concepts of things would be made available through text. For this purpose we define a variant of *Rule B* as follows: use the weakest picture whose representation is implied by the text such that figural and reference objects in locative expressions can be bound to schematic representations of spatial objects in the picture. This last condition can be verified with an algorithm for multimodal reference resolution, through with correference relation between symbols of different modalities can be established[21]. We refer to this new rule as *Rule B'*. The effect of this rule would be to select a picture including the relevant spatial relations between the objects in the textual expressions, but at the same time, the one with the least conceptual load in the database. A similar effect could be achieved with a variant of *Rule A'*, in which the smallest intersection between the representations of text and picture that satisfies the binding condition is chosen instead. *Rules A'* and *B'* are illustrated in Figure 5.

To appreciate the difference between the original and the new pictures selection rules consider that if the user selects the text *squeeze a small amount of 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 appropriated also, despite the little informative content that it carries. Figure 5a illustrates information about the concepts of the objects involved in the relation and also about the relation in which they stand to each other, but Figure 5b emphasize only the relational information, as the conceptual information is supplied by the text, and there might be many situations in which this latter picture would be applicable.

In summary, logical representations can be used for selecting pictures by approximating the meaning of a reference text and the meanings of pictures in a picture's database. However, the formulation of the selection rules must take into account not only the logical properties of the representations, but also the purpose of illustration, in particular in relation to the *what* and *where* distinctions.

We conclude this paper with a reflection potentially useful for the interpretation and generation of multimodal presentations, and also for the interpretation and generation of computer graphics. The mind is like a prism that splits perceived objects in different dimensions, and process every dimension by specialized structures, avoiding in this way the need of carrying the full informational weight of all aspects of the perceived world in each aspect of the interpretation. Multimodal information provides additional dimensions or aspects of the objects of the world, with the corresponding number of dimensions in which objects are represented. Computations about the different dimensions of objects can proceed in parallel, distributing greatly the computational effort; objects, on the other hand, are not fully reconstructed to be presented to the mind as wholes, as only the integrative process established by binding processes allows us to perceive sensations as cognitive wholes. These observations, although speculative at this stage, can be considered for the design of multimodal presentations that take the perceptual abilities of people into account, and also for developing new strategies for multimodal generation systems.

6. Acknowledgments

The author gratefully thanks Kees van Deemter, John Lee and Joerg Schirra for useful comments and suggestions, and to Ivan Meza for technical assistance.The author also acknowledges the support Conacyt grant 400316-5-27948-A.

References

1. Biederman, I. (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. 4, 5

2. Brachman, J. R. and Schmolze, J. G. (1985). An Overview of the KL-ONE Knowledge Representation System. Cognitive Science 9, pp. 171–216 2

3. Brachman, R. J., Fikes, R. E and Levesque, H. (1985). KRYPTON: A Functional Approach to Knowledge Representation. In Readings in Knowledge Representation. Edited by R. J. Brachman and H. L. Levesque. pp. 411-430. Morgan Kaufmann Publishers, Inc. Los Altos, California. 2

4. Brachman, R. J., McGuinness, D. L., Patel-Scheider, P. and Borgidaa, A. (1999). Reducing CLASSIC to practice: Knowledge representation theory meets reality. Artificial Intelligence, 114, 203-237. 2

5. 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, pp. 717–726. 6

6. Van Deemter , K. (1998). Retrieving Pictures for Document Generation. In Proc. of Fourteenth Workshop on Language Technology, University of Twente, The Netherlands, pp.117–128. 3

7. Van Deemter and Power, R. (2000). Authoring Multimedia Documents using WYSIWYM. Proceedings of COLING, 2000. 2

8. Dowty, D. R., Wall, R. E. and Peters, S. (1985). Introduction to Montague Semantics. D. Reidel Publishing Company, Dordrecht, Holland. 6

9. Farah, M. K., Hammond, D. Levine, and R. Calvanio. (1988). Visual and Spatial Mental Imagery: Dissociable Systems of Representation, Cognitive Psychology 20, pp. 439–462. 6

10. Herkovits, A. (1985). Semantics and Pragmatics of Locative Expressions. Cognitive Science 9, pp. 341–378. 5, 6

11. Feiner, S. K. and McKeown, K. R. (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. 1

12. Grice, P. (1975). Logic and Conversation. In P. Cole and J. L. Morgan (Eds.). Studies in Syntax Vol. 3. Speech Acts. Academic Press, NY. 3

13. Jackendoff, R. (1987). On Beyond Zebra: The relation of linguistic and visual information. Cognition, 26, pp. 89–114. 5

14. Jackendoff, R. (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. 5

15. Jackendoff, R. and Landau, B. (1992). Spatial Language and Spatial Cognition, in Languages of the Mind: Essays on Mental Representation. pp. 99–124. The MIT Press. 5, 6

16. Jolicoeur, P., M. A. Gluck, and S. M. Kosslyn (1984). Picture and Names: Making the connection. Cognitive Psychology 16, pp. 243–275. MIT Press. 4

17. Mann, W. C. & Thompson, S. A. (1988). "Rhetorical Structure Theory: Towards a functional theory of text organization", Text 8(3), pp. 243–281. 1

18. Marr, D. (1982). Vision. San Francisco: Freeman. 4

19. Moore, J. 1995. Participating in Explanatory Dialogues: interpreting and responding to questions. A Bradford Book, The MIT Press, Cambridge. 1

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

21. Pineda, L. A. and Garza, G. (2002). A Model for Multimodal Reference Resolution. Computational Linguistics 26(2), pp. 139–193. 2, 7, 9

22. Schirra, Jörg R.J. (1993). A Contribution to Reference Semantics of Spatial. Prepositions: The Visualization Problem and its Solution in VITRA, in Zelinsky-Wibbelt, Cornelia, Eds. The Semantics of Prepositions – From Mental Processing to Natural Language. Processing, pp. 471–515. Mouton de Gruyter. 5, 6

23. Power, R. and Scott, D. (1999). Multimodal Authoring using Feedback Texts. In Proc. of COLING/ACL conference, Montreal. 2

24. Tarr, M. J. and Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine, in Tarr and Bülthoff (eds.) Object Recognition in Man, Monkey and Machine, MIT Press, Cambridege, Mass. 5

25. 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, pp. 549–586. Cambridge, Mass.: MIT Press. 6

26. Wahlster, W., André, E., Finkler, W. and Rist, T. (1993). Plan-based integration of natural language and graphics generation, Artificial Intelligence 63, pp. 387–427, Elsevier.