A Model for Multimodal Reference Resolution

Luis. A. Pineda
Institute for Electrical Research
Unit of Informatic Systems
AP 1-475, Cuernavaca, Mor., México
luis@sgi.iie.org.mx

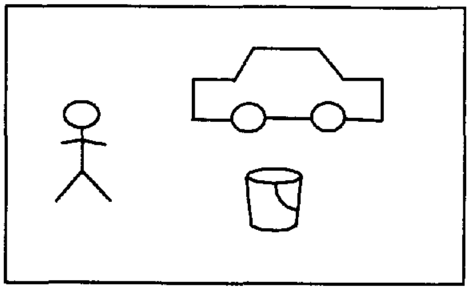
E. Gabriela Garza
Institute for Electrical Research
Unit of Informatic Systems
AP 1-475, Cuernavaca, Mor., México
ggarza@sgi.iie.org.mx

Abstract

In this paper a discussion on multimodal referent resolution is presented. The discussion is centered on the analysis of how the referent of an expression in one modality can be found whenever the contextual information required for carrying on such an inference is expressed in one or more different modalities. In particular, a model for identifying the referent of a graphical expression when the relevant contextual information is expressed through natural language is presented. The model is also applied to the reciprocal problem of identifying the referent of a linguistic expression whenever a graphical context is given. In Section 1 of this paper the notion of modality in terms of which the theory is developed is presented. The discussion is motivated with a case of study in multimodal reference resolution. In Section 2 a theory for multimodal representation along the lines of Montague's semiotic programme is presented. In Section 3, an incremental model for multimodal reference resolution is illustrated. In Section 4 a brief discussion of how the theory could be extended to handle multimodal discourse is advanced. Finally, in the conclusion of the paper, a reflexion on the relation between spacial deixis and anaphora is advanced.

1 Reference and Multimodality

Consider Figure 1 (adapted from an example presented by Thomas Rist in the past workshop on IMMPS at ECAI 96) in which a message is expressed through two different modalities, namely text and graphics.



"He washed it"

Figure 1

The figure illustrates a kind of reasoning required to understand multimodal presentations: in order to make sense of the message, the interpreter must realize what individuals are referred to by the pronouns *he* and *it* in the text. For the sake of argument, it is assumed that the graphical symbols in the figure are understood directly in terms of a graphical lexicon, in the same way that the words *he*, *it* and *washed* are understood in terms of the textual lexicon. It can easily be seen that given the graphical context *he* should resolve to the man, and *it* should resolve to the car. However, this inference is not a valid deduction since the information inferred to is not contained in the overt graphical context and the meaning of the words involved.

One way to look at this problem is as a case of anaphoric inference. Consider that the information provided by graphical means can be expressed also through the following piece of dicourse: *There is a man, a car and a bucket. He washed it.* With Kamp's discourse representation theory (DRT) (Kamp 1981, Kamp et. al. 1993) a discourse representation structure (DRS) in which the reference to the pronoun *he* is constrained to be the man can be built. However, the pronoun *it* has two possible antecedents, and for selecting the appropriate one, conceptual knowledge is required. In particular, the knowledge that a man can wash objects with water, and that water is carried on in buckets must be employed. If these concepts are included in the interpretation context like DRT conditions (which should be retrieved from memory rather than from the normal flow of discourse), the anaphora can be solved. In terms of this analogy, situations like the one illustrated in Figure 1 have been labeled as problems of anaphor with pictorial antecedent in which the interpretation context is built not from a preceeding text but from a graphical representation which is introduced with the text (André et al., 1994).

Consider now the reciprocal situation shown in Figure 2 (adapted from Rist as above) in which a drawing is interpreted as a map thanks to the preceeding text. The dots and lines of the drawing, and their properties, do not have an interpretation and the picture in itself is meaningless. However, given the context introduced by the text, and also considering the common sense knowledge that Paris is a city of France, and Frankfurt a city of Germany, and that Germany lies to the east of France (to the right), it is possible to infer that the denotations of the dots to the left, middle and right of the picture are Paris, Saarbrücken and Frankfurt, respectively, and that the dashed lines denote borders of countries, and in particular, the lower segment denotes the border between France and Germany. In this example, graphical symbols can be thought of as "variables" of the graphical representation or "graphical pronouns" that can be resolved in terms of the textual antecedent. Here again, the inference is not a valid deduction as the graphical symbols could be given other interpretations or non at all.

"Saarbrücken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt."

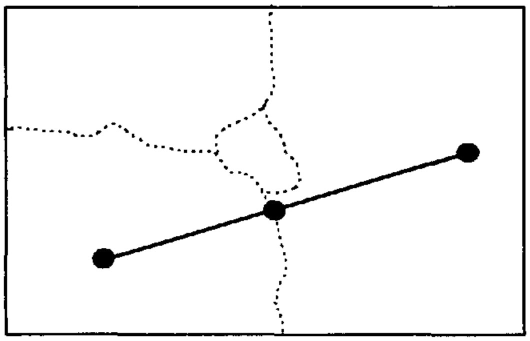


Figure 2

The situation in Figure 2 has been characterised as an instance of a pictorial anaphor with linguistic antecedent, and further related examples can be found in (André et al., 1994). This situation, however, cannot be modeled that easily in terms of Kamp's DRT because the "pronouns" are not linguistic objects, and there is not a straight forward way to express in a the discourse representation structure that a dot representing "a variable" in the graphical domain has the same denotation as a natural language name or description introduced from text in a DRS. Furthermore, consider that the situation in Figure 1 can be thought of as anaphoric only if we ignore the modality of the graphics, as was done above, but if the notion of modality is to be considered at all in the analysis, then the situation in Figure 1 poses the same kind of problems as the one in Figure 2. In general, graphical objects, functioning as constant terms or as variables, introduced as antecendents or as pronouns, cannot be expressed in a DRT, as the rules constructing these structures (the so-called DRS-construction rules) are triggered by specific syntactic configurations of the natural language in which the information is expressed.

An alternative view on this kind of problems consists in looking at them in terms of the traditional linguistic notion of deixis (Lyons, 1968). This notion has to do with the orientational features of language which are relative to the spatio-temporal situation of an utterance. In this regard, and in connexion with the notion of graphical anaphor discussed above, it is possible to mention the deictic category of demonstrative pronouns: words like *this* and *that* which permit us to make reference to extralinguistic objects with the help of pointing acts. Ambiguity of this kind of words is not unusual, as they function not only as deictic or demonstrative pronouns but also as anaphoric, if they are preceeded by a linguistic context, and even as determiners with a deictic component as in expressions like *this car*.

Consider, for instance, that the text in Figure 1 could be substituted by the expression *this washed this* where the first *this* is supported by pointing to the man and the second by pointing to the car. So, in this line of discussion, it is possible to think of the pronoun *it* in Figure 1 as a deictic pronuon which refers to the world, although in an indirect manner through a symbol of a different modality. More generally and according to Kamp (Kamp, 1981):

> ...deictic and anaphoric pronouns select their referents from certain sets of antecedently available entities. The two pronouns uses differ with regard to the

> nature of these sets. In the case of a deictic pronoun the set contains entities that belong to the real world, whereas the selection set for an anaphoric pronoun is made up of constituents of the representation that has been constructed in response to antecent discourse.

For the purpose of this discussion, it is interesting to question what the nature of the sets mentioned above can be. In normal deictic situations the use of a demonstrative pronoun is accompanied by a pointing act to an object that can be perceived directly through the visual modality, and as a result of such a visual interpretation process, the object is represented internally by the subject, however, not necessarily through a linguistic representation, but in a representation of a different modality. According to this, the notion of modality is a representational notion, and not a sensory notion, as it is normally considered in psychological discussion. In the former sense, a representation is a set of expressions of a formal language; that is to say, with a lexicon and with well-defined syntactic and semantic structures. The interpretation conventions of an expression of a given modality are determined by the interpreter of the language. Reasoning with information expressed in both of the modalities is achieved with the help of a translation relation that is similar to the relation of translation between natural languages.

This view of multimodal representation and reasoning can be formalized in terms of Montague's general semiotic programme (Dowty, 1985). Each modality in the system can be captured through a particular language, and relations between expressions of different modalities can be modeled in terms of translation functions from basic and composite expressions of the source modality into expressions of the object modality. In a system of this kind, interpreting examples in Figures 1 and 2 in relation to the linguistic modality consists in interpreting the information expressed through natural language directly when enough information is available, and completing the interpretation process by means of translating expressions of other modalities into the linguistic one. Consider Figure 3 —following (Pineda, 1989, 1996) and (Santana et al., 1997)— in which a multimodal representational system for linguistic and graphical modalties is illustrated.

The circles labeled **G** and **L** in Figure 3 represent the sets of expressions of the graphical and natural languages respectively; the functions ρ_L and ρ_G stand for the translation mappings between the two languages. The circle labeled with **P** represents the set of graphical symbols consituting the graphical modality proper (i.e the actual symbols on a piece of paper or on the screen). Note that two sets of expressions are considered for the graphical modality: the expressions in **G** belong to the formal language in which graphics are represented and reasoned about, and are thought of as the "form" of the overt graphical symbols whose "substance" (in the Saussurean sense) contained in **P** cannot be manipulated directly. The set **W** stands for the world and together with the functions $\mathbf{F_P}$, $\mathbf{F_G}$, $\mathbf{F_L}$ constitutes a multimodal system of interpretation. The order pair $<W, F_L>$ defines the model $\mathbf{M_L}$ for the natural language, and the order pair $<W, F_G \circ F_P>$ defines the model M_G for the graphical language.

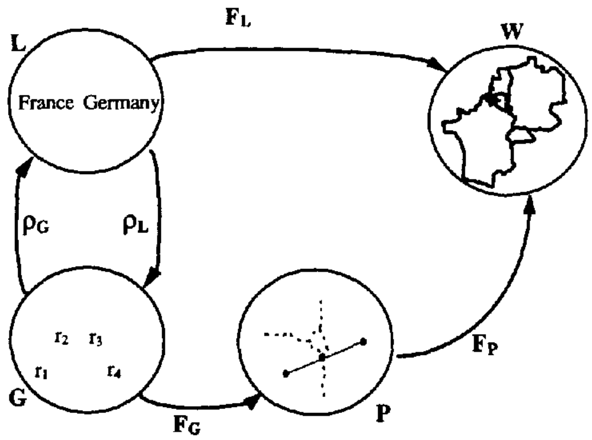


Figure 3

In order to illustrate how this multimodal system of interpretation works consider, for instance, that the denotations of the picture of a man and the word *he* in Figure 1 is the same individual in the world; in the same way, the denotations of the word *Saarbrücken* and the dot on the intersection between the straight line and the lower segment of curve representing the border between France and Germany in Figure 2 are also the same, which is the city of Saarbrücken itself. So, if one asks *who is he?* looking at Figure 1, the answer is found by computing $\rho_L(he)$ whose value is the picture of the man on the figure. Once this computation is performed the picture can be highlighted or signaled by other graphical means. If one points out the middle dot in Figure 2 at the time the question *what is this?* is asked, on the other hand, the answer is found by applying the function ρ_G to the pointed dot, whose value would be the word *Saarbrücken*.

It should be clear that if all theoretical elements illustrated in Figure 3 are given, questions about multimodal scenarios can be answered through the interpretation process; that is to say, evaluating expressions of a given modality in terms of the interpreters of the languages involved and the translation functions.

However, when one is instructed to interpret a multimodal message, like Figures 1 and 2, not all information in the scheme of Figure 3 is available. In particular, the translation functions ρ_L and ρ_G are not known, and the crucial inference of the interpretation process has as its goal to induce these functions. Such an inference can be thought of as the same process that the one involved in solving the so-called linguistic anaphor with pictorial antecedent and the pictorial anaphor with linguistic antecedent. It would also be equivalent to finding out deictic references if "the visual world" is thought of as represented through expressions of the graphical representation modality. These are three different ways of looking at the same problem.

It is important to highlight that in order to induce ρ_L and ρ_G the information overtly provided in the multimodal message is usually not enough. As will be discussed below in this paper, such a process will also require to consider the grammatical structrure of the languages involved, the definition of translations rules between languages, and conceptual knowledge stored in memory about the interpretation domain.

Another consequence of the scheme in Figure 3 is that it provides the basis for generating referring expressions of a given modality in terms of information provided in other modalities. Consider that basic constants or composite expressions of the languages **G** and **L** can be translated to basic or composite expressions of the other language, depending on the definition of the translation function. So, if ones needs to refer linguistically to a graphical configuration, for instance, it would only be required to find an expression of **G** which expresses all graphical attributes of the desired object in the most simple fashion, and then translate it to its corresponding expression in **L**. The resulting natural language expression could be used directly or embbeded in a larger natural language expressions containing words that refer to abstract objects or properties. To illustrate this point consider the natural language text *Saarbrüken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt.* This sentence contains the definite description *the intersection between the border between France and Germany and a line from Paris to Frankfurt,* which in turns contains a number of simplier (basic and composite) referring expressions. Finding the graphical referent of these expressions requires the identification of dots, lines and curves (and parts of curves) in the map that have the same referent. However, the map in Figure 2 has graphical entities that have an interpretation but are not named in the text (consider Figure 4 in which the graphical entities of Figure 2 have been labeled). For instance, Belgium is represented by region r_4, and the curve r_6 represents the border between France and Belgium. Once the picture has been interpreted one would be entitle to ask not only for graphical objects that have been named, but also for any meaningful graphical object. So, if one points to the curve c_6 in Figure 2, one answer provided could be *The border between France and Belgium.* As some graphical objects named by constants of the graphical language do not have a proper natural language name, the translation function ρ_G must associate a basic constant of **G** with a composite description of **L**. The process of inducing such a translation function produces the corresponding referring expressions too.

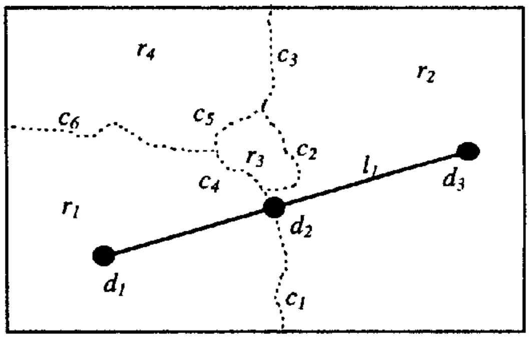


Figure 4

In the rest of this paper, some preliminar results of how this programme can be carried out are presented. In Section 2, a formalization of the languages **L** and **G** with their corresponding translation functions and semantic interpretation, along the lines of Montague's general semiotic programme, is presented. In this section the process of multimodal interpretation and reasoning is explained, and the translation of expressions of one modality in terms of the other is illustrated. However, such a process can be carried out only if the translation functions are known and, as was mentioned above, that is not normally the case. In Section 3, some initial results on how such functions can be induced in terms of the multimodal message, constraints on the interpretation conventions of the modalities, and constraints on the general knowledge about the domain, are presented. In this section the process of generating graphical and linguistic referring expressions, which is associated with the induction of the translation functions, is also illustrated. In Section 4 a preliminar discussion on the feasibility of extending Kamp's DRS with multimodal structures is presented. Finally, in the conclusion a tentative reflexion on the relation of anaphora and spacial deixis on the light of such a kind of theory is advanced.

2 A Multimodal Interpretation System

In this section the definition of the syntax and semantics of the languages **L** and **G** to express the multimodal message of Figure 2 is presented. The language **L** is designed to produce expressions useful to refer to objects, properties and relations commonly found in discourse about maps. In particular, the natural language expressions of Figure 2 can be constructed in a compositional fashion within **L**.The language **G**, on the other hand, is expressive enough to refer to geometrical objects, properties and relations found in drawings. The definitions of **L** and **G** follow closely the general guidelines of Montague's semiotic programme. As a first step of the syntactic definition the set of categories or types is stated. For each type of a language a corresponding type in the other language is defined. Basic constants of the source language can be mapped either to basic or composite expressions of the corresponding type in the object language; in a similar fashion a composite expression of the source language can be mapped into a basic or composite expression of the object language. A number of basic constants for each of these types is defined and the combination rules for producing composite expressions are stated. Associated to each syntactic rule a translation rule mapping the expression formed by the rule to its translation to the other language is defined. In the same way that the interpretation of the natural language expressions in the PTQ system (Dowty, 1985) is given indirectly through the translation to intensional logic, which has a model-theoretic semantic interpretation, the interpretation of expressions of **L** is given indirectly through its translation to expressions of **G**, as shown in Figure 3. The interpretation of expressions of **G**, in turn, is explicitly given through the model $\mathbf{M}_G$. The interpretation function $\mathbf{F}_L$ states the normal meaning for English words, and $\mathbf{F}_P$ is determined by transitivity once the translation function between **G** and **L** is defined, and no further formalization for $\mathbf{F}_L$ and $\mathbf{F}_P$ is presented in this paper. Another simplifying assumption rests on the consideration that the interpretations of all expressions included in these languages depend only on the current graphical state and no intensional types are included in the definition of **L** and **G**. However, this analysis can be extended on the lines of intensional logic if to deal with a more comprehensive fragment of English is required. In Section 2.1 the definition of **L** is presented and in Section 2.2 the language **G** and its interpretation function are formally defined.

2.1 Definition of the Language L

The language **L** is designed to produce expressions like *Saarbrücken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt* in a compositional fashion.This means that all basic constants like *France* and *Germany*, and also all subexpressions of the former sentence, like *the border between France and Germany* or *a line from Paris to Frankfurt* can also be produced. In addition, language **L** can produce expressions like *France is a country*, *Frankfurt is a city of Germany* or *Germany is to the east of France* which express common sense knowledge required in the interpretation of maps. Next, the definition of **L** is presented.

The set of syntactic categories of **L** is as follows:

1. The basic syntactic categories of **L** are t, IV and CN, where t is the category of sentences, IV is the category of intransitive verbs and CN is the category of common nouns.

2 If A and B are syntactic categories then A/B is a category.

Traditional syntactic categories of natural language like transitive verbs (TV), terms (T), propositional phrases (PP) and determiners (T/CN) can be derived from the basic categories.

For each syntactic category of **L** there is a corresponding type in **G**. The correspondence between linguistic categories and geometrical types resembles the translation from English to Intensional Logic (Dowty, 1985) and it is defined in terms of the function f as follows:

1. $f(t) = t$.
2. $f(CN) = f(IV) = <e, t>$.
3. For any categories A and B, $f(A/B) = <f(B), f(A)>$.

The following table illustrates the basic constants of **L** with their category names, category definition and the corresponding type in the graphical language:

Basic constant	Category name	Category definition	Corresponding type in G
Paris, Frankfurt, Saarbrücken, France, Germany	*T*	*t/IV*	<<*e*, *t*>, *t*>
city, country, border, line, intersection, east	*CN*	*CN*	<*e*, *t*>
be, lie at, be to	*TV*	*IV/(t/IV)*	<<<*e*, *t*>, *t*>, <*e*, *t*>>
a, the	*T/CN*	*(t/IV)/CN*	<<*e*, *t*,>, <<*e*, *t*>, *t*>>
	PP	*CN/CN*	<<*e*, *t*>, <*e*, *t*>>
	IV	*IV*	<*e*, *t*>

Figure 5

As can be seen in Figure 5, simple terms like the names of cities and countries translate into characteristic functions of sets of individuals. This graphical type is interpreted as the set of properties (or predicates holding in the interpretation state) that an individual named by the term has. So, as a city is represented through a dot in the graphical domain, the translation of Paris, for instance, is the set of properties that the dot representing Paris has in the intepretation state. Common nouns are translated into graphical predicates: *city* translates into the set of dots representing cities. Transitive verbs are translated into functions taking predicates as their arguments and producing sets of individuals as their values: the verb phrase *be a city*, for instance, translates into a set of dots representing cities. Determiners, prepositional phrases and intransitive verbs function in a similar fashion, although there are no basic constant of the last two categories, as prepositional words are introduced syncategorematically and intransitive verb phrases are always composite expressions in this grammar. In Figure 6 the translation for all basic constants of **L** into **G** is presented. The interpretation of the expressions in the column for **G** are clarified below in this paper when **G** is formally defined.

Next, the syntactic rules of **L** and the translation rules to **G** are presented. Each rule is presented in a box containing the purpose of the rule, the syntactic rule itself with examples of expressions that can be formed with the rule, and finally the translation rule of expressions formed by the rule to their corresponding expressions in **G**. Following Montague, syntactic rules and the syntactic operations for combining symbols (for instance, F_1) associated to each rule are separated. In the following, P_C is the set of expressions of catergory C.

L	**G**
Paris	$\lambda P[P(d_1)]$
Frankfurt	$\lambda P[P(d_3)]$
Saarbrücken	$\lambda P[P(d_2)]$
France	$\lambda P[P(r_1)]$
Germany	$\lambda P[P(r_2)]$
city	*dot*
country	*region*
border	*curve*
line	*line*
intersection	*intersection*
east	*right*
be	$\lambda \mathrm{P}\lambda x \mathrm{P}(\lambda y[x=y])$
lie at	*lie_at*
be to	*in_zone*
a	$\lambda P\lambda Q\exists x[P(x) \wedge Q(x)]$
the	$\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y))]$

Figure 6

SENTENCES

S1$_L$. If $\alpha \in P_T$ and $\delta \in P_{IV}$, then $F_1(\alpha,\delta) \in P_t$, where $F_1(\alpha, \delta) = \alpha\ \delta^*$, y δ^* is the result of replacing the first *verb* inδ by its third person singular present form.

Examples: -Paris is a city of France
-Germany is to the east of France
-Saarbrücken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt

T1$_L$. If $\alpha \in P_T$ and $\delta \in P_{IV}$, and α, δ translate into α', δ', respectively, then $F_1(\alpha,\delta)$ translates into $\alpha'(\delta')$.

TRANSITIVE VERB PHRASES

S2$_L$. If $\delta \in P_{TV}$ and $\beta \in P_T$, then $F_2(\delta, \beta) \in P_{IV}$, where $F_2(\delta, \beta) = \delta\ \beta$.

Examples: -be a city
-be to the east of France

T2$_L$. If $\delta \in P_{TV}$ and $\beta \in P_T$, and δ, β translate into δ', β', respectively, then $F_2(\delta , \beta)$ translates into $\delta'(\beta')$.

TERMS

S3$_L$. If $\delta \in P_{T/CN}$ and $\zeta \in P_{CN}$, then $F_3(\delta, \zeta) \in P_T$, where $F_3(\delta, \zeta) = \delta^*\ \zeta$, and δ^* is δ except in the case where δ is *a* and the first word in ζ begins with a vowel; here, δ^* is *an*.

Examples: -the border
-the border between France and Germany
-a city
-a city of France
-a line from Paris to Frankfurt
-the east of France

T3$_L$. If $\delta \in P_{T/CN}$ and $\zeta \in P_{CN}$, and δ, ζ traslate into δ', ζ', respectively, then $F_3(\delta, \zeta)$ translates into $\delta'(\zeta')$.

COMMON NOUNS

S4$_L$. If $\beta \in P_{PP}$ and $\delta \in P_{CN}$, then $F_2(\delta, \beta) \in P_{CN}$.

Examples: -city of France
-east of France
-border between France and Germany
-line from Paris to Frankfurt
-intersection between the border between France and Germany and a line from Paris to Frankfurt

T4$_L$. If $\beta \in P_{PP}$ and $\delta \in P_{CN}$, and δ, β translate into δ', β', respectively , then $F_2(\delta, \beta)$ translates into $\beta'(\delta')$.

of PREPOSITIONAL PHRASES

S5$_L$. If $\alpha \in P_T$, then $F_4(\alpha) \in P_{PP}$, where $F_4(\alpha) = of\ \alpha$.

Example: of France

T5$_L$. If $\alpha \in P_T$, and α translates into α', then $F_4(\alpha)$ translates into $of^*(\alpha')$ where of^* is a short-hand definition for either of the expression (1) or (2) of **G**:

(1) $\lambda x_{<<e,t>,t>}\ \lambda y_{<e,t>}\ \lambda z_e[y(z) \wedge inside(x)(z)]$
(2) $\lambda x_{<<e,t>,t>}\ \lambda y_{<e,t>}\ \lambda z_e[zone\ (x)(y)(z)]$.

For instance, $of^*(\alpha)$ is obtained by applying of^* to α' as follows:
$\lambda x_{<<e,t>,t>}\ \lambda y_{<e,t>}\ \lambda z_e[y(z) \wedge inside(x)(z)]\ (\alpha')$

which can be reduced to
$\lambda y_{<e,t>}\ \lambda z_e[y(z) \wedge inside(\alpha')(z)]$.

Although *of* has been introduced syncategorematically in **L** for simplicity, it could have been defined as a basic constant of some category of **L** and its translation to **G** would have been a composite expression of some graphical type.

between PREPOSITIONAL PHRASES

S6$_L$. If α, $\beta \in P_T$, then $F_5(\alpha, \beta) \in P_{PP}$, where $F_5(\alpha, \beta) = between\ \alpha\ and\ \beta$.

Examples: -between France and Germany
-between the border between France and Germany and a line from Paris to Frankfurt

T6$_L$. If α, $\beta \in P_T$, and α, β translate into α', β', respectively, then $F_5(\alpha, \beta)$ translates into $between^*(\alpha')(\beta')$ where $between^*$ is a short-hand definition for either of the expression (1) or (2) of **G**:

(1) $\lambda x_{<<e,t>,t>}\ \lambda y_{<<e,t>,t>}\ \lambda z_{<e,t>}\ \lambda u_e\ [z(u) \wedge curve\_between(x)(y)(u)]$
(2) $\lambda x_{<<e,t>,t>}\ \lambda y_{<<e,t>,t>}\ \lambda z_{<e,t>}\ \lambda u_e\ [z(u) \wedge intersection\_between(x)(y)(u)]$.

from-to PREPOSITIONAL PHRASES

S7$_L$. If α, $\beta \in P_T$, then $F_6(\alpha, \beta) \in P_{PP}$, where $F_6(\alpha, \beta) = from\ \alpha\ to\ \beta$.

Example: from Paris to Frankfurt

T7$_L$. If α, $\beta \in P_T$, and α, β translate into α', β', respectively, then $F_6(\alpha, \beta)$ translates into $from\_to^*(\alpha')(\beta')$ where $from\_to^*$ is a short-hand definition for the following expression of **G**:

$\lambda x_{<<e,t>,t>}\ \lambda y_{<<e,t>,t>}\ \lambda z_{<e,t>}\ \lambda u_e\ [z(u) \wedge line\_from\_to(x)(y)(u)]$

2.2 Definition of the Language G

In this section, the syntax and semantics or the graphical language are formally defined, as well as the rules for translating graphical expressions back into natural language. The types of language **G** are defined as follows:

(1) *e* is a type (graphical objects).

(2) *t* is a type (truth values).

(3) If *a* and *b* are any types, then <*a*, *b*> is a type.

(4) Nothing else is a type.

Let V_s be the set of variables of type *s*, C_s the set of basic constants of type *s*, and E_s the set of well-formed expressions of graphical type *s*. The basic constants are presented in Figure 7.

Basic constant	Type	Corresponding category in **L**
d_1, d_2, d_3, r_1, r_2, r_3, r_4, c_1, c_2, c_3, c_4, c_5, c_6, l_1	*e*	*none*
dot, region, curve, line, intersection, right	*<e, t>*	*CN*
lie_at, in_zone	*<<<e, t>, t>, <e, t>>*	*TV*
=	*<e, <e, t>>*	*none*
∧, ↔	*<t, <t, t>>*	*none*
inside	*<<<e, t>, t>, <e, t>>*	*TV*
curve_between	*<<<e, t>, t>, <<<e, t>, t>, <e, t>>>*	*none*
intersection_between	*<<<e, t>, t>, <<<e, t>, t>, <e, t>>>*	*none*
line_from_to	*<<<e, t>, t>, <<<e, t>, t>, <e, t>>>*	*none*
zone	*<<<e, t>, t>, <<e, t>, e>>*	*none*
*of**	*<<<e, t>, t>, <<e, t>, <e, t>>>*	*none*
between, from_to**	*<<<e, t>, t>, <<<e, t>, t>, <<e, t>, <e, t>>>>*	*none*

Figure 7

The symbols *of**, *between** and *from_to** are abbreviations for the corresponding expressions in G as mentioned above in the definition of rules $T5_L$, $T6_L$ and $T7_L$.

In the same way that the translation of basic constants of **L** into **G** where given with the purpose to understand the translation rules $T1_L$ to $T7_L$, in Figure 8 and 9 the reciprocal translations are given.

G	L
dot	*city*
region	*country*
curve	*border*
line	*line*
intersection	*intersection*
right	*East*
lie_at	*lie at*
in_zone	*be to*

Figure 8

Note that constants of **G** in Figure 8 translate into basic constants of **L**; however, the translation shown in Figure 9 are more complex as composite expressions of **G** can translate into basic or composite expressions of **L**. Consider that expressions 1 to 6 of **G** in Figure 9 represent the graphical objects d_1, d_2, d_3, r_1, r_2 and c_1 in Figure 4. The reason to represent these objects with a type-rised is that the language **G** is designed to allow quantification over the graphical domain. So, instead of referring directly to the dot representing Paris, the corresponding expression denotes the set of geometrical properties that the dot representing Paris has. According to this, expression 1 (of **G**) in Figure 9 denotes the set containing all sets of dots of the drawing in which d_1 is included; thus, if *P* is the set of all dots representing cities, d_1 is included in *P* (that is to say, *P* is a property of d_1), but if *P* is the set of all dots representing cities in Germany then d_1 is not included in *P*.

	G	L
1.	$\lambda P[P(d_1)]$	*Paris*
2.	$\lambda P[P(d_2)]$	*Frankfurt*
3.	$\lambda P[P(d_3)]$	*Saarbrücken*
4.	$\lambda P[P(r_1)]$	*France*
5.	$\lambda P[P(r_2)]$	*Germany*
6.	$\lambda P[P(c_1)]$	*the border between France and Germany*
7.	$\lambda P\lambda x P(\lambda y[x=y])$	*be*
8.	$\lambda P\lambda Q\exists x[P(x) \wedge Q(x)]$	*a*
9.	$\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$	*the*

Figure 9

The syntactic definition of **G** is as follows.

1. If $\alpha \in C_s$, then $\alpha \in E_s$.
2. If $\mu \in V_s$, then $\mu \in E_s$.
3. If $\alpha \in E_{<a,\, b>}$ y $\beta \in E_a$, then $\alpha(\beta) \in E_b$.
4. If $\alpha \in E_a$ y $u \in V_b$, then $\lambda u[\alpha] \in E_{<a,\, b>}$.
5. If $\lambda u[\alpha] \in E_{<a,\, b>}$, and $\beta \in E_a$, then $\lambda u[\alpha](\beta) \in E_b$.
6. If $\mu \in V_s$ and $\beta \in E_t$ then $\exists\mu(\beta) \in E_t$.
7. If $\mu \in V_s$ and $\beta \in E_t$ then $\forall\mu(\beta) \in E_t$.

Note that all expressions of **L** can be translated into **G**; however, **G** is a very expressive language and only a subset of well-formed expressions of **G** has translation into **L**. The definition of this last subset the format used for introducing **L** is also used.

SENTENCES

S1$_G$. If $\alpha \in E_{<<e,\, t>,\, t>}$ y $\delta \in V_{<e,\, t>}$, then $\alpha(\delta) \in E_t$.

Examples:

- $\lambda P[P(d_1)]\ (\lambda P\lambda x P(\lambda y[x=y])\ (\lambda P\lambda Q\exists z[P(z) \wedge Q(z)]\ (of^*(\lambda P[P(r_1)])(dot))))$
- $\lambda P[P(r_2)]\ (in\_zone\ (\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]\ (of^*(\lambda P[P(r_1)])(right))))$
- $\lambda P[P(d_2)]\ (lie\_at$
  $(\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$
  $(between^*(\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$
  $(between^*(\lambda P[P(r_1)])\ (\lambda P[P(r_2)])(curve)))$
  $(\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](from\_to^*(\lambda P[P(d_1)])\ (\lambda P[P(d_3)])(line)))$
  $(intersection)\ )\ )\ )$

T1$_G$. If $\alpha \in E_{<<e,\, t>,\, t>}$ and $\delta \in V_{<e,\, t>}$, and α, δ translate into α', δ' in **L**, respectively, then, $\alpha(\delta)$ translates into α' δ'', where δ'' is the result of replacing the firt *verb* in δ' for its third person singular present form.

Translation of the Examples:

- *Paris is a city of France*
- *Germany is to the east of France*
- *Saarbrücken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt*

TRANSITIVE VERB PHRASES

S2$_G$. If $\delta \in E_{<<<e,t>,t>,<e,t>>}$ and $\beta \in E_{<<e,t>,t>}$ then $\delta(\beta) \in E_{<e,t>}$.

Examples:

- $\lambda P\lambda x P(\lambda y[x=y])\ (\lambda P\lambda Q\exists x[P(x) \wedge Q(x)]\ (dot))$
- $in\_zone\ (\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]\ (of^{*}(\lambda P[P(r_1)])(right)))$

T2$_G$. If $\delta \in E_{<<<e,t>,t>,<e,t>>}$ and $\beta \in E_{<<e,t>,t>}$, and δ, β translate into δ',β', respectively, then, δ'(β') translates into δ' β'.

Translation of the Examples:

-be a city
-be to the east of France

TERMS

S3$_G$. If $\delta \in E_{<<e,t,>,<<e,t>,t>>>}$ and $\zeta \in E_{<e,t>}$, then $\delta(\zeta) \in E_{<<e,t>,t>}$.

Examples:

- $\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)](curve)$
- $\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)](between^{*}(\lambda P[P(r_1)])(\lambda P[P(r_2)])(curve))$
- $\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](dot)$
- $\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](of^{*}(\lambda P[P(r_1)])(dot))$
- $\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](from\_to^{*}(\lambda P[P(d_1)])(\lambda P[P(d_3)])(line))$
- $\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)](of^{*}(\lambda P[P(r_1)])(right))$

T3$_G$. If $\delta \in E_{<<e,t,>,<<e,t>,t>>>}$, $\zeta \in E_{<e,t>}$, and δ, ζ translate into δ', ζ', respectively, then δ(ζ) translates into δ'' ζ', where δ'' is δ' except in the case where δ' is *a* and the first word in ζ begins with a vowel; here, δ'' is *an*.

Translation of the Examples:

-the border
-the border between France and Germany
-a city
-a city of France
-a line from Paris to Frankfurt
-the east of France

COMMON NOUNS

S4$_G$. If $\beta \in E_{<<e,t>,<e,t>>}$, and $\delta \in E_{<e,t>}$, then $\beta(\delta) \in E_{<e,t>}$.

Examples:

- $of^*(\lambda P[P(r_1)])(dot)$
- $of^*(\lambda P[P(r_1)])(right)$
- $between^*(\lambda P[P(r_1)])\ (\lambda P[P(r_2)])(curve)$
- $from\_to^*(\lambda P[P(d_1)])\ (\lambda P[P(d_3)])(line)$
- $between^*(\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$
  $(between^*(\lambda P[P(r_1)])\ (\lambda P[P(r_2)])(curve)))$
  $(\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](from\_to^*(\lambda P[P(d_1)])\ (\lambda P[P(d_3)])(line)))$
  $(intersection)$

T4$_G$. If $\beta \in E_{<<e,t>,<e,t>>}$, and $\delta \in E_{<e,t>}$, and β, δ translate into β', δ', respectively, then $\beta(\delta)$ translates into $\delta'\ \beta'$.

Translation of the Examples:

-*city of France*
-*east of France*
-*border between France and Germany*
-*line from Paris to Frankfurt*
-*intersection between the border between France and Germany and a line from Paris to Frankfurt*

of PREPOSITIONAL PHRASES

S5$_G$. If $\alpha \in E_{<<e,t>,t>}$ then $of^*(\alpha) \in E_{<<e,t>,<e,t>>}$.

Example: $of^*(\lambda P[P(r_1)])$

T5$_G$. If $\alpha \in E_{<<e,t>,t>}$, and α translates into α', then $of^*(\alpha)$ translates into *of* α'

Translation of the Example: *of France*

between PREPOSITIONAL PHRASES

S6$_G$. If α, $\beta \in E_{<<e,t>,t>}$ then $between^*(\alpha)(\beta) \in E_{<<e,t>,<e,t>>}$.

Examples:

- $between^*(\lambda P[P(r_1)])(\lambda P[P(r_2)])$
- $between^*(\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]\ (between^*(\lambda P[P(r_1)])(\lambda P[P(r_2)])(curve)))$
  $(\lambda P\lambda Q\exists x[P(x) \wedge Q(x)](from\_to^*(\lambda P[P(d_1)])\ (\lambda P[P(d_3)])(line)))$

T6$_G$. If α, $\beta \in E_{<<e,t>,t>}$ and α, β translate into α', β', respectively, then $between^*(\alpha)(\beta)$ translates into *between* α' *and* β'.

Translation of the Examples:

-*between France and Germany*
-*between the border between France and Germany and a line from Paris to Frankfurt*

from-to PREPOSITIONAL PHRASES

S7$_G$. If α, $\beta \in E_{<<e,t>,t>}$, then *from_to*$^*(\alpha)(\beta) \in E_{<<e,t>,<e,t>>}$.

Example: *from_to*$^*(\lambda P[P(d_1)])\ (\lambda P[P(d_3)])$

T7$_G$. If α, $\beta \in E_{<<e,t>,t>}$ y α, β translate into α', β', respectively, then *from_to*$^*(\alpha)(\beta)$ translates into *from* α' *to* β'.

Translation of the Example: *from Paris to Frankfurt.*

The semantics for the language is given in a model-theoretic fashion as follows.

Let A be the set of graphical individuals A = {d$_1$, d$_2$, d$_3$, r$_1$, r$_2$, r$_3$, r$_4$, c$_1$, c$_2$, c$_3$, c$_4$, c$_5$, c$_6$, l$_1$, right-side, right-r$_1$, right-r$_2$, right-r$_3$, right-r$_4$}, where d$_1$ to l$_1$ are the graphical entities shown in Figure 4, right-side is "the right" and right-r$_i$ is the zone at the right side of region r$_i$. Let D_x be the set of possible denotations for expressions of type *x*, such that D_e = A, D_t = {1, 0}, and, for any types *a* and *b*, $D_{<a,b>} = D_b{}^{D_a}$ (i. e., the set of all functions from D_a to D_b). Let *F* be an interpretation function that assigns to each constant of type *a* a member of D_a. The interpretation (assigned by *F*) of the constants *dot, region, curve, line, right, intersection* are the sets containing the corresponding graphical objects. The interpretation of the constants *lie_at, in_zone, inside, curve_between, intersection_between, line_from_to, zone* (whose types are shown in Figure 6) are geometrical functions. If the arguments of these functions have an appropriate geometrical type (dot, region, curve, etc.) expressions containing these constants can be properly interpreted through a geometrical algorithm; however, if some of the arguments are not of the right kind of geometrical object, then expressions containing these constants have no denotation in **G** and, as a consequence, their translation into **L** lack a denotation too. These conditions can be computed with the help of the type-predicates for geometrical objects in **G**.

Following Montague, the interpretation of variables is defined in terms of an assignment function *g*. It is also adopted the notational convention by which the semantic value or denotation of an expression α with respect to a model *M* and a value assignment *g* is expressed as $[[\alpha]]^{M,g}$.

The semantic rules for interpreting the language **L** are the following:

1. If $\alpha \in C_s$, then $[[\alpha]]^M = F(\alpha)$.
2. If $\mu \in V_s$, then $[[\mu]]^{M,g} = g(\mu)$.
3. If $\alpha \in E_{<a,b>}$, and $\beta \in E_a$, then $[[\alpha(\beta)]]^{M,g} = [[\alpha]]^{M,g}([[\beta]])^{M,g}$
4. If $\alpha \in E_a$ and $u \in V_b$, then $[[\lambda u[\alpha]\ ]]^{M,g}$ is that function *h* from D_b into D_a such that for all objects *k* in D_b, *h*(*k*) is equal to $[[\alpha]]^{M,g}$.
5. If $\alpha \in E_a$, $u \in V_b$, and $\beta \in E_b$, then $[[\lambda u[\alpha](\beta)]]^{M,g}$ is equal to $[[\alpha(u/\beta)]]^{M,g}$, where $\alpha(u/\beta)$ is the result of replacing all ocurrences of *u* for β in α.
6. If $\mu \in V_s$ y $\beta \in E_t$ then $[[\exists\mu(\beta)]]^{M,g} = 1$ iff for some value assignment *g*' such that *g*' is exactly like *g* except possibly for the individual assigned to μ by *g*', $[[\beta]]^{M,g'} = 1$.
7. If $\mu \in V_s$ y $\beta \in E_t$ then $[[\forall\mu(\beta)]]^{M,g} = 1$ iff for every value assignment *g*' such that *g*' is exactly like *g* except possibly for the individual assigned to μ by *g*', $[[\beta]]^{M,g'} = 1$.

With this, the specification of the system of multimodal interpretation presented in Figure 3 is concluded. In this system it is possible to express natural language and graphics and translate expresssion between each other as stated in Section 1. It is also possible to interpret multimodal messages in which part of information is expressed in one modality but some information is carried out in the other modality. One advantage of the system is that a natural language question can be answered by considering the graphics; for instance, if one asks (with a suitable extension to the language) *what is the distance between Paris and Frankfurt?* the answer could be obtained by translating the question into the graphical domain where the distance sought could be computed in terms of the geometry (assuming that the map is drawn at a given scale) and the numerical value could be translated back into natural language. In addition, if reasoning models acting upon representations of each of the modalities were stated, problems could be solved in the modality requiring the lower reasoning effort.

Another advantage of the system is that it permits to rule out natural language expressions which are well-formed syntactically but are, nevertheless, meaningless. The expression *a city of France* is well-formed and has a well-defined reference; however, the expression *a city of Paris*, with the same syntactic form, is not well-defined semantically. This last expression can be rule out as ungrammatical as its translation into the graphical language does not have an interpretation in terms of the geometry. If a condition to the effect that a expression of **L** is grammatical only if its translation into **G** has a well-defined denotation, the graphical domain imposes a kind of selectional restriction which simplifies greatly the syntactic definition of **L**.

At this point, one warning note is in order: the graphical language is probably too expressive and more complex than required. However, having explicit quantification in the graphical domain can pose some interesting questions. Traditionally graphics is considered appropriate for representing concrete situations and natural language is better to express abstractions. However, this is not necessarily the case: consider the example of Figure 1 in which the drawings of a man, a car and a bucket were taken to represent concrete individuals. This is so because *the* man and *the* car were taken to be the antecedents of the anaphoric pronouns *he* and *it*. In a suitable graphical language of the kind developed here a definite reference to such graphical objects can be made as follows:

$\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$ (🯅).

However, some pressupositions are involved in this interpretation choice as the graphical symbols could also be taken to be representing any individual or even the set of all individuals of the kind. In the language **G** all of these readings of the drawings can be expressed as shown in Figure 10.

The multimodal interpretation system developed here does not solve the question of which interpretation should be preferred to, and this question can only be resolved most probably at a pragmatic level; however, the distinction can be made and it should be taken into consideration in a general theory of graphical interpretation. Consider, for instance, whether the drawing of a car on a road sign preventing cars from parking should be interpreted as a particular car, any car or all cars.

Possible Reading	Graphical expression
the man	$\lambda P\lambda Q\exists y[\forall x[P(x) \leftrightarrow x=y] \wedge Q(y)]$ (🯅)
a man	$\lambda P\lambda Q\exists x[P(x) \wedge Q(x)]$ (🯅)
every man	$\lambda P\lambda Q\forall x[P(x) \rightarrow Q(x)]$ (🯅)

Figure 10

The theory also suggests an intriguing path of exploration related to interactive issues. In the same way that natural language expressions can be input directly through the interface, concrete graphical symbols are normally placed on the screen by graphical input devices. The question is, however, whether it is possible and useful to input expressions of **G**, like the ones in Figure 10, directly. At the moment this issue is left for further research.

One last consideration is that the system of multimodal interpretation presented here provides a sound representational scheme to refer in an uniform way to symbols on the screen and to the objects in the world that they represent. In particular, the ambiguity of natural language expressions making interwoven reference to objects on the screen and their interpretation can be placed in a clear representational setting.

3 Incremental Interpretation

In the theory developed in Section 2 it was assumed that the translation of the basic constants of all categories from **L** to **G** and vice versa were known, and then multimodal interpretation and reasoning were possible; however, in the interpretation of multimodal messages, natural language and graphics are input from different sources, and working out the meaning of a multimodal message is by no means trivial. As was discussed in Section 1, solving the graphical anaphora, finding out the reference of deictic pronouns and inducing the translation function are related problems that need to be solved for the interpretation of multimodal messages. Consider, for instance, the situation of reading a book with words and pictures: when the associations between text and graphical symbols is realized by the reader, the message as a whole has been properly understood. However, it cannot be expected that such an association can be known beforehand.

Inducing this translation function is similar to the computer vision problem of interpreting drawings. A related antecedent is the work on the logic of depiction (Reiter et al., 1987) in which a logic for the interpretation of maps to be applied in computer vision and intelligent graphics is developed. It is argued that any adequate representation scheme for visual (and computer graphics) knowledge must mantain the distinction between knowledge of the image (the graphics) and knowledge of the scene (its interpretation), and about the depiction relation. In Reiter's system two sets of first order logical sentences representing the scene and the image are employed, and express, respectively, the conceptual and geometrical knowledge about hand drawn sketch maps of geographical regions. The depiction relation corresponds to the translation function between constants of **L** and **G** discussed above. An interpretation in Reiter's system is defined as a model, in the logical sense, of both sets of sentences and the depiction relation, and interpreting a drawing consists in finding out all possible models of such sets of sentences.

Although computing the set of models of a set of first order sentences is computationally untractable problem, the entities constituting a drawing conform, normally, a finite set which is often small. So, the possibility of computing the set of models of a drawings is a matter for empirical research. In particular, Reiter's system employs a constraint satistaction algorithm to find out all possible interpretation of maps, and the output of his system is a set of labels for curves or chains as rivers, roads or shores, and for areas as land regions or water regions. As was mentioned above, to find the translation functions between **G** and **L** is a similar problem with the same kind of complexity.

As a side effect of working out the translation between basic graphical and linguistic constants, a method for generating natural language expressions that refer to graphical objects and configurations is at hand. Consider that a natural language description can have both simple and composite referring expression that translate into basic graphical constants, and inducing the linguistic translation of a graphical term which has not been named is the same as generating a linguistic description for such an object. Next, an algorithm for constructing the translations is illustrated.

As a preliminar consideration it is important to highlight that such translations cannot be built with the overt information expressed through the multimodal message. For working out the interpretation of Figure 2, for instance, it is required, in addition to the text and graphics, knowledge about the geography of Europe and also knowledge about interpretation conventions of maps. To find out the translation such knowledge must be employed.

The conventions about the interpretation of maps are expressed as a correspondence between graphical and conceptual types, for instance, that dots represent cities and regions represent countries. General knowledge about maps, either geometrical or conceptual, will constraint the possible translation.

The algorithm for computing the translation function has two parts; the purpose of the first is to assign a graphical constant to all terms in the overt textual message according to the grammar of **L**, and the second is to assign a referring expression to the remaining graphical constants of the drawing. For the example in Figure 2, the output of the first part is shown in Figure 11.

L	*G*
Paris	d_1
Saarbrücken	d_2
Frankfurt	d_3
France	r_1
Germany	r_2
the border between France and Germany	c_1
a line from Paris to Frankfurt	l_1
the intersection between the border between France and Germany and a line from Paris to Frankfurt	d_2

Figure 11

The second part would assign a description to the remaining graphical constants as shown in Figure 12.

G	*L*
r_3	*a country*
r_4	*a country*
c_2	*a border*
c_3	*a border*
c_4	*a border*
c_5	*a border*
c_6	*a border*

Figure 12

For the definition of the algorithm a function table for representing the set of possible functions from graphical to linguistic constants of the corresponding semantic types is defined. The interpretation conventions for maps are stated through the order pairs in the following set: *I* = {<dot, city>, <region, country>, <curve, border>, <line, line>}. The function table for the first pair in relation to Figure 2 and 4 is shown in Figure 13.

Saarbrücken			
Frankfurt			
Paris			
	d_1	d_2	d_3

Figure 13

As can be seen the function table in Figure 13 relates all dots to all cities in the text. As any dot can represent any city, but different cities are represented by different dots, each dot must be associated to one city by filling the box in which the dot intersects the corresponding city in the function table. Considering that once a dot has been assigned to a city, the row corresponding to that city cannot be filled out for the other dots. According to this, if there are *n* cities, the first dot receiving an interpretation can be assigned in *n* different ways (it can represents one of the *n* cities), the second in *n*-1 different ways, etc. As a consequence, each function map represents *n*! possible interpretation functions (if all cities are represented).

The first step in the algorithm is to identify all graphical and linguistic basic constants from the overt message and draw the function tables for the interpretation conventions set *I*. In our example basic linguisctic constants referring to graphical objects (proper names) name cities and countries, and only a function table for region representing countries (in addition to Figure 13) is considered as shown in Figure 14.

Germany				
France				
	r_1	r_2	r_3	r_4

Figure 14

The next step is to fill out tables in Figures 13 and 14 with all possible interpretations that are consistent with the overt knowledge and also the background knowledge about the interpretation task, in this case knowledge about the geography of Europe. The general knowledge to be considered for this example is shown in Figure 15. Note that clauses 1 to 6 are general knowlege of geography, but clause 7 is introduced explicitly in the multimodal message. Note as well that there might be a considerable amount of knowledge about the geography of Europe which is not included in Figure 15. However, how knowledge is brought about to the interpretation process is beyond the scope of this paper. The only consideration is that the range values of the function tables are the main indices that somehow retrieve the information from memory.

1.	France is a country
2.	Germany is a country
3.	Paris is a city of France
4.	Frankfurt is a city of Germany
5.	Saarbrücken is a city of Germany
6.	Germany is to the east of France
7.	Saarbrücken lies at the intersection between the border between France and Germany and a line from Paris to Frankfurt

Figure 15

For filling out the function tables th set of all constraints should be considered. As shown by Reiter (Reiter et al., 1987) all possible models can be found —for a finite set of graphical symbols— with a constraint satisfaction algorithm. Along this line, we are exploring strategies to find out the set of models incrementally extending the function tables filling out one column of one table at a time by considering one constraint at a time. This is done in a way that the extended model satisfies all constraints considered so far. The process continues until all function tables are filled out.

For the example, propositions 3 to 7 are considered to produce the function tables illustrated in Figure 16. Note, in particular, that if proposition 7 is not considered a diagonal model for the intepretation of dots would also be admissible.

Germany		x		
France	x			
	r_1	r_2	r_3	r_4

Saarbrücken		x	
Frankfurt			x
Paris	x		
	d_1	d_2	d_3

Figure 16

The next step of the algorithm consists in identifying the translations of composite referring expressions to complete table in Figure 11. This can be done with the help of the grammar, the translation functions, the semantic interpretation of **G** and the translations already computed in Figure 16. In fact, once the translation of basic constants is known the translation of composite terms using those constants can be found compositionally in terms of the translation rules from **L** to **G** and the interpretation rules of **G**. Consider, additionally, that the translation for basic constants for other categories is given beforehand as shown in Figure 6. For instance, the translation of the composite expression *the border between France and Germany* into **G** is

$$\lambda P \lambda Q \exists y[\forall x[P(x) \leftrightarrow x{=}y] \wedge Q(y)] \\ (\mathit{between}^{*}\ (\lambda P[P(r_1)])\ (\lambda P[P(r_2)])\ (\mathit{curve}))$$

which can be reduced into

$$\lambda Q \exists y[\forall x[\mathit{between}^{*}\ (\lambda P[P(r_1)])\ (\lambda P[P(r_2)]) \\ (\mathit{curve})\ (x) \leftrightarrow x{=}y] \wedge Q(y)]$$

to produce the final value which is, due to the type rising of terms, the set of properties that c_1 has, and for simplicity we take it to be the constant c_1.

Next, the second part of the algorithm producing the translation of the set *UNNAMED* of graphical constants that have no name, as shown in Figure 12, is described. In order to carry out this process, the first step is to identify the types of all constants in *UNNAMED* with the help of the type predicates of *G*. For each constant it is required to identify all geometrical functions producing objects of the constant type. So, the same constant may be produced by a number of ways depending on the geometrical functions available for producing objects of the constant type. The next step consists in identifying all combinations of expressions that can be used as arguments of geometrical functions, forming in this way a set of expressions that can produce the constant at hand. Each of these expressions is interpreted. From this process two kinds of outputs can be expected: either the expression has a well-defined value or it does not. Expressions having a proper value must be combined with a graphical quantifier, and probably with other expressions of **G** (to produce the *between** term from the geometrical function *curve_between*). The resulting term can be translated back into the natural language, producing in this way the corresponding description.

Suppose it is desired to find a referring expression of the constant c_1 of graphical type *curve*. Considering all geometrical functions denoted by the basic constants of **G**, only *curve* and *curve_between* can produce curves. The constant *curve* denotes the set of curves on the drawing, so the expressions $\lambda P \lambda Q \exists x[P(x) \wedge Q(x)](\mathit{curve})$ —*a border*— and $\lambda P \lambda Q \exists y[\forall x[P(x) \leftrightarrow x{=}y] \wedge Q(y)](\mathit{curve})$ —*the border*— can be formed; the interpretation of the former expression results in the set of properties that one curve or another has, including the properties of curve c_1; however, the intepretation of the last expression results in an empty set as there is more than one curve in the drawing. So, only the former expression generated by the constant *curve* is a possible description of c_1. If the constant *curve_between* is considered, the expressions shown in Figure 16 can be obtained (the expressions denoting empty sets —as $\mathit{between}^{*}(\lambda P[P(r_1)])\ \lambda P[P(r_1)])\ \mathit{curve})$— where omitted). As can be seen, only the first of these expressions refers to the curve denoted by c_1.

Expression of **G**	*Interpretation (set of properties of)*
$\mathit{between}^{*}(\lambda P[P(r_1)])\ (\lambda P[P(r_2)])\ (\mathit{curve})$	c_1
$\mathit{between}^{*}(\lambda P[P(r_1)])\ (\lambda P[P(r_3)])\ (\mathit{curve})$	c_4
$\mathit{between}^{*}(\lambda P[P(r_1)])\ (\lambda P[P(r_4)])\ (\mathit{curve})$	c_6
$\mathit{between}^{*}(\lambda P[P(r_2)])\ (\lambda P[P(r_3)])\ (\mathit{curve})$	c_2
$\mathit{between}^{*}(\lambda P[P(r_2)])\ (\lambda P[P(r_4)])\ (\mathit{curve})$	c_3
$\mathit{between}^{*}(\lambda P[P(r_3)])\ (\lambda P[P(r_4)])\ (\mathit{curve})$	c_5

Figure 16

Thus, the expressions *a border*, *a border between France and Germany* and *the border between France and Germany* are possible descriptions for c_1. The pragmatic choice of which expression is the most appropriate in an interpretation context goes beyond the scope of this paper; here we are only concerned with the computation of the set of expressions that can be correctly produced in terms of the multimodal representation.

This procedure has one additional complication that has to be taken into account. Once a graphical object has been produced by the procedure mentioned above it can also extend the set of the argument combinations of the graphical functions referring to other well-formed graphical objects. Note that it is possible to produce very large expressions and even infinite ones with the unconstrained recursive application of the procedure. It could be possible to produce, for instance, *the border between France and Germany* *between France and Germany between France and Germany*. This expression is well-formed and denotes the border between France and Germany. In order to prevent these kind of very large expressions an algorithm for generating the set of simplest but maximally expressive expressions of a graphical language has been proposed (Santana, 1995).

4 Multimodal Discourse Representation Theory

The ability to interpret individual multimodal messages is a prerrequisite for interpreting sequences of multimodal messages occuring in the normal flow of interactive conversations. In the same sense that discourse theories, like DRT, are designed to interpret sequences of sentences, it is desirable to have a theory in which sequences of multimodal messages can be understood. Such a kind of theory would have to support anaphoric and deictic resolution models in an integrated fashion, and would have to be placed in a larger pragmatic setting in which intentions and presupositions are considered, and in which mechanisms to retrieve knowledge from memory are also taken into account. To work out such a theory is quite an ambitious goal, however, in the same way that DRT focuses in internal structural processes that govern anaphoric resolution, it is plausible to consider a multimodal discourse representation theory (MDRT) to cope with referential aspects of multimodal communication. In the same way that DRT postulates discourse representation structures in which referents and conditions are introduced incrementally through the interpretation of the incoming natural language discourse through the application of construction rules, it is plausible to conceive similar multimodal discourse representation structures (MDRS) whose referents and conditions would be introduced by modality depending construction rules acting upon the expressions of the corresponding modality. The definition of such an extention of DRT is a long-term goal of this work.

A consequence of the notion of modality that has been developed so far is that expressions referring to graphical objects and relations are well-defined in a suitable language, and could be included as referents and conditions in the proposed MDRS. In these structures, DRS-conditions extracted from different modalities would be kept in separate partitions, but the discourse referents would be abstract objects common to the whole MDRS. The formalization of modalities in terms of representation languages would permit to extend DRT, allowing to handle different modalities, as long as the conditions and referents were introduced by construction rules that triggered by specific syntactic configurations of the representation language of the modality in question.

In summary, the definition of this kind of structures would be possible if the following three questions could be answered: how information of different modalities can be incorporated into a MDRS, how discourse referents common to expressions of different modalities can be identified, and lastly, how simplification of conditions involving different modalities can be carried out. The suggestion is that these three problems can be solved in terms of the scheme shown in Figure 3 and Section 2, and the interpretation process illustrated in Section 3. For the moment these issues are left for further work.

5 Conclusions

In this paper a theory of representation and interpretation for multimodal messages, and a model for multimodal reference resolution has been presented. First, it was discussed how this problem can be stated in terms of so-called linguistic anaphor with pictorial antecedents, or pictorial anaphor with linguistic antecedents. It was argued that in more traditional lines such a problem can be thought of, alternativelly, as the resolution of spacial indexical referents. It was also argued that with a representational theory of modality, one in which the notion of modality is captured in terms of a formal language and its interpreter, a third interpretation of the problem of multimodal reference resolution can be given. In this last view, solving multimodal references can be thought of as inducing a translation function between basic constants of the modalities involved. The representation and interpretation machinery for carrying on this third view was formally developed along the lines of Montague's semiotic programme and its associated general theory of translation. It was also illustrated an algorithm for finding out such a translation relation when text and graphics are introduced through independent input channels, and the translation between constants must be induced dynamically. Finally, it was suggested to extend Kamp´s DRT with multimodal discourse structures (MDRS) in order to model the referential aspects of the kind of multimodal discourse that is likely to occur in interactive multimodal conversation. This extension would permit to capture an aspect of spacial deixis which is currently beyond the scope of DRT. If the notion of multimodal discourse representation structures is developed along the lines suggested in this paper, Kamp´s demarcation between anaphoric and deictic uses of pronouns could be formally captured, as the sets of antecedents taken from the world would be incorparted as referents and conditions of a MDRS: while the antecedents for anaphoric pronouns taken from preceeding text are accesible for pronouns, deictic antecedents would be accesible via translation functions.

References

Elisabeth André, Thomas Rist. 1994. Referring to World Objects with Text and Pictures, technical report, German Research Center for Artificial Intelligence (DFKI).

David R. Dowty, Robert E. Wall, Stanley Peters. 1985. *Introduction to Montague Semantics*. D. Reidel Publishing Company, Dordrecht, Holland.

Hans Kamp. 1981. A Theory of Truth and Semantic Representation. *Formal Methods in the Study of Language*, 136 pp. 277-322, Mathematical Centre Tracts.

Hans Kamp, Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publisher, Dordrecht, Holland.

John Lyons. 1968. *Introduction to Theoretical Linguistics*, Cambridge University Press, Cambridge.

Luis Pineda. 1989. Graflog: a Theory of Semantics for Graphics with Applications to Human-Computer Interaction and CAD Systems. PhD thesis, University of Edinburgh, U.K.

Luis Pineda. 1996. Graphical and Linguistic Dialogue for Intelligent Multimodal Systems. In *G. P. Facinti and T. Rist editors, WP32 Proceedings, 12th European Conference on Artificial Intelligence ECAI-96,* Hungary, August. Budapest University of Economic Sciences.

Raymond Reiter, Alan K. Mackworth. 1987. The Logic of Depiction, *Research in Biological and Computational Vision*, University of Toronto.

J. Sergio Santana. 1995. Quantification Issues in a Graphical Language, *Monthly progress report on the IIE/University of Salford in-house PhD Programme*, November.

J. Sergio Santana, Sunil Vadera, Luis Pineda. 1997. The Coordination of Linguistic and Graphical Explanation in th Context of Geometric Problem-solving Tasks, *technical report on the IIE/University of Salford in-house PhD Programme* (submitted to *Computational Linguistics*).