

# Discourse Structure of Text-Graphics Documents

Robert P. Futrelle

College of Computer Science 161CN  
Northeastern Univ., 360 Huntington Ave.  
Boston, MA 02115  
1-617-373-4239  
futrelle@ccs.neu.edu

Anna Rumshisky

College of Computer Science 161CN  
Northeastern Univ., 360 Huntington Ave.  
Boston, MA 02115  
1-617-373-7920  
arum@ccs.neu.edu

## ABSTRACT

In order to analyze, generate and use documents containing both text and graphics, it is important to have a theory of their structure. We argue that it is possible to develop a semantics for graphics, as well as text, and generate an integrated representation of text/graphics discourse, building on previous theories of text discourse. A major component of our theory is an integrated natural language / visual language lexicon that allows people to understand bimodal discourse. Another major component of text/graphics discourse is the *role of the reader*. That is, the reader constantly exercises choices to shift his or her attention between the text and graphics while reading/viewing. We demonstrate some computational approaches to the automation of building discourse structure at the level of syntax. This is followed by the construction of semantics via logical forms. The result of this work is a Text-Graphics Discourse Theory, TGDT.

## Keywords

Discourse theory, documents, reading, parsing, natural language, text, viewing, graphics, diagrams, Diagram Understanding System, DUS, lexicon, semantics, logical form, inference load, Restricted Focus Viewer, RFV, TGDT.

## 1. INTRODUCTION

In order to develop powerful systems for the analysis and generation of multimodal documents, it is useful to have theories that go beyond the individual elements to relate the elements in the larger whole. Such *discourse theories* have been developed for natural language [6,16], but much less has been done for multimodal documents. A theory of multimodal discourse could be applied to analysis or generation; we focus on analysis. The domain for this work is text-graphics documents (TG) such as research papers in science and engineering. The particular aspect of discourse we focus on is arguably the most important, *coreference*. Coreference can operate on an intratext basis, as in anaphora or the use of definite descriptions, within graphics as in the use of identical elements as co-referring items, or as TG coreference such as in descriptive or naming constructs in text that identify aspects of the graphics.

Our work attempts to characterize the elements and relations of TG coreference. Our primary assumption is that readers and writers of such documents employ an internal multi-modal lexicon that relates text and graphics [13]. Without such an assumption, we would be at a loss to explain how people can talk about what they see ("I see a cat.") or how they can identify visually perceived objects, given their verbal descriptions ("Look at the big red chair.")

In this paper we focus on the parsing of text and graphics and the integration of the resulting analysis into a coordinated whole, with coreference resolved. We use machine-based parsers for text and for graphics, then manually build logical forms for the semantics of each, and then integrate the two into a representation at the level of logical forms. Our work complements numerous psychological studies on reading combined text and graphics [10]. It differs from that work in that it builds on previous theories of natural language discourse as well as constructing concrete representations. It also continues earlier work in our group [21].

The resolution of TG coreference allows inferences to be made that would not otherwise be possible. In Figure 1 and its caption, the resolution of TG coreference is a necessary and sufficient condition for inferring, "The large box is heavy."

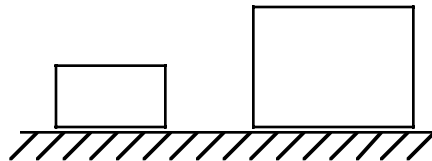


Figure 1. The box on the right is heavy.

## 2. MIXED-MODE DISCOURSE

### 2.1 The Notion of Discourse

Because of its importance, discourse has been studied extensively, focusing on natural language [5,14]. Humans also make extensive use of the visual modality in discourse -- e.g., images in technical articles, books, newspapers, television and lectures. There have been few studies, much less theories, of discourse involving language *and* images.

### 2.2 The Reading/Viewing Sequence Problem

One of the most obvious differences between text and graphics is that there is no well-defined sequential order for "reading" a graphic or for shifting attention between text and graphics during reading. In text discourse the assumption of sequential movement through text is made, with few exceptions. For text, it is common to employ constructs such as pronominal anaphora in which a pronoun is used to refer to a *previous*

item. In spite of the various orders in which a TG document might be read, readers are still able to build an integrated understanding of content. In any theory of TG discourse we must then take seriously the *role of the reader* in shaping the theory. We suggest that representations are built as reading proceeds and are integrated incrementally or later in a "batch" mode. The behavior of readers of text-graphics documents has been examined in extensive psychology experiments [10,12]. But those studies typically do not build detailed representations of the type developed here. An incremental mode appears to be the norm, discovering coreference and building and revising integrated representations while reading. The incremental strategy is presumably chosen because it reduces the inference load on the reader [11].

### 2.3 Observations of Shifts in Attention During the Reading of Text/Graphics Documents

Attention shifts between text and graphics have been documented in eye-tracking studies [12]. We presume that attention shifts are related to specific natural language processes such as new entity introduction, reference resolution for definite noun phrases, etc. In the graphic domain there are similar processes involving the inspection of diagram entities, larger structures, connectivity and other geometrical relations.

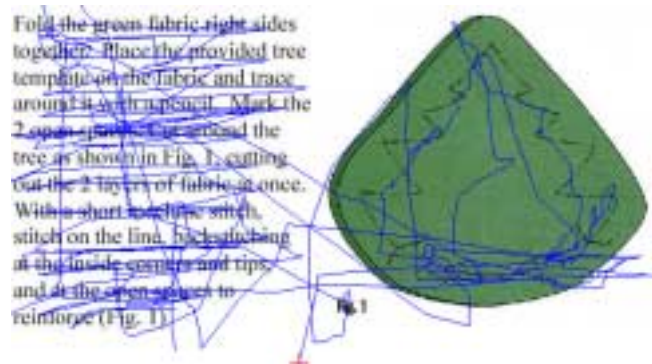
If the reading is for general informational purposes, then the reader presumably will try to choose shifts that minimize the inference load. If the reader has a particular question in mind, then the attention shifts are presumably related to some problem-solving strategy. This should lead to increased attention being paid to the more relevant units of the text and graphics.

In preliminary studies, we have presented subjects with TG material and a question to be answered, and then monitored their attention shifts using the Restricted Focus viewer or RFV [4]. Unlike conventional eye-tracking systems, the RFV presents the material on the screen so that only a small rectangular region of it is in focus, with the material outside of the central region blurred. The position of the in-focus rectangle is controlled by the subject moving the mouse. No additional eye-tracking instrumentation is used. By its very nature, the RFV can never get out of alignment. The RFV allows scripting to control the timing of the material presented as well as logging the subject's "scan-path" and button responses. It can also generate a variety of reports, including replaying the scan path at controlled rates. The RFV is written in Java, and was run under Solaris on a Sun Ultra 10. The RFV, including source code, is freely available at the URL included in the reference [4].

In the trials, each subject was presented with the TG material in Figure 2 after first reading the following statement:

"You will be presented with the first paragraph of an article from the magazine, "Quilt It", describing how to make a cloth Christmas tree. At the end of the experiment, you will be asked to explain how the cloth tree should be stuffed."

We observed that the subjects would return ("regress") to specific portions of the text during a run. This suggests that they develop a *spatial map of the text*. The example above is discussed further in Sec. 7.



**Figure 2.** The cumulative scan-path of the focus of attention of a subject during a problem-solving session employing the Restricted Focus Viewer, RFV. Note that the subject spent a major portion of the time focused on the two noun phrases referring to the "open spaces" and to the portions of the figure near the bottom right where the two open spaces are in the fabric construction. There were five shifts of attention between the text and graphics during this session.

### 2.4 Coreference Items in TG Discourse

Our theory first identifies the elements in text and graphics that can co-refer and then describes how they interact.

In text, the most important types of such elements are:

- *Names* of recognizable graphic objects ("circle", "cat")
- *Characterizations or descriptions* of graphic objects' intrinsic properties ("the red rectangle")
- *Geometric relational descriptions* ("the river to the west of Bigtown")

In graphics, the most important types of such elements are:

- Recognizable objects ("arrow")
- Assignment of intrinsic properties to objects (dashes and width of a line)
- Arranging objects (placing an "axle" circle at the center of a "wheel" circle)

Assuming that the text and graphics are separately analyzed and characterized by the elements above, there are two aspects of the integration process. The first deals with objects and the second with relations.

### 2.5 Integration of Text and Graphics

Integration of text and graphics elements is done to build a coordinated representation of TG discourse. The first level of integration involves lexical items referring to objects. Natural language terms such as "triangle" are related to graphical depictions of triangles. Lexical lookup is non-trivial, because the "language" used in the two modalities may differ. We assume that some type of extended thesaural mechanism is available, e.g., one that could establish coreference between a class of smooth convex closed regions in the graphics domain and the natural language term "oval".

The other major components in integrated representations are relations or predications, for example, a descriptive property such as "red" and its graphical counterpart. A more complex example would be the predicate "above" and its graphical

correlate that would involve the placement of one graphic object at a greater vertical position than another (referred to the normal viewing direction for the graphic). Figure 1 gives an example of a relational predicate in the phrase, "the box on the right" and its corresponding graphic realization.

### 3. THE NATURAL LANGUAGE / VISUAL LANGUAGE LEXICON

#### 3.1 The Need For and Existence Of an Integrated Lexicon

People can describe and often name what they see [13]. They can also visually locate and attend to objects in their view that are described to them or named. So they possess links between their visual and natural language faculties, and presumably some integrated mental modules comprising both modalities. People can reason about visual entities, often combined with reasoning about natural language, e.g., "Do you think my big leather chair would fit through this door?" Using a simple model of these integrated representational faculties, we can construct representations of mixed-mode discourse. To keep our presentation in bounds, we will couch this work in terms of the *lexicons* of language and vision. We will discuss the nature of the two separate lexicons and show how this leads naturally to an integrated lexicon. We will call these the natural language lexicon, NLex, the graphical lexicon, GLex, and the integrated natural language/visual language lexicon, ILex.

One could argue that a graphical entity such as a rectangle is highly ambiguous -- it could have thousands of distinct meanings depending on the context. It appears that a powerful disambiguation procedure is required. We would argue that it is more appropriate to view a rectangle as having little or no meaning beyond its geometric properties. The process of giving specific meaning to such a shape is based on the *context* as understood by the viewer. For example, experience with *shoe store* contexts develops mental representations that contain *shoebox* entities and these in turn can be linked to the otherwise unspecific graphical rectangles in a scene. In this view ambiguity resolution for a complex shape such as a dog proceeds in exactly the same manner.

The task of establishing coreference in a TG discourse has to be based on the viewer's knowledge, specifically by reference to the viewer's integrated lexicon, ILex. We will also describe operations required for establishing coreference that go beyond simple lexical lookup, such as symbolic and visual reasoning. In the discussions that follow, we will use double quotes to denote items in the NLex, "rectangle", underlining/boxes to denote graphical entities in GLex, rectangle, (as well as graphical relations, such as left-of) and a subscript to denote items in the ILex, rectangle<sub>1</sub>. Below, we freely mix entities from the three lexicons to emphasize their equal stature and roles.

#### 3.2 Objects

There are both simple and complex visual objects that are named by natural language tokens and recognized as such. These include the basic shapes such as, square, straight-line<sub>1</sub>, "ellipse", etc. Complex objects that are part of the viewer's world such as arrow, "dog", automobile<sub>1</sub>, etc., are also in both lexicons. Unary properties and descriptors such as yellow, "large", smooth<sub>1</sub>, etc., are also used to describe and identify objects for coreference purposes.

#### 3.3 Object Structure

Complex objects have structure that includes particular shapes as well as components. A violin has a distinctive shape as well as named parts such as the bridge and fingerboard<sub>1</sub>. A viewer with sufficient knowledge of the components may be able to both name and identify them. Less knowledgeable viewers may not be able to name them (purfling<sub>1</sub>, or the treble foot of the bridge) but can identify them when given a description or a graphical deixis element such as a callout (a line or arrow pointing to an object or object part with a text label at the other end of the line).

There are other important aspects of object structure that have no specialized names attached to them but can operate to establish coreference, such as definite descriptions. The unary descriptors mentioned earlier can focus attention on certain parts of an object's structure, e.g., "the *red* components".

#### 3.4 Relations

An important part of the description and recognition processes for graphics are the graphical concepts and natural language terms for *geometrical relations*. This includes entities such as near, "above", etc. They can operate as intra-object and inter-object descriptors. Thus one can talk about the "right-hand side" of a single object or an object to the "right" of another, using the same relative orientation term. Comparative terms denote inter-object relations, such as smaller<sub>1</sub> or brighter.

Descriptions that focus on particular aspects of a picture may go beyond single lexical items, can be arbitrarily complex and may require some reasoning on the viewer's part, e.g., "The house is the red brick one on the left side of the picture with the truck parked in front of it."

#### 3.5 An Integrated Lexicon

The naming and descriptive devices described above demonstrate that the phenomenon of coreference is a rich one. This shows that the mental lexicon that a viewer possesses has a large integrated store of items and constructs that tie together the visual world and the world of natural language. In summary, we have identified some of the classes of entries in the integrated natural language/visual-language lexicon, ILex:

1. Named geometrical forms ("rectangle", rectangle)
2. Unary descriptors ("flat", flat)
3. Named complex objects ("door", door)
4. Named components ("handle", handle)
5. Intra-object relations ("top", top)
6. Inter-object relations ("between", between)

Beyond single lexical items there are a variety of devices available to a viewer for both describing and identifying an object or a portion or component of an object or object collection ("lower left corner"). These have corresponding complex relations in the visual domain (lower°left°corner). The list above integrates the items discussed in Sec. 2.4 into the more concise structure of a lexicon.

The structure of the integrated lexicon, ILex, derives naturally from the discussion above. In the simplest view, the ILex is a

collection of pairs of the form "<natural language term>" and <corresponding°graphical°percept/concept>, e.g.,

rectangle<sub>1</sub> = "rectangle" <----> rectangle.

## 4. SYNTACTIC PARSING OF TEXT AND GRAPHICS

### 4.1 Parsing In General

We approach the parsing of text and graphics in the same way. We describe natural language by a lexicon and a grammar and compute the syntactic structures of sentences using a conventional parser. In the graphics domain, we focus on diagrams made up of discrete components rather than raster images. Figure 1 is an example of this, made up of two rectangles, a long horizontal line and sixteen short oblique lines. We design *context-based constraint grammars* that describe diagrams and then generate syntactic analyses of the diagrams using our Diagram Understanding System parser [8,9]. The relations between the constituents (right-hand-side elements of a production) contain unary and multi-element constraints such as *short* or *near*. The major difference is at the lexical level, since the primitives or leaf elements in a graphics grammar are drawn from a small set of basic geometrical objects such as lines, polygons, and positioned text. This is in contradistinction to natural language which has thousands of tokens with distinct senses.

### 4.2 Natural Language Parsing

Natural language parsing, for the examples discussed in this paper, was done using a straightforward implementation of a top-down chart parser. The parser was designed following the Earley algorithm described in [15], implemented in Java and run under Sun Solaris. We used the standard Penn Treebank POS tagset, and a slight variation of the syntactic tagset. The right-recursive grammars used were designed to ensure maximal proximity to the bracketing in Penn Treebank style [18]. The final parses used were selected from the small number of alternative parses produced by the Earley parser.

A simple example of one of the parses is one for the fabric construction example in Sec. 7, "Mark the 2 open spaces.":

```
(S
  (VP
    (VB Mark)
    (NP (DetP (DT the))
      (NBar (ADJP (CD 2))
        (NBar (ADJP (JJ open))
          (NBar (NNS spaces))))))))
```

### 4.3 Diagram Parsing

It is important to explain some of the challenges facing graphics parsing before describing the details. One challenge is to determine the context of the graphics. Diagrams can be highly ambiguous because of the lack of token specificity, and without guidance from the context, impossible to disambiguate. This is clearly illustrated in Figure 3. This figure could easily be any one of the following: Three coupled cars of a train, a sequence of three gene regions, or three buildings connected by walkways.

In our work to date we have solved the domain problem by assuming that the context, typically the document in which the graphics is embedded, severely limits the interpretation of the graphics. For example, we have designed grammars for gene diagrams, and applied them to diagrams which we know

are gene diagrams. We have not yet attempted to automate the determination of contexts.

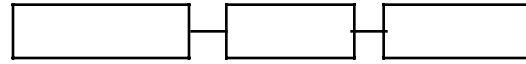


Figure 3. A diagram that has a variety of interpretations.

For a given domain, we write grammars that have rules (productions) such as the following, which is one rule in a finite-state automaton diagram grammar [8]:

```
(Labeled-arrow -> Arrow Label
  (Arrow)
  (Label (touch Arrow '?')
    :select (min
      (distance
        (center Label)
        (arrow-back Arrow))))))
```

This rule states that a Labeled-arrow is made up of an Arrow and a Label. The Arrow constituent is defined by still another production not shown (it is not a primitive). The Label is also defined by another production (basically it must be numeric text). The important parts of the body of this rule are the *constraints*. The first constraint, (touch Arrow '?') states that the Label must be drawn from a set of graphical objects that touch the Arrow (are very near it). The second, :select constraint, states that if there is more than one Label found, it should be the one closest to the shaft of the Arrow (the arrow-back).

The parsing proceeds by a top-down, depth-first search of the space of possible solutions. The body of the rule is processed in the order stated (Arrow before Label in the rule above). Careful choice of this order and careful design of the constraints allows the parsing to be quite efficient, typically parsing diagrams of a few hundred elements in a few seconds. The parser used here was implemented in Macintosh Common Lisp and run on a G3 Macintosh [8].

## 5. SEMANTICS FOR ESTABLISHING TEXT-GRAPHICS COREFERENCE

### 5.1 Semantics and Logical Form

Logical form is a structure representing the invariant meaning of an utterance that is independent of certain variations in the syntactic structure ([22] and Chap. 8 of [1]). Logical form typically uses first-order logic extended to deal with natural language issues. Logical form includes objects, predications, and various generalized quantifiers.

### 5.2 Semantics of Text

A syntactic parse of text is not a useful representation of its semantics, the meaning of the text being analyzed. In English, the two forms, "The rod presses against the wheel." and "The rod is pressed against the wheel." are semantically equivalent but have different syntactic structures. An invariant form that represents the semantics of both can be built as a predicate-argument structure with the predicate **press** and arguments "against", "rod" and "wheel":

(**press manner:** against; **agent:** rod; **object:** wheel)

This representation is the variant of logical form which we will use in this paper. (Our examples do not require or use quantifiers.) Once both the text and related graphical items are represented as logical forms, they can be integrated. The

syntactic parsing of both modalities is automated, but the transformations to logical form are done manually.

Logical forms are derived from text parses, typically by semantic information attached to the syntactic rules, which are used to construct the semantics of the larger forms (root of the parse tree) by composing the semantics of the child nodes -- a bottom-up compositional semantics.

### 5.3 Semantics and Logical Form -- Graphics

As with text, the syntactic parse of a diagram must be rewritten to express its semantics properly. Figures 4A and 4B show two drawings that can both be described by the text discussed in the previous section.

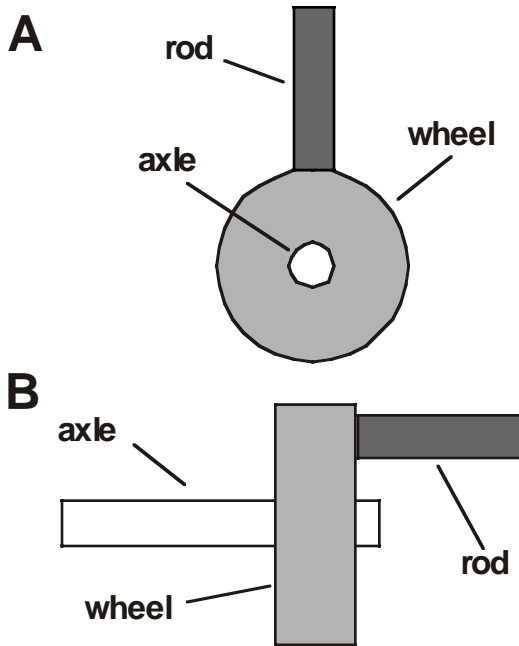


Figure 4. Illustrations of two distinct mechanisms that both contain a rod pressing against a wheel.

The predicates needed to build logical forms from graphics parses are often present in the constraints that are used in our grammars. For Figure 4A, a typical grammar might include the following production,

```
PRW1: Circle-attached -> Circle Rect
      (Circle)
      (Rect (touch Circle '?')
       constraint: (rectanglep Rect))
```

and for Figure 4B,

```
PRW2: Rectangle-attached -> Rect1 Rect2
      (Rect1 constraint: (rectanglep Rect1))
      (Rect2 (touch Rect1 '?')
       constraints: (rectanglep Rect1)
                  (different Rect1 Rect2))
```

The point to note here is that the touch predicate is included in both of these productions, and it is this predicate that will be transformed into the logical form predicate **press**. Nevertheless, the graphics semantic analysis still requires additional effort in order to build the appropriate logical form. Graphics depicts circles and rectangles, not wheels and rods

explicitly, as language is able to. Contextual and domain knowledge is required to properly interpret geometrical items and structures as more specific entities. Some of this could be extracted from the text we analyzed above. The result can be represented as associations of the following form, for Fig. 4A,

Circle = "wheel" and Rectangle = "rod"

Similarly, graphical adjacency as represented by touch needs to be related to "pressed" or "presses". When this is done, using the constraint for production PRW1, the result is a logical form very similar to the logical form for text,

(touch/press agent: Rect/"rod"; object: Circle/"wheel")

### 5.4 Integration of Text and Graphics Semantics

Examining the separate text and graphics logical forms above makes it clear that they can be unified, resulting in the *integrated logical form*,

(touch/press manner: "against"; agent: Rect/"rod"; object: Circle/"wheel")<sub>1</sub>

The unification process adds the preposition of manner, "against". It is obvious that the final result represents more than the simple sum of the two separate logical forms, since bindings have been established between objects in the two modalities. For example, the query "Highlight the object that is pressed against." could be responded to by graphical highlighting of the Circle (aka "wheel").

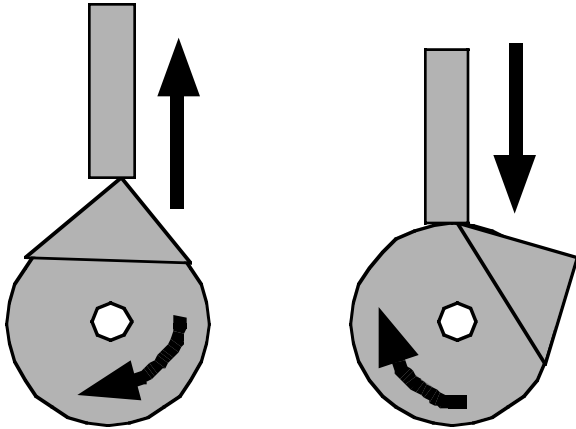
## 6. INTEGRATION OF A MECHANISM EXPLANATION

The description of the nature and action of a cam is shown in Figure 5, adapted from "The New Way Things Work" [19].

The grammar below parses both the right and the left portions of the diagram. The productions for the lower level constituents are omitted. The grammar distinguishes between the Attached-1 constituents which are in direct contact with the wheel, and the Attached-2 constituents which are in contact indirectly, because they touch Attached-1 items. The Attached constituents are required to be Polygons and the Outer-wheel and Axle are required to be Circles. The triangular projection portion of the cam is an Attached-1 item whereas the rod is an Attached-2 constituent in the left figure and an Attached-1 constituent in the right one.

```
PC1: Wheel-assembly -> Wheel Attached-1 Attached-2
      (:non-sharable Attached-1)
      (Wheel)
      (Attached-1 (touch '? Wheel))
      (Attached-2 (touch '? Attached-1))
```

```
PC2: Wheel -> Outer-wheel Axle
      (Outer-wheel)
      (Axle (inside Outer-wheel '?')
       :constraint
        (concentric (Circle (Outer-wheel self))
                    (Circle (Axle self))))
```



**Figure 5. The Cam example, showing the cam and rod in two positions.**

The caption of the original figure in [19] is, "The egg-cracker uses a cam, a device which in its most basic form is simply a fixed wheel with one or more projections. A rod is pressed against the wheel, and as the wheel rotates, the rod moves out and in as the projection passes."

The development of the integrated representation can be illustrated by showing the logical forms derived from the parses of the diagram and the first sentence of the caption [19]. We focus on the portion referring to the wheel and projections, "... a fixed wheel with one or more projections."

(with agent: "wheel" object: "projections") -- text

(touch agent: wheel object: Attached-1) -- graphics

Note that the graphical logical form is derived from the constraint in the production rather than the constituent structure, as we noted earlier. The two forms require further semantic processing to bring them into an integrated form, using the appropriate sense of "with", resulting in,

(touch agent: "wheel" object: "projections")

Unification of the text and graphical forms yields the integrated form,

(touch/touch agent: wheel/"wheel" object: Attached-1/"projections")<sub>1</sub>

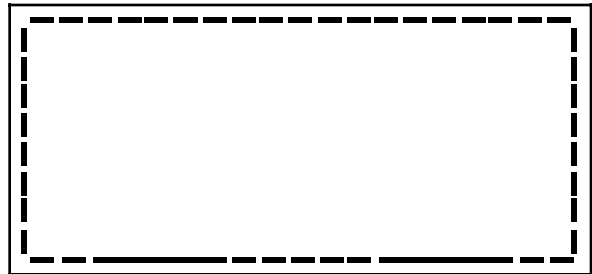
An important part of this result is the establishment of the co-referring items Attached-1 and "projections". The more difficult point that the rod is not a "projection" would have to be resolved using a qualitative physics argument about the motion of the rod and the fact that it is an Attached-1 constituent in only one configuration.

## 7. INTEGRATION OF A FABRIC CONSTRUCTION EXPLANATION

A magazine article has described the construction of a fabric mobile to hang on a Christmas tree, [3] (our Figure 2). It is constructed from two layers of fabric the shape of a Christmas tree, which are stitched together leaving two small openings through which stuffing can be inserted. We developed a graphics grammar with seven productions and parsed the simplified diagram in Figure 6 showing the stitching and two openings.

Two of the most relevant productions are,

```
PF1: Pattern -> All-Stitching Unattached
      (All-Stitching)
      (Unattached (touch All-Stitching '?))
PF2: Unattached -> Set(Line)
      (:element-constraints
       (> (length Line)
        (* 4 (stitch-length All-Stitching))))
```



**Figure 6. A fabric construction showing the stitching and two openings on the bottom edge through which stuffing can be inserted. Simplified from [3].**

The openings are represented in the grammar as Unattached, portions of the boundary (not stitched together). The openings are represented by solid lines in the figure, as they were in the original. The grammar distinguishes between stitching and openings based on the fact that openings are noticeably longer than stitch lines (otherwise they couldn't function as openings!). In this grammar each opening must be at least four times the *stitch-length* attribute of the *All-Stitching* constituent. In this example, Unattached is a set containing the two horizontal lines on the lower boundary in the figure.

A portion of the text accompanying the original figure is, "Using the full-size pattern on the pull-out sheet, make a template of the tree. Fold the green fabric right sides together. Place the tree on the fabric and trace around it with a pencil. Mark the 2 open spaces."

The development of the logical forms for text and graphics and their integration for the construction task is similar to the cam example. An important part of the integration is the unification unattached/"opening".

## 8. ANALYSIS OF DATA GRAPHS

Data graphs are of paramount importance in reporting scientific data of all types. We have demonstrated automated parsing of data graphs in our earlier work [8,9]. Though the semantics of the "background" of data graphs is not difficult to understand, e.g., the scale lines and tick marks and labels, the semantics of the data itself is quite variable and context-dependent. For example, in one data set a maximum or a plateau in the data values may be of importance and in another, the amount of noise in the data may be. Establishing coreference is not impossible though, since terms such as "maximum", "asymptote", "linear region" and the like can potentially be identified by a numerical and statistical analysis of the data values themselves. This would require data analysis functions to be integrated into the semantic analysis. In the logical forms a reference such as maximum would include some data value or values, possibly a small region around the maximum rather than a single data point.

## 9. A THEORY OF TEXT/GRAPHICS DISCOURSE

### 9.1 Prior Work

The work on natural language discourse is voluminous and can only be touched on briefly. For text discourse, *Centering Theory* has as its goal the modeling of the local coherence of discourse [11]. Briefly, centering posits the existence of a set of *forward-looking centers*,  $C_f$ , of an utterance, and single *backward-looking centers*,  $C_b$ , in later utterances. A  $C_f$  set might include individuals and a  $C_b$  could be the pronoun "she". Centers are semantic objects determined by a combination of factors including the intentional structure created by the author and the attentional state of the reader. If a document adheres to the centering constraints it will decrease the inference load on the reader during the construction of a mental representation of the discourse. On the surface, Centering Theory appears to rely heavily on the sequential nature of spoken or written discourse. But more generally, it is concerned with the ultimate goal of constructing a representation of the entire discourse, and this allows us to relate Centering to our approach. The primary point of contact between our theory and Centering is that both theories share the goal of trying to identify those aspects of discourse that reduce the inference load on the reader, that makes one discourse design better than another. The major difference that our theory brings is that there is an important role for the reader in deciding when and how to negotiate attention shifts between the text position of the moment and some portion of the graphics. Thus, in text-graphics discourse it is not only the author's design of the discourse that determines the flow but the reader's active involvement in controlling the flow.

*Discourse Representation Theory* (DRT) [16], builds a *discourse representation structure* (DRS) derived from syntactic analysis of utterance units, normally sentences. In the DRS, elements introduced in later utterances are bound to earlier ones (coreference), giving coherence to the discourse. The DRS is a diagrammatic structure closely related to first-order predicate calculus. Beyond the coreference issues that are basic to all discourse analysis, DRT deals with more complex issues of conjunction, implication, scoping, tense, and more.

A discourse theory built around maps and natural language descriptions has been developed [20]. It focuses on compass directions and connectivity and does not attempt to parse the maps themselves. Work on the generation of text to accompany figures has been done which requires careful construction of referring expressions, paying attention to saliency and focus [2]. The figures were taken as givens in that work and were not represented as parsed structures. There has also been work on the generation of referring expressions which are definite descriptions of real-world objects. But the objects are described using attribute-value structures, rather than directly using graphics [7].

### 9.2 Text-Graphics Discourse Theory (TGDT)

Our theory, TGDT, builds on earlier approaches to text discourse but adds four novel components: 1. Representation of the semantics of graphics. 2. Non-sequential access by the reader (attention shifts). 3. Bi-modal, text-graphics lexicons, mental models and reasoning by the reader. 4. The production of an integrated model of the discourse. In this paper we have presented examples of all of these components.


#### 9.2.1 Semantics of Graphics

Graphics and its structure and content require analysis by the reader based on the reader's visual reasoning repertoire, both in terms of the objects per se (e.g., arrows) and domain knowledge (e.g., arrows as dimensioning devices). We have shown that it is possible to parse graphics and build semantic representations for them.

#### 9.2.2 Non-Sequential Access of Text and Graphics

There is simply no well-defined order of processing of text and graphics by the reader. For written text, Centering Theory is focused on the author who should construct a discourse that attempts to minimize the inferential load on the reader. An important insight in our theory is that there is an active *role for the reader* in text/graphics discourse: We claim that the reader will try to minimize the inference load required to construct a model of the discourse by shifting his or her attention from material in one modality to material in the other as the reading proceeds. Readers may direct their attention to particular portions of the graphics material at various times, e.g., in the cam example, they might first focus on the cam then on some of the text, and then on the rod. We have done initial studies of subjects' attention shifts using the Restricted Focus Viewer, RFV. Our early data shows that subjects appear to develop spatial maps of both graphics and text.

#### 9.2.3 Bi-Modal Mental Models

We assert that a person must possess a mental representation of the world that integrates the text and graphics modalities. This is evident from the fact that people can talk about what they see and conversely, they can visually attend to something described in language. At the simplest level, we assert the existence of an integrated lexicon, ILex, that contains entries such as  $cat_i$ , tied to the utterance "cat" as well as to mental constructs corresponding to a visual form, . In addition, there must be an integration of reasoning about language and the visual world when inference is required such as might be triggered by the utterance, "the left-most red ball".

#### 9.2.4 An Integrated Discourse Representation

We have argued that an integrated representation of discourse can be constructed by unifying the separate logical forms that express the semantics of the text and the semantics of graphics. We have developed automated parsers for both diagrams and text. We have then manually constructed semantic representations for both as logical forms. By identifying coreferential items across the two modalities, we have shown that integrated representations can be created that support additional inferences beyond those based on the semantics of the two modalities taken separately.

## 10. SUMMARY AND CONCLUSIONS

Science and technology as we know it could not exist without graphics [17]. Therefore it is of utmost importance to understand the discourse structure of text/graphics documents. We have argued in this paper that it is possible to develop a semantics for graphics and generate an integrated model of a text/graphics discourse, building on previous theories of text discourse. A major component of our theory is that we posit an integrated natural language / visual language lexicon and other attendant mental modules that allow people to understand such bimodal discourse. We have pointed out a major difference in the character of text discourse and text/graphics discourse, and that is the *role of the reader* in the latter. The reader constantly exercises choices to shift his

or her attention between the text and the graphics components of the discourse during reading. We have also demonstrated some computational approaches to the automation of building discourse structure, at the level of syntax.

This paper is the first of a series based on our research and as a consequence, covers a large number of topics in only modest depth. Future papers will focus on specific aspects in greater depth.

## ACKNOWLEDGEMENT

We thank the anonymous reviewers for helpful suggestions, especially concerning the experimental psychology literature. This work was supported in part by the National Science Foundation, grant no. IIS-9978004 to RPF.

## REFERENCES

- [1] Allen, J., *Natural Language Understanding*, 574 pp., Benjamin/Cummings Publishing Company, Menlo Park, CA, 1994.
- [2] Andre, E., and T. Rist, Referring to World Objects with Text and Pictures, in *Coling 94*, pp. 530-534, 1994.
- [3] anonymous, The Gift Tree, in *Quilt It for Christmas*, pp. 34-35, 1999.
- [4] Blackwell, A.F., A.R. Jansen, and K. Marriott, Restricted Focus Viewer: A Tool for Tracking Visual Attention <http://www.csse.monash.edu.au/~tonyj/RFV/>, in *Theory and Application of Diagrams*, edited by M. Anderson, P. Cheng, and V. Haarslev, pp. 162-177, Springer, Edinburgh, 2000.
- [5] Brown, G., *Speakers, Listeners and Communication : Explorations in Discourse Analysis*, Cambridge Univ. Press, Cambridge, UK, 1995.
- [6] Brown, G., and G. Yule, *Discourse analysis*, 288 pp., Cambridge University Press, Cambridge, England, 1983.
- [7] Dale, R., and E. Reiter, Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions, *Cognitive Science*, 18, 233-263, 1995.
- [8] Futrelle, R.P., The Diagram Understanding System Demonstration Site, <http://www.ccs.neu.edu/home/futrelle/diagrams/demo-10-98/>, 1998.
- [9] Futrelle, R.P., and N. Nikolakis, Efficient Analysis of Complex Diagrams using Constraint-Based Parsing, in *ICDAR-95 (Intl. Conf. on Document Analysis & Recognition)*, pp. 782-790, Montreal, Canada, 1995.
- [10] Glenberg, A.M., Langston, W.E., Comprehension of Illustrated Text: Pictures Help to Build Mental Models, *Journal of Memory and Language*, 31, 129-151, 1992.
- [11] Grosz, B.J., A.K. Joshi, and S. Weinstein, Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics*, 21 (2), 203-225, 1995.
- [12] Hegarty, M., Just, M. A., Constructing Mental Models of Machines from Text and Diagrams, *Journal of Memory and Language*, 32 (6), 717-742, 1993.
- [13] Jackendoff, R., Semantics and Cognition, in *The Handbook of Contemporary Semantic Theory*, edited by S. Lappin, pp. 540-559, Blackwell, Oxford, 1996.
- [14] Joshi, A.K., B.L. Webber, and I. Sag, *Elements of Discourse Understanding*, Cambridge Univ. Press, Cambridge, UK, 1981.
- [15] Jurafsky, D., and J.H. Martin, *Speech and Natural Language Processing*, Prentice-Hall, Upper Saddle River, NJ, 2000.
- [16] Kamp, H., and U. Reyle, *From Discourse to Logic*, 713 pp., Kluwer Academic, Dordrecht, 1993.
- [17] Latour, B., Drawing things together, in *Representation in Scientific Practice*, edited by M. Lynch, and S. Woolgar, pp. 19-68, MIT Press, Cambridge, MA, 1990.
- [18] Linguistic Data Consortium, Bracketing Guidelines for Treebank II Style Penn Treebank Project, <ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/>, 1995.
- [19] Macaulay, D., *The New Way Things Work*, 400 pp., Houghton Mifflin, Boston, 1998.
- [20] Pineda, L., and G. Garza, A Model for Multimodal Reference Resolution, in *Referring Phenomena in a Multimedia Context and their Computational Treatment*, pp. 99-117, Assoc. Computational Linguistics, Madrid, 1997.
- [21] Selfridge, D., Using Correspondence to Integrate Knowledge Represented in Multiple Domains, (MS thesis) Northeastern University, 1993.
- [22] van Eijck, J., and H. Alshawi, Logical Forms, in *The Core Language Engine*, edited by H. Alshawi, pp. 11-39, MIT Press, Cambridge, MA, 1992.