# Imagistic Modeling in Story Understanding

**Eric Bigelow**
Univ. of Rochester
ebigelow@u.rochester.edu

**Daniel Scarafoni**
Univ. of Rochester
dscarafo@u.rochester.edu

**Lenhart Schubert**
Univ. of Rochester
schubert@cs.rochester.edu

**Alex Wilson**
Univ. of Rochester
alexwilson@rochester.edu

## Abstract

We argue that the knowledge acquisition bottleneck in story understanding has a neglected, but very important dimension, namely *imagistic modeling*. This requires a large repertoire of object prototypes, along with their poses, configurations and behaviors, enabling incremental model construction and, crucially, querying of object relationships during story understanding. To motivate the need for imagistic modeling, we consider some simple first-reader stories, and how we envisage the interplay between symbolic and imagistic modeling. More concretely, we report our initial work on building an imagistic modeling system (IMS) based on the Blender graphical software, with some of the capabilities required for story understanding.

## 1 Introduction

Genuine story understanding by machine is still thwarted by the knowledge acquisition (KA) bottleneck. Before we can overcome that formidable obstacle, whether by machine learning or other methods, we need at least to identify the kinds of knowledge representations required for understanding. While a great deal of AI research has addressed the question of what sorts of symbolic representations could support story comprehension, much less attention has been devoted to the potential role of mental imagery in that process, despite the long-standing interest in such imagery in human understanding among cognitive scientists and neuroscientists (e.g., Johnson-Laird 1983, Paivio 1986, Kosslyn 1994).

In the following, we illustrate how the need for imagistic representations arises in even the simplest first-reader stories. We then outline an approach to integrated symbolic and imagistic modeling and inference, applying this to the previously introduced motivating examples. In that outline, we take for granted the imagistic functionalities, but subsequently we describe our preliminary imagistic modeling system (IMS), and assess the extent to which it can (and cannot) provide the functionalities required. Finally, we discuss related work on building imagistic models of text, and then reiterate our conclusions and sketch our plans for further work.

The reported work does not in itself alleviate the KA bottleneck. On the contrary, we are arguing that the challenge is even larger than one might infer from most work on KA, which tends to treat the term as synonymous with acquisition of relational or rule-like knowledge. But in underscoring the need for extensive imagistic knowledge in deep story understanding, and outlining the possible form and use of such knowledge, we hope to be providing a better understanding of the challenges facing the natural language understanding community.

## 2 The Need for Imagistic Modeling

> "linguistic terms may not in the first place describe or represent meanings as such, but rather serve as triggers for activating concepts of human experience, which are far richer and more flexible than any lexical entry or formalization could possibly represent."
>
> – Thora Tenbrink

The following is a simple story for beginning readers (from Lesson XXXI, Harris et al. 1889):

1. Oh, Rosy! Do you see that nest in the apple tree?
2. Yes, yes, Frank; I do see it.
3. Has the nest eggs in it, Frank?
4. I think it has, Rosy.
5. I will get into the tree.

6. Then I can peep into the nest.
7. Here I am, in the tree.
8. Now I can see the eggs in the nest.
9. Shall I get the nest for you, Rosy?
10. No, no, Frank! Do not get the nest.
11. Do not get it, I beg you.
12. Please let me get into the tree, too.
13. Well, Rosy, here is my hand.
14. Now! Up, up you go, into the tree.
15. Peep into the nest and see the eggs.
16. Oh, Frank! I see them!
17. The pretty, pretty little eggs!
18. Now, Frank, let us go.

One point where the need for imagistic modeling arises particularly clearly is at sentences 3 and 4. We know from the preceding two sentences that both Rosy and Frank see the nest. Yet it is clear from sentences 3 and 4 that they cannot see whether there are eggs in the nest – a fact needed to make sense of their subsequent dialogue and actions. Why is this? One could of course assume symbolic knowledge to the effect that to see whether eggs or other small objects are in a bird's nest in a tree, one needs to be up in the tree. But this would be "cheating", in the sense of supplying overly specific knowledge just to deal with a specific story. To see this, imagine the following variant of the first three sentences:

1'. Rosy, do you see the basket hanging in the cherry tree? Father left it there for picking cherries.
2'. Yes, Frank; I do see it.
3'. Has the basket cherries in it, Frank?

In a "targeted symbolic knowledge" approach, we would now need to assert that to see small objects in a basket in a tree, one needs to be up in the tree! A more general symbolic approach, applicable to more varied situations, would be to assert that seeing the contents of a topless container requires that one be near it, with one's head above it. But there are still problems with such an approach: First, there are objects that can occlude smaller objects much as a container does, but are not ordinarily classified as containers, such as small boats (lacking a deck), pickup trucks and other vehicles with an open cargo area, tires (on the ground), shoes and boots, hollowed-out tree stumps, ditches, blossoms such as those of water lilies or tulips (as might occur in children's stories about frogs, bees, spiders, or other small creatures), hedged or walled-in yards, skateboarding bowls, open-air stadiums, volcanic craters, etc.

In fact, occlusion is a much more general phenomenon, not restricted to concavities; the same geometric principles account for the invisibility of small objects on a table for a toddler (while objects under the table are visible to her), or on the other side of a sufficiently high wall, or of a squirrel on the far side of a large tree trunk, or of persons and objects inside a house (and not at a window) from an outside perspective, or of stars high in the sky from the interior of a car or house, or of the sun behind clouds, and so on. The pervasiveness of occlusion alone already provides a strong case for use of three-dimensional models to support story understanding, noting that with a little imagination we can easily construct stories in which any of the occlusions just mentioned are criterial in understanding the stories.

We noted that to see into a nest, the viewer should not only have a sufficiently high vantage point, but also be *near* the nest. This brings us to the second problem with a purely symbolic approach: "Near" and "next to" relations are very difficult to track symbolically, as has long been apparent in STRIPS-like planners, even in simple rooms-and-blocks domains (Fikes & Nilsson 1971). Such planners often have a "go-to" action whose predicted effect is to bring the robot "next to" (or near) an object, as required for manipulating or moving it. But what *other* "next-to" relations are altered? Often the assumption is made that all but the target "next-to" relation will become false or at least unknown. So even if the robot moves from one end of a full-length wall shelf to the other (arranging books, perhaps), it will not know that it remains close to the same wall, at the same time ceasing to be next to the wall at one end of the shelf, and ending up next to the wall at the other end. A physical robot with sensory equipment can make up for the inadequacies in symbolic reasoning, though at a large computational cost. A much better method is to model its immediate environment (Roy 2005), and to the extent that readers can understand examples like the one just given, the same goes for the environments of characters in stories.

While it would not be impossible to embed such geometric models directly in a logic that allows for quantitative object properties and relationships and even motion, doing so would clog the inference machinery with masses of details only a few of which would be relevant at a given moment

in a story comprehension process. This problem would be aggravated by the *frame problem*, i.e., the need to effectively infer what relationships remain unaltered by the motions or actions of a few agents or objects. While the frame problem can be dealt with fairly effectively with methods that merely add a set of explanatory axioms comparable in size to the set of action-effect axioms that predict change (e.g., Reiter 1991), tracking change and nonchange of numerous quantitative facts in the form of logical predications as a story progresses would be thoroughly impractical.

What these considerations point to is the need for a *specialist* system for modeling spatial relations, supporting a general symbolic understanding and reasoning system but using techniques that exploit the very special nature of the spatial relationships and interactions of complex objects. This is our motivation for designing an imagistic modeling system (IMS). As such, our endeavor is similar to the kind of hybridization strategy that has been pursued in logic-based systems and programming languages that allow for support by taxonomic, temporal, or arithmetic constraint solvers, and more broadly in the tradition of specialist-supported reasoners such as several of those in (Melis 1993).

Before illustrating how we envisage the interplay between a symbolic story processing system and our IMS, we quote two more brief stories in which the need for inferring spatial relations is quite apparent:

> A little girl went in search of flowers for her mother. It was early in the day, and the grass was wet. Sweet little birds were singing all around her. And what do you think she found besides flowers? A nest with young birds in it. While she was looking at them, she heard the mother bird chirp, as if she said, "Do not touch my children, little girl, for I love them dearly." (McGuffey 2005, Lesson XLII)

> This is a fine day. The sun shines bright. There is a good wind, and my kite flies high. I can just see it. The sun shines in my eyes; I will stand in the shade of this high fence. Why, here comes my dog! He was under the cart. Did you see him there? What a good time we have had! (McGuffey 2005, Lesson XXIX)

In the first story, note the contrast with our opening story: We infer that the girl's attention is generally turned downward, perhaps by reference to the prototypical gaze direction of someone seeking objects on the ground. So the nest she spots is likely to be situated close to or on the ground, within a field of view lower than the girl's head. This is confirmed by her ability to look at the young birds, as well as by the mother bird's perception that the girl might touch them. In the second story, the prototypical configuration of a person holding a high-flying kite at the end of a long string, and the position of the sun high in the sky, are essential to understanding why the sun shines in the kite-flyer's eyes. Further, how the shade of a high fence might shield the kite-flyer from the sun, and why a dog under a cart may or may not be noticeable, are best understood from (at least rough) physical models of these entities.

We now outline the way in which an IMS could support the understanding process in a story understanding system, with reference to our opening story.

## 3 Interleaving Linguistic and Imagistic Processing

We will assume that we can obtain correct parse trees and logical forms (LFs) for simple stories as such as those we have mentioned. In fact, we are converging on this goal after some small adjustments in the stories. For example, We change Rosy's name to *Rosie* to prevent the Charniak parser from treating it as an adjective, and we change the question *"Has the nest eggs in it, Frank?"* to the American form *"Does the nest have eggs in it, Frank?"*. After some automatic postprocessing of the parse tree, for example to mark prepositional phrases with their type and to insert traces for dislocated constituents, our semantic interpreter successively produces (i) an initial LF by compositional rules; (ii) an indexical LF in which quantifiers and connectives are fully scoped and intrasentential anaphors are resolved; (iii) a deindexed LF with a speech act predicate and with explicit, temporally modified episodic (event or situation) variables; and (iv) a set of canonicalied formulas derived from the previous stage by Skolemization, negation scope narrowing, equality substitutions, separation of top-level conjuncts, and other operations. The following are the parse trees and the set of formulas derived from sentence (1) of our lead story:

```
PARSE TREES:
``````````
 (FRAG (INTJ (UH OH)) (, ,) (NP (NNP ROSIE)) (. !)),

 (SQ (AUX DO) (NP (PRP YOU))
  (VP (VB SEE) (NP (DT THAT) (NN NEST))
   (PP-IN (IN IN) (NP (DT THE)
                      (NN APPLE) (NN TREE)))) (. ?))
```
```

```
CANONICAL FORMULAS, WITH HEARER IDENTIFIED AS ROSIE:
````````````````````````````````````````````````````
 (SPEAKER DIRECT-OH-TOWARDS*.V ROSIE.NAME),

 (NEST8.SK NEST.N), (NEST8.SK NEW-SALIENT-ENTITY*.N),
 (TREE9.SK ((NN APPLE.N) TREE.N)),
 (TREE9.SK (NN APPLE.N) TREE.N),
 ((SPEAKER ASK.V ROSIE.NAME
   (QNOM (YNQ (ROSIE.NAME SEE.V NEST8.SK)))) [C] E7.SK),
 ((SPEAKER ASK.V ROSIE.NAME
   (QNOM (YNQ (ROSIE.NAME SEE.V
    (THAT (NEST8.SK IN.P TREE9.SK)))))) [D] E7.SK)
```

Note that for browsability we generate our formulas in infix form, i.e., the predicate follows the first (subject) argument. Also the unnamed objects referred to have been given Skolem names, such as TREE9.SK, and the corresponding type-predications, such as (TREE9.SK ((NN APPLE.N) TREE.N)), have been separated out. Also "seeing the nest in the tree" has been approximated with the conjunction of "seeing the nest" and "seeing *that* the nest is in the tree"; correspondingly the speaker's question has been broken into two parts. YNQ is a yes-no question-forming operator (with intensional semantics) and QNOM is a reification operator mapping a question intension to an abstract individual. (Together, (QNOM (YNQ ...)) are equivalent to (WHETHER ...).) The '[c]' and '[d]' operators respectively relate a sentence to a situation it characterizes, and a situation that it partially describes. (We omit references to our representational, interpretive, and inference tools for anonymity reasons.)

We could now use our inference engine to infer that the speaker's question presupposes that the speaker sees the nest, and sees that it is in the tree, and that this presupposition can be accommodated, i.e., NEST8.SK is in fact in TREE9.SK, and hence that it is (presumably) in the crown of TREE9.SK. The knowledge needed for the presupposition inference can be formally stated using the metasyntactic capabilities of our knowledge representation, and the inference that the nest is in the crown of the tree would be based on the stereotyped knowledge that a nest in a tree is normally in the crown of the tree.

Additional stereotyped knowledge is needed before imagistic modelling can begin to contribute to understanding: Trees are generally outdoors and standing upright, rooted in the ground; persons who can see a small outdoor object (such as a nest) are typically outdoors as well and near the object; further, they are usually upright on the ground, unless we have reason to believe otherwise.

At this point, we can begin to generate a model scene: Two children, one called Rosie, are outdoors on the ground, near an apple tree with a nest in it, and they can see the nest. (That the persons are children is a meta-inference from the genre of the book the story is taken from.) Our modeling system has successfully constructed this scene (and some other simple ones), also predicting via evaluation of a "can-see" predicate that the children cannot see *into* the nest. While our implementation has not gone beyond this point, the prediction is clearly crucial for verifying an implicature of sentence 3 (*Does the nest have eggs in it, Frank?*), namely that Rosie cannot see into the nest, despite being able to see the nest. Further symbolic story processing at sentence 5 leads to the inference that Frank will end up in the crown on the apple tree if he climbs it, and by modeling this future situation we could confirm his expectation expressed by sentence 6, again using "can-see". Sentence 7 verifies the (tentative) inference that Frank carried out the intention expressed in sentence 5, and sentence 8 confirms both Frank's expectation and the model's prediction.

Without going into further details, we hope it is clear that (as already noted) full understanding of sentences 12-16 likewise depends on supporting symbolic knowledge and reasoning about the effects of, and intentions behind, physical actions and speech with an evolving model of the described situations, allowing prediction and verification of spatial relationships by "inspection", such as visibility relationships.

## 4 Imagistic Modeling System

Our modeling environment used the Blender system (www.Blender.org), a free Gnu software system for designing and rendering complex 3D scenes. It provides a means of easily creating object models, attaching properties and constraints to those models, and performing operations in a 3D space. Blender is primarily written in python and supports native development of python scripts using its API. Python modules executed in the Blender scripting environment may access and manipulate object information in real time.

Based on input from our inference engine, the specialist constructs a 3D model for given entities and relations derived from a portion of a story describing some situation. This model may then be queried by the inference engine for spatial relations in the scene, such as 'Can-See', 'Near', 'Under', 'In', and 'Within'. Both placement and
```

querying rely primarily on computed *acceptance areas* (AAs – in general 3D regions) in the vicinity of objects. Entities are obtained as instances of object prototypes in a database, stored in the form of 3D Blender files and XML trees. Each object is comprised of a series of part models arranged relative to an empty parent object. Default poses and configurations are also stored for objects, and these are invoked with a certain probability given that instances of all objects in a configuration are in the scene. For example, a scene with a set of chairs and a table might be automatically arranged such that the chairs encircle the table.

Based on a class structure allowing for Scene, Entity, and Predication classes, a Scene modeling a story situation contains pointers to all active Entities and Predications. Entity instances contain a pointer to the Blender object representing the entity as well as methods for drawing and manipulating the object, and pointers to all predications relevant to the entity's parameters. A Predication is an instantiated spatial predicate applied to particular argument entities and includes methods imported from the predicate library.

Each predicate in the predicate library has two distinct functions, one for object placement and the other for querying. Both are based on associated functions that compute AAs for parameterized 3D objects, drawing on a sublibrary of primitive spatial operations.

For object placement, an AA is represented as a tuple of maximum and minimum values for the x, y, and z dimensions. For example, suppose that an object A whose maximum dimension is 1.0 bu (Blender units) is specified as being near a second object B whose maximum dimension is 2.0 bu, and object B is located at (2, 2, 0). A reasonable AA for placing A based on the predication [A near B] might then be: ((-1,5), (-1,5), (-3,3)). This area would be randomly sampled to determine A's modeled location.

Each query function returns a numerical value ranging from 0 to 1 depending on how well the predication is satisfied. Widely different approaches are used to query predications depending on which predicate is instantiated. 'Near', for example, performs a simple mathematical calculation depending on object sizes and distance from one another. 'Under' generates a temporary block representing its AA and returns a value depending on what proportion of the second object's volume intersects this block.

A rejection technique based on predicate querying is also used to aid in object placement. After AA bounds are determined and an object is placed, predications affected by its parameters are queried. If any of these returns a value of 0, the object is placed again.

These methods have proved adequate for the simple situations we have so far considered.

## 5 Related Work

Existing research in qualitative spatial reasoning emphasizes the utility of acceptance areas and qualitative variance in predicate meanings as a function of object shape and size (Hernández, 1994; Hernández et al., 1995; Hernández, 1997; Clementini et al., 1997; Cohn, 1997; Cohn and Hazarika, 2001; Tappan, 2004). Kadir et al. (2011) use Blender as their basis for modeling scenes from natural language input, utilizing similar methods to object placement as ours. WordsEye is a powerful text-to-scene conversion system capable of modeling complex visual scenes based on layers of spatial predications (Coyne and Sproat, 2001; Coyne et al., 2010; Coyne et al., 2010). However, existing text-to-scene conversion systems such as this differ crucially from ours in that they are designed only for predicate-based scene construction and not for querying.

Soar's Spatial and Visual System (SVS) described by Wintermute (2009) is conceptually similar to the IMS. Predicate-specified AAs are used to create an imagistic scene representation, which can be then used to query further predicates. SVS differs from our system in that it is only built to model very simple scenes and it is not designed to handle linguistic input (Wintermute, 2010).

## 6 Conclusion

The very simple first-reader stories we looked at illustrated the need for spatial knowledge about humans, ordinary objects, and their typical poses and configurations in story understanding. The spatial models for this purpose need not be as fine-grained as those intended for creating realistic, pleasing 3D scenes. But the number of ordinary object types is very large, and their typical kinds of configurations even larger. Thus developing a large imagistic knowledge base – ultimately encompassing motion – should be considered an important challenge by the NLU community.

## References


Eliseo Clementini, Paolino Di Felicea, and Daniel Hernández. 1997. Qualitative representation of positional information. *Artificial Intelligence*, 95(2):317–356.

Anthony G. Cohn. 1997. Qualitative spatial representation and reasoning techniques. *KI-97: Advances in Artificial Intelligence*, 1–30. Springer, Berlin, Heidelberg.

Anthony G. Cohn and Shyamanta M. Hazarika. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1):1–29.

Bob Coyne and Richard Sproat. 2001. WordsEye: An Automatic Text-to-Scene Conversion System. *Proceedings of SIGGRAPH 2001*.

Bob Coyne, Richard Sproat, and Julia Hirschberg. 2010. Spatial relations in text-to-scene conversion. *Computational Models of Spatial Language Interpretation, Workshop at Spatial Cognition*.

Bob Coyne, Owen Rambow, Julia Hirschberg, and Richard Sproat. 2010. Frame semantics in text-to-scene generation. *Knowledge-Based and Intelligent Information and Engineering Systems*, 375–384. Springer, Berlin, Heidelberg.

Daniel Hernández 1994. *Qualitative representation of spatial knowledge*, volume 804= .

Daniel Hernández, Eliseo Clementini, and Paolino Di Felicea. 1995. *Qualitative distances*, 45–57. Springer Berlin Heidelberg.

Daniel Hernández 1997. Qualitative vs. fuzzy representations of spatial distance. *Foundations of Computer Science: Potential Theory - Cognition. Lecture Notes in Computer Science*, volume 1337: 389–398. Springer, Verlag, Berlin.

Philip N. Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Rabiah Abdul Kadir, Abdul Rahman Mad Hashim, Rahmita Wirza, and Aida Mustapha. 2011. 3D Visualization of simple natural language statement using semantic description. *Visual Informatics: Sustaining Research and Innovations*, 28(1):114–133. Springer, Berlin, Heidelberg.

Ken Kahn. 1979. Creation of Computer Animation from Story Descriptions. *Ph.D. Thesis, AI Tech. Report 540*. AI Lab, MIT, Cambridge, MA

Stephen M. Kosslyn. 1994. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA.

William H. McGuffey. 2005 (original edition 1879). McGuffey's First Eclectic Reader. *JOURNAL NAME*, 28(1):114–133. John Wiley and Sons, New York.

Erica Melis. 1993. Working Notes of the 1993 IJCAI Workshop on Principles of Hybrid Representation and Reasoning. *International Joint Conferences on Artificial Intelligence*.

Marvin Minsky. 1974. *A Framework for Representing Knowledge*. from http://web.media.mit.edu/ minsky/papers/Frames/frames.html.

Amitabha Mukerjee, Kshitij Gupta, Siddharth Nautiyal, Mukesh P. Singh, and Neelkanth Mishra. 2000. Conceptual description of visual scenes from linguistic models. *Image and Vision Computing*, 18(2):173-187.

Allan Paivio. 1986. *Mental representations: a dual coding approach*. Oxford University Press, Oxford, England.

Raymond Reiter. 1991. The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, 359–380. Academic Press, New York.

Dan Tappan. 2004. Monte Carlo Simulation for Plausible Interpretation of Natural-Language Spatial Descriptions. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, USA.

Samuel Wintermute. 2009. An Overview of Spatial Processing in Soar/SVS. *Report CCA-TR-2009-01*, U. Michigan Center for Cognitive Architecture.

Samuel Wintermute. 2010. Using imagery to simplify perceptual abstraction in reinforcement learning agents. *Ann Arbor, 1001*, 48109-2121.

Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. 2004. Mental Imagery for a Conversational Robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3):1374–1383.

Richard E. Fikes and Nils J. Nilsson. 1972. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3):189–208.