# Knowledge-Based Spatial Reasoning for Scene Generation from Text Descriptions

**Dan Tappan**

College of Engineering, Idaho State University
921 S. 8th Ave., Stop 8060
Pocatello, ID 83209-8060
tappdan@isu.edu

**Abstract**

This system translates basic English descriptions of a wide range of objects in a simplistic zoo environment into plausible, three-dimensional, interactive visualizations of their positions, orientations, and dimensions. It combines a semantic network and contextually sensitive knowledge base as representations for explicit and implicit spatial knowledge, respectively. Its linguistic aspects address underspecification, vagueness, uncertainty, and context with respect to intrinsic, extrinsic, and deictic frames of spatial reference. The underlying, commonsense reasoning formalism is probability-based geometric fields that are solved through constraint satisfaction. The architecture serves as an extensible test-and-evaluation framework for a multitude of linguistic and artificial-intelligence investigations.

## Introduction and Background

A simple description like *a large dog is in front of a cat and near a small tree* explicitly specifies only a tiny fraction of the details that a corresponding image contains. Most of the content comes from an implicit, commonsense, contextual understanding of the words. Such spatial reasoning, like most intelligent processes, is a difficult computational task to emulate despite its apparent, intuitive simplicity for humans (Herskovits 1986, Tversky 2000). What makes the problem especially troublesome is that computers lack our intangible knowledge of the world and powerful abilities to reason intelligently over it. This work addresses the primary aspects of these issues in terms of what to represent and how to represent it. It uses a simple representation of a description in conjunction with a knowledge base of relevant spatial details to define the declarative form of a valid solution. A constraint satisfaction algorithm then generates any number of corresponding interpretations with plausible positions, orientations, and dimensions for the objects.

Four knowledge-based spatial issues are the focus: *underspecification*, or the lack of complete details in a description, requires background knowledge to supply implicit information; *vagueness*, or the imprecise nature of descriptions, requires knowledge that defines a range of plausible interpretations; *uncertainty*, or the lack of commitment to a particular interpretation, requires knowledge of preferences over this range; and *context*, or the different interpretation of objects in certain combinations, requires knowledge to identify and interpret such patterns.

These issues are considered for three frames of spatial reference (Olivier and Tsujii 1994). The *intrinsic* (object-centered) frame generally applies to objects that have an accepted front, like dog. The *extrinsic* (environment-centered) frame and the *deictic* (viewer-centered) frames are generally the opposite case for objects without such a front, like tree. They correspond to the viewer's position being explicitly stated or loosely implied, respectively.



## Knowledge Representation

A description consists of nouns, adjectives, prepositions, and various support words. The nouns refer primarily to animals and plants within a zoo scenario because they exhibit a variety of interesting and visually appealing spatial characteristics. The adjectives play a role in the contextually appropriate determination of size. The prepositions are 58 spatial relations for position (e.g., *in front*, *left*, *north*, *between*), distance (e.g., *inside*, *near*, *far*), and orientation (e.g., *facing toward*, *away from*, *north*).

### Explicit Representation

The explicit knowledge in a description is represented with a semantic network of object nodes, attribute nodes, and directed relation arcs, which map closely to nouns, adjectives, and prepositions, respectively. For example, Figure 1a depicts the semantic network for *Loki is a small retriever; the tree is north of Loki; Loki is facing the tree.*

### Implicit Representation

To understand the meaning of the description even superficially requires deeper analysis into what the objects

are and how their spatial rules apply to them (Davis 1990). This implicit, commonsense background knowledge is represented in a knowledge base that is similar to an inheritance hierarchy in object-oriented programming. It currently contains 108 concepts that either inherit their contents from their ancestors or define/override them. A simplified example appears in Figure 1b. Linking the semantic network to the knowledge base provides the objects with the appropriate rules for interpreting their position and orientation relations and dimension attributes.

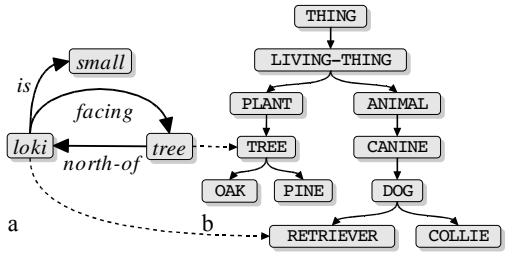



Figure 1: Semantic Network and Knowledge Base

## Spatial Relations

Each spatial relation is associated with one or more circular, two-dimensional fields of 100 rings and 32 sectors that have two complementary parts (Yamada 1993, Gapp 1994, Olivier and Tsujii 1994, and Freska 1992). The *geometry* specifies where another object can and cannot appear with respect to the object in the center. Most relations use variants of the wedge and ring fields in Figures 2a-b. The *topography* overlays a probability distribution on the geometry to specify preferences in placement, as Figures 2c-d show. Fields may also be combined with the standard logical operators *and*, *or*, *xor*, and *not* to represent compositional linguistic expressions like *in front of and far from*.

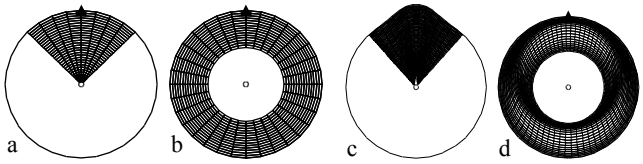



Figure 2: Geometry and Topography of Wedge and Ring

## Spatial Reasoning

The intelligent, commonsense aspects of the spatial reasoning are actually performed earlier by establishing their contextually appropriate, qualitative constraints. Generating a solution from them is now a straightforward, mechanical process of quantitative constraint satisfaction using a greedy, backtracking strategy to generate and test positions and orientations for every pair of objects in a relationship.

## Interactive Visualization

The graphical output is a three-dimensional, interactive world, in which the viewer can move to any vantage point and perspective. It is also possible to query the objects on their underlying representations and constraints, etc. Various display modes depict supporting details like the geometry and topography of the fields, as well as alternative solutions. Figure 3 renders *the dog is south of the tree and near the panther; the panther is to the right of the dog; and the elk is near the maple tree and midrange from and facing away from the pond*.

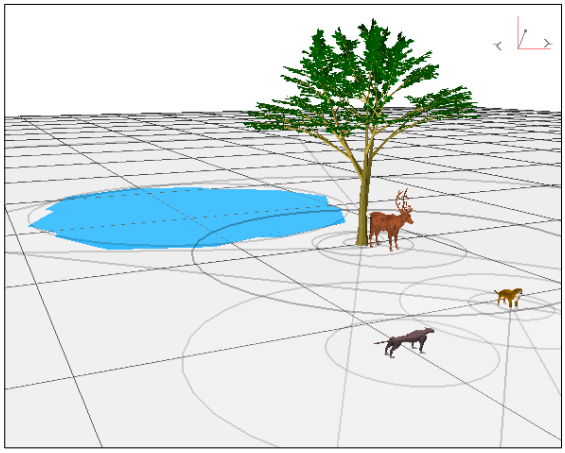


Figure 3: Sample Visualization

## References


Davis, E. 1990. *Representations of Commonsense Knowledge*. Morgan Kaufmann, San Mateo: CA.

Freska, C. 1992. Using Orientation Information for Qualitative Spatial Reasoning. In Frank, A.; Campari, I.; and Formentini, U., eds. *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, LNCS 639, Springer-Verlag, Berlin.

Gapp, K. 1994. Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. In *Proceedings of AAAI-94*, 1393-1398. Seattle, WA.

Herskovits, A. 1986. *Language and Spatial Cognition: An interdisciplinary Study of the Prepositions in English*. Cambridge: Cambridge University Press.

Olivier, P.; and Tsujii, J. 1994. A computational view of the cognitive semantics of spatial prepositions. In *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, New Mexico.

Tversky, B. 2000. Levels and structure of spatial knowledge. In *Cognitive Mapping: Past, present and future*, Kitchin, R.; and Freundshuh, S., eds. London and New York: Routledge.

Yamada, A. 1993. Studies on Spatial Description Understanding Based on Geometric Constraints Satisfaction. Ph.D. diss., University of Kyoto.

# Demo Storyboard
# Knowledge-Based Spatial Reasoning for Scene Generation from Text Descriptions

**Dan Tappan**

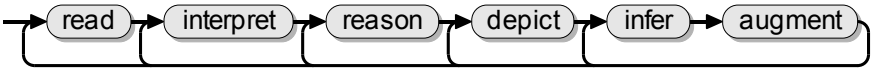
College of Engineering, Idaho State University
921 S. 8th Ave., Stop 8060
Pocatello, ID 83209-8060
tappdan@isu.edu

**Abstract**

This system translates basic English descriptions of a wide range of objects in a simplistic zoo environment into plausible, three-dimensional, interactive visualizations of their positions, orientations, and dimensions. It combines a semantic network and contextually sensitive knowledge base as representations for explicit and implicit spatial knowledge, respectively. Its linguistic aspects address underspecification, vagueness, uncertainty, and context with respect to intrinsic, extrinsic, and deictic frames of spatial reference. The underlying, commonsense reasoning formalism is probability-based geometric fields that are solved through constraint satisfaction. The architecture serves as an extensible test-and-evaluation framework for a multitude of linguistic and artificial-intelligence investigations.

## 1. Process Overview

From the viewer's perspective, the demo can be run as a straightforward input-processing-output model. The input is restricted English text; the processing is various applications of knowledge representation and reasoning; and the output is one or more interactive, three-dimensional visualizations. In this respect, it accommodates a quick 60-second presentation for viewers who are only superficially interested. For those wanting more depth, each of the stages decomposes into significant, lower-level details. This system is designed as a test-and-evaluation platform, so its internals are meant to be exposed, studied, and modified, etc. The final output feeds back into any of the earlier stages for further analysis:




## 2. Read

The input comes in a packaged form called a *vignette*. It contains the properly formatted English text, as well as configuration settings for any experiments to run. For the short-form demo, a number of predetermined vignettes will be available for selection. The long-form demo will allow changes to the settings based on the viewer's interests.

Here is an example description, which specifies the objects in play, and their attributes and spatial interrelations:

```
The scene contains a tree, a zebra named Zeus, and a giraffe.

Zeus is in front of the giraffe.
Zeus is at the fringe of the tree.
The giraffe is in front of the tree.
The tree is in front and left of the giraffe.

The tree is small.
The giraffe is big.
```

Flexible English parsing is not the focus of this work, so the wording is stylistically dry.

There 108 concepts available for the objects, most of which can modified with complementary adjectives of size like *big*, *small*, *long*, *short*, etc. They can be constrained with any combination of 58 spatial relations for position and orientation. A typical description does not contain more than a few objects and relations because humans do not communicate complex, engineering-style configurations and dependencies in ordinary language.

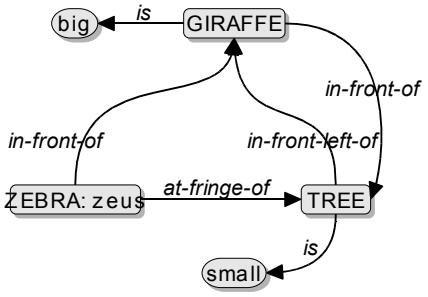
| # | Concept | # | Concept | # | Concept |
|---|---|---|---|---|---|
| 1 | alligator | 37 | garter-snake | 73 | park-bench |
| 2 | american-alligator | 38 | gator | 74 | pen |
| 3 | american-crocodile | 39 | geographic-thing | 75 | pickup-truck |
| 4 | amphibian | 40 | giant-manta | 76 | pig |
| 5 | anaconda | 41 | giraffe | 77 | pine-tree |
| 6 | animal | 42 | golden-eagle | 78 | pink-salmon |
| 7 | ape | 43 | gray-wolf | 79 | pit bull |
| 8 | aquatic-animal | 44 | great-egret | 80 | plant |
| 9 | aquatic-plant | 45 | great-white-shark | 81 | pond |
| 10 | artificial-thing | 46 | hammerhead-shark | 82 | primate |
| 11 | aspen-tree | 47 | hippo | 83 | python |
| 12 | atlantic-octopus | 48 | horse | 84 | rabbit |
| 13 | birch-tree | 49 | human | 85 | raft |
| 14 | bird | 50 | iguana | 86 | red-wolf |
| 15 | blue-whale | 51 | kangaroo | 87 | redwood-tree |
| 16 | bullfrog | 52 | killer-whale | 88 | reptile |
| 17 | bush | 53 | lake | 89 | rhino |
| 18 | cactus | 54 | leopard | 90 | river |
| 19 | cage | 55 | lily-pad | 91 | rodent |
| 20 | camel | 56 | lion | 92 | saguaro |
| 21 | canine | 57 | lizard | 93 | saint-bernard |
| 22 | cat | 58 | loch-ness-monster | 94 | salmon |
| 23 | cherry-tree | 59 | mallard-duck | 95 | sea-monster |
| 24 | coho-salmon | 60 | mammal | 96 | shark |
| 25 | colobus-monkey | 61 | man | 97 | snake |
| 26 | corral | 62 | manta | 98 | snapping-turtle |
| 27 | crocodile | 63 | maple-tree | 99 | swine |
| 28 | dog | 64 | marsupial | 100 | tree |
| 29 | domestic-cat | 65 | mexican-wolf | 101 | turtle |
| 30 | elephant | 66 | mondopod | 102 | ungulate |
| 31 | elk | 67 | monkey | 103 | whale |
| 32 | emperor-penguin | 68 | mountain-gorilla | 104 | white-pelican |
| 33 | feline | 69 | octopus | 105 | wild-cat |
| 34 | fern | 70 | pacific-octopus | 106 | willow-tree |
| 35 | fish | 71 | palm-tree | 107 | wolf |
| 36 | fountain | 72 | panther | 108 | zebra |

3. **Interpret**

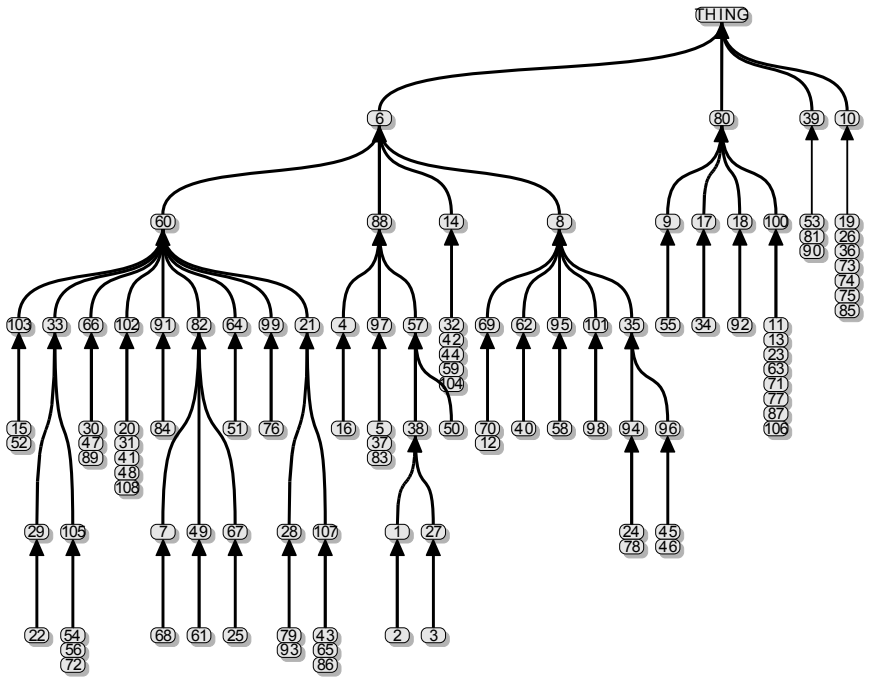
The interpretation stage involves transforming the English description into a more concise form without the extraneous linguistic elements. For example, the text in (2) first reduces to:

```
tree zebra/zeus giraffe

zeus in-front-of giraffe
zeus at-fringe-of tree
giraffe in-front-of tree
tree in-front-left-of giraffe

tree small
giraffe big
```

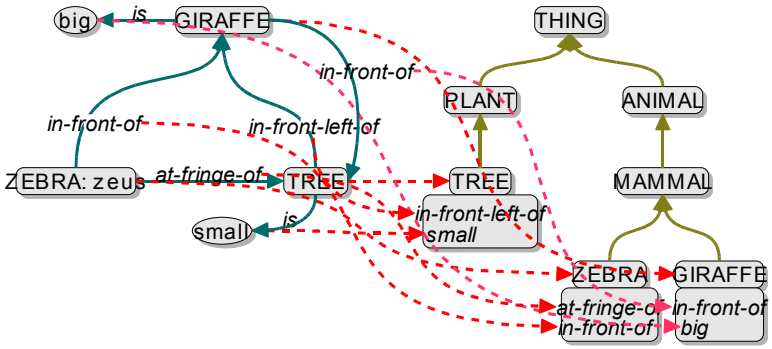
This intermediate text representation translates into a semantic network, which is graphically presented to the viewer:




Each object in the semantic network links to its corresponding concept in the table above. These concepts reside in an inheritance-based taxonomy of concepts from general to specific. This is also graphically presented to the viewer, and each node can be expanded to show more information:



The semantic network contains the explicitly stated details about the objects in the description. The knowledge base contains the implicit, background details about the concepts the objects refer to. The final element of the interpretation stage, which is graphically presented to the viewer, links the two to provide a fuller picture of the underlying meaning:

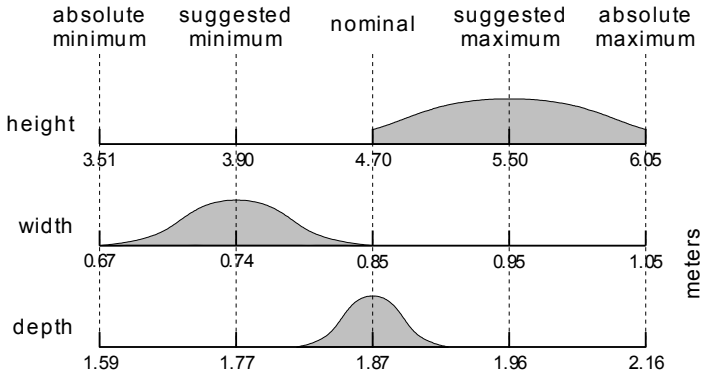



4. **Reason**

The reasoning stage evaluates the objects in the semantic network with respect to their conceptual definitions in the knowledge base, as shown by the dashed arrows above. The definitions fall into two categories.

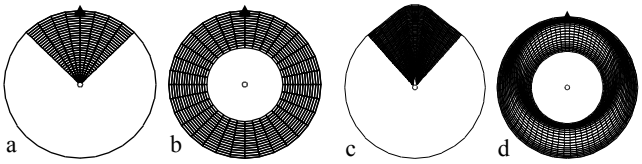
Dimension Definitions

Dimensions are defined as probability distributions over the height, width, and depth intervals for each object. A graphical definition editor is available to display and adjust these:

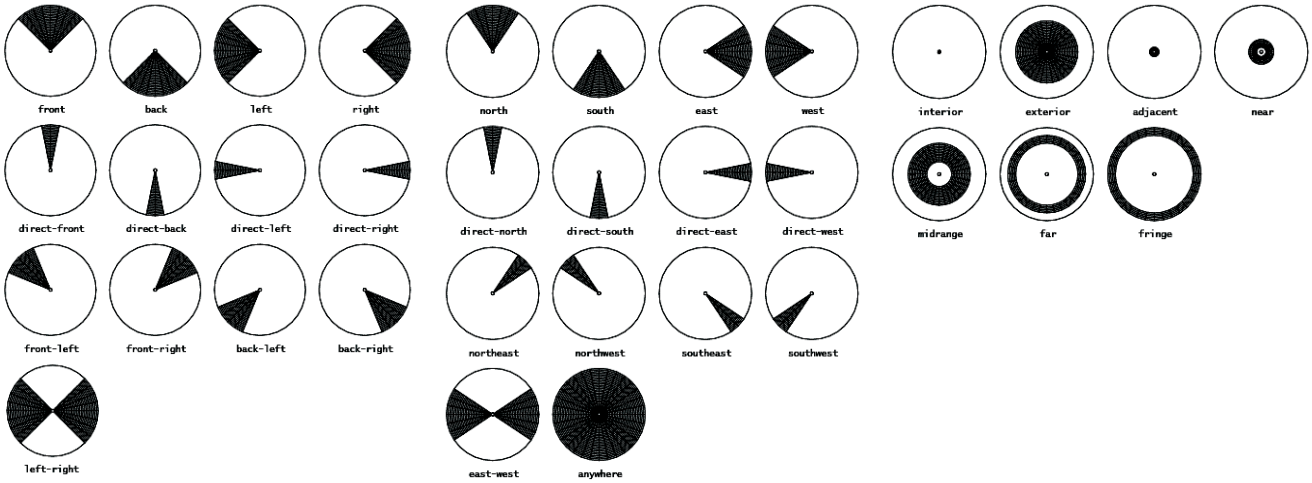



Position and Orientation Definitions

Position and orientation definitions use circular fields of sectors and rings to specify where other objects may appear with respect to the object in their center. Each field has two complementary parts, which specify the legal positions (a-b), and the preferred positions (c-d):

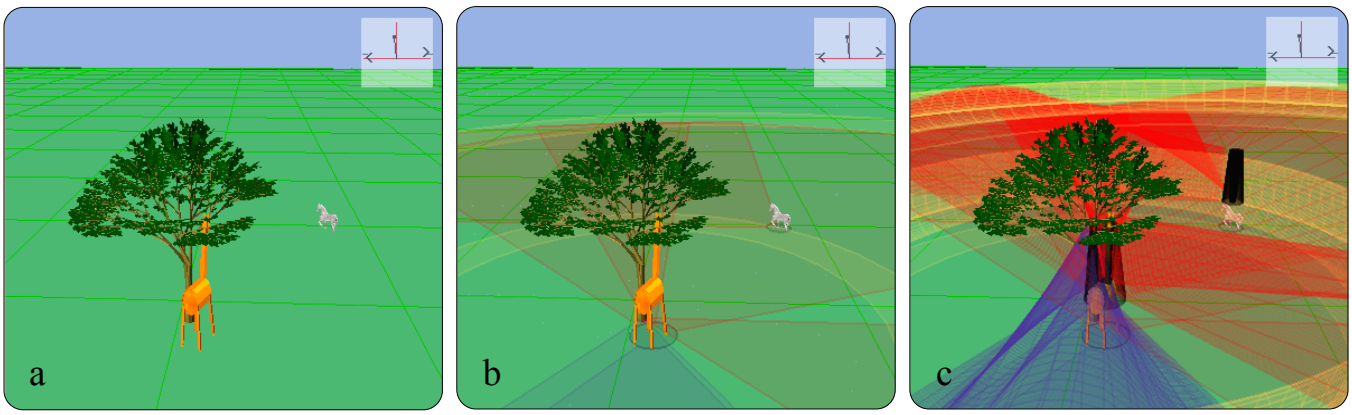



Most of the fields appear as follows:

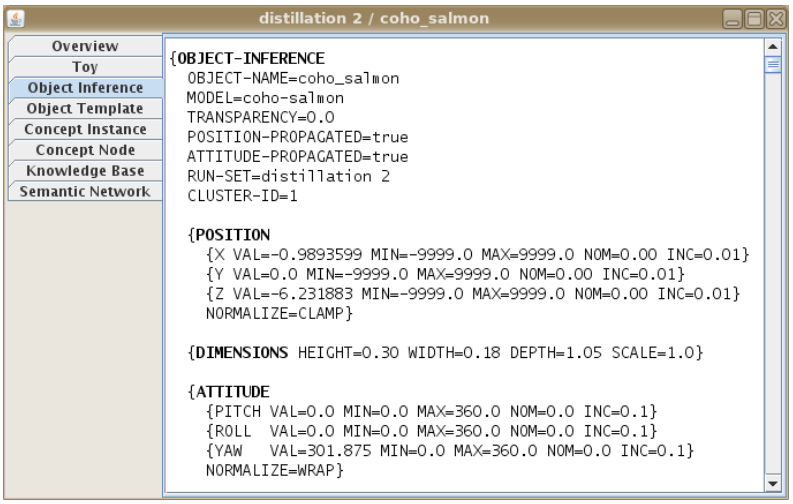


5. **Depict**

The visualization stage represents one or more interpretations (as specified in the vignette settings) for the text. The basic form (a) presents only the solution. Form (b) overlays the geometries, and form (c) overlays the topographies. The viewer can fly around to any vantage point and perspective.




Clicking on objects reveals extensive details of their underlying representations:




6+7. **Infer and Augment**

Once the positions, orientations, and dimensions have been determined, the spatial definitions associated with each object are applied in basically a reverse process to infer spatial information that was not present in the description. This mechanism of knowledge generation is extremely powerful. For example, the Zeus example has only 6 stated relationships, but this process reveals another 66. These can be back-propagated into the original semantic network to augment it with a wealth of additional knowledge, which is presented to the viewer:

tree southwest-of world-center
tree far-from world-center
tree local-in-front-of giraffe
tree local-in-front-left-of giraffe
tree global-in-back-of giraffe
tree global-directly-in-back-of giraffe
tree north-of giraffe
tree directly-north-of giraffe
tree outside giraffe
tree near giraffe
tree has-more-height giraffe
tree has-less-width giraffe
tree has-less-depth giraffe
tree local-in-front-of zeus
tree local-directly-in-front-of zeus
tree global-left-of zeus
tree west-of zeus
tree outside zeus
tree near zeus
tree has-more-height zeus
tree has-less-width zeus
tree has-less-depth zeus

giraffe south-of tree
giraffe directly-south-of tree
giraffe at-fringe-of tree
giraffe facing tree
giraffe has-more-width tree
giraffe has-more-depth tree
giraffe has-less-height tree
giraffe south-of world-center
giraffe far-from world-center
giraffe local-left-of zeus
giraffe local-in-front-left-of zeus
giraffe global-in-front-left-of zeus
giraffe southwest-of zeus
giraffe outside zeus
giraffe midrange-from zeus
giraffe facing zeus
giraffe directly-facing zeus
giraffe has-more-height zeus
giraffe has-more-width zeus
giraffe has-more-depth zeus

zeus east-of tree
zeus at-fringe-of tree
zeus facing tree
zeus directly-facing tree
zeus has-more-width tree
zeus has-more-depth tree
zeus has-less-height tree
zeus south-of world-center
zeus far-from world-center
zeus local-in-front-of giraffe
zeus local-directly-in-front-of giraffe
zeus global-in-back-right-of giraffe
zeus northeast-of giraffe
zeus outside giraffe
zeus near giraffe
zeus has-less-height giraffe
zeus has-less-width giraffe
zeus has-less-depth giraffe

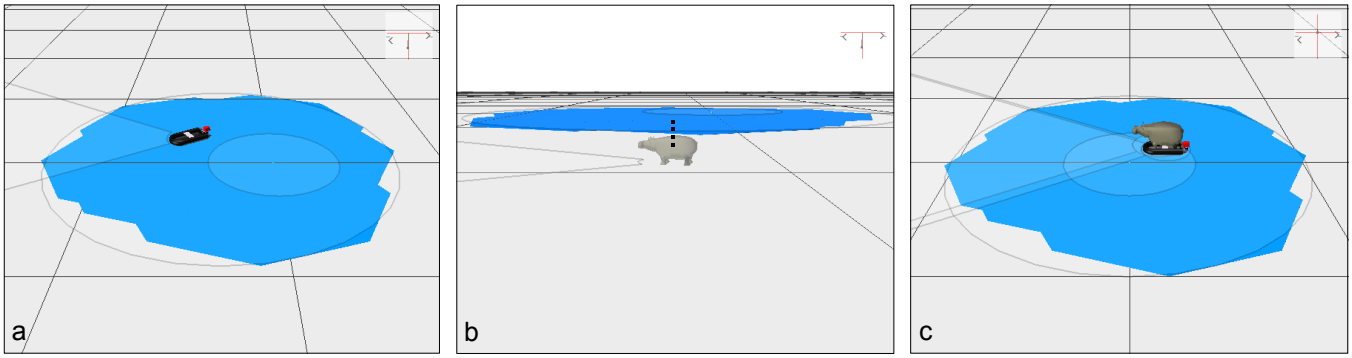
world-center global-in-back-of giraffe
world-center north-of giraffe
world-center at-fringe-of giraffe
world-center global-in-back-of zeus
world-center north-of zeus
world-center at-fringe-of zeus

**Example Visualizations**

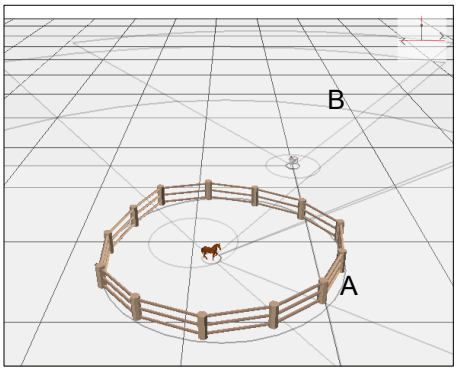
These screenshots have their contrast set for printing; the demo variants (like 5a-c) are much more colorful.

a) *The raft is in the lake*
b) *The hippo is in the lake*
c) *The hippo is in the raft, and the raft is in the lake*

(showing the effects of context on the relation *in*)

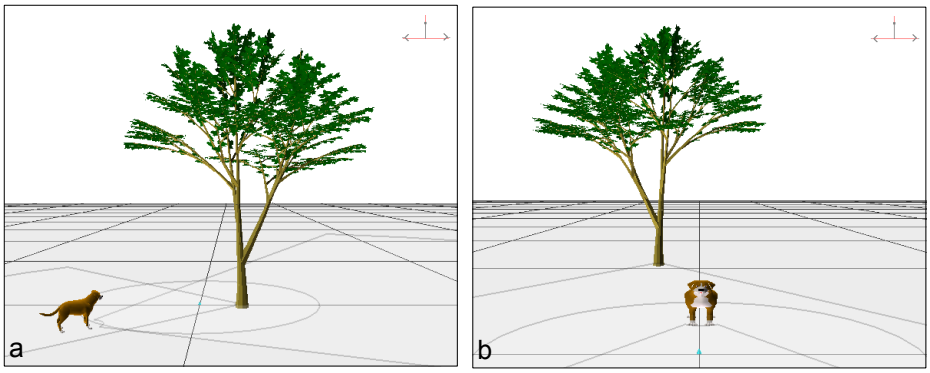



a) *The horse is inside the corral*
b) *The zebra is outside the corral*

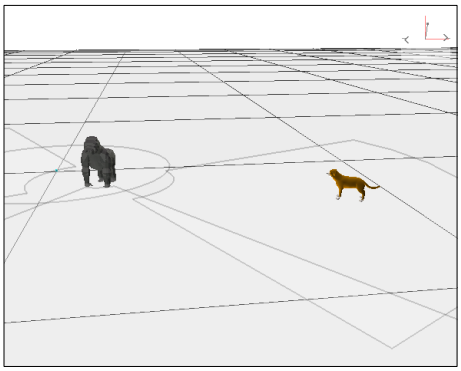



a) *The tree is in front of the dog*
b) *The dog is in front of the tree*

(showing the difference between intrinsic and deictic interpretations)

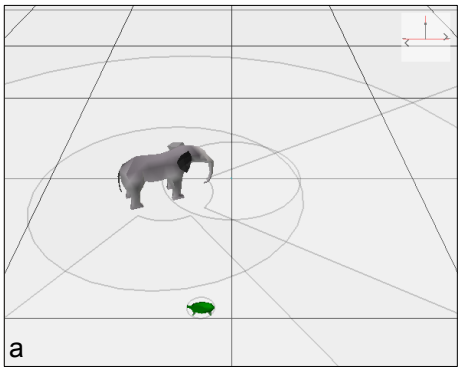


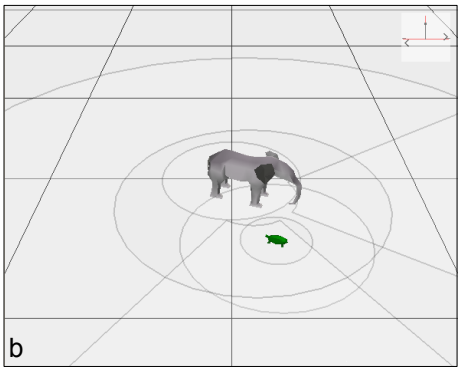
*The dog is to the side of the gorilla*



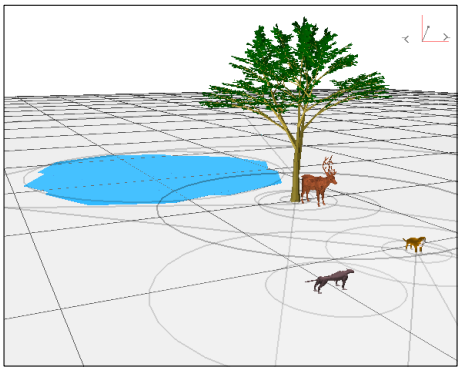
a) *The turtle is near the elephant*
b) *The elephant is near the turtle*

(showing the effects of context and size on the relation *near*)

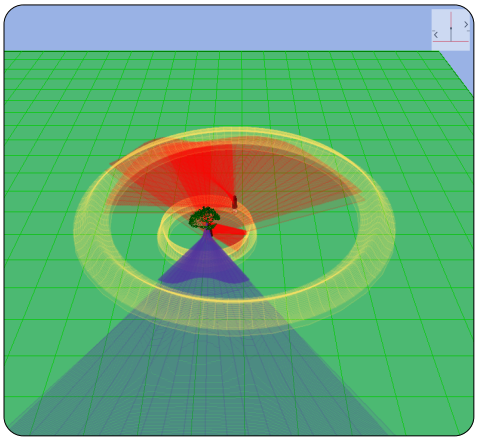


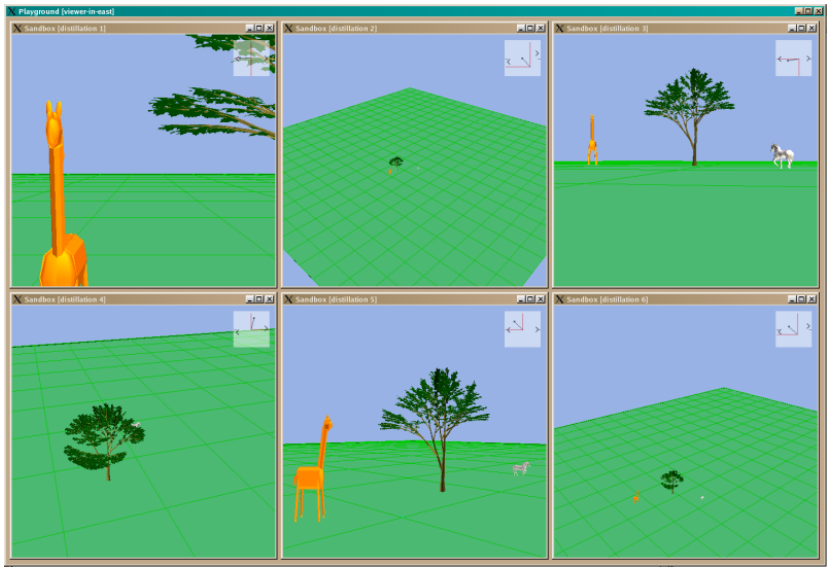




*The dog is south of the tree and near the panther; the panther is to the right of the dog; and the elk is near the maple tree and midrange from and facing away from the pond*



An overview and collage of the Zeus example:





Summary

Knowledge-Based Spatial Reasoning for Scene Generation from Text Descriptions

Dan Tappan

Idaho State University

This system translates basic English descriptions of a wide range of objects in a simplistic zoo environment into plausible, three-dimensional, interactive visualizations of their positions, orientations, and dimensions. It combines a semantic network and contextually sensitive knowledge base as representations for explicit and implicit spatial knowledge, respectively. Its linguistic aspects address underspecification, vagueness, uncertainty, and context with respect to intrinsic, extrinsic, and deictic frames of spatial reference. The underlying, commonsense reasoning formalism is probability-based geometric fields that are solved through constraint satisfaction. The architecture serves as an extensible test-and-evaluation framework for a multitude of linguistic and artificial-intelligence investigations.

Hardware and Software Requirements

I will provide my own laptop with all the required elements for the demo.