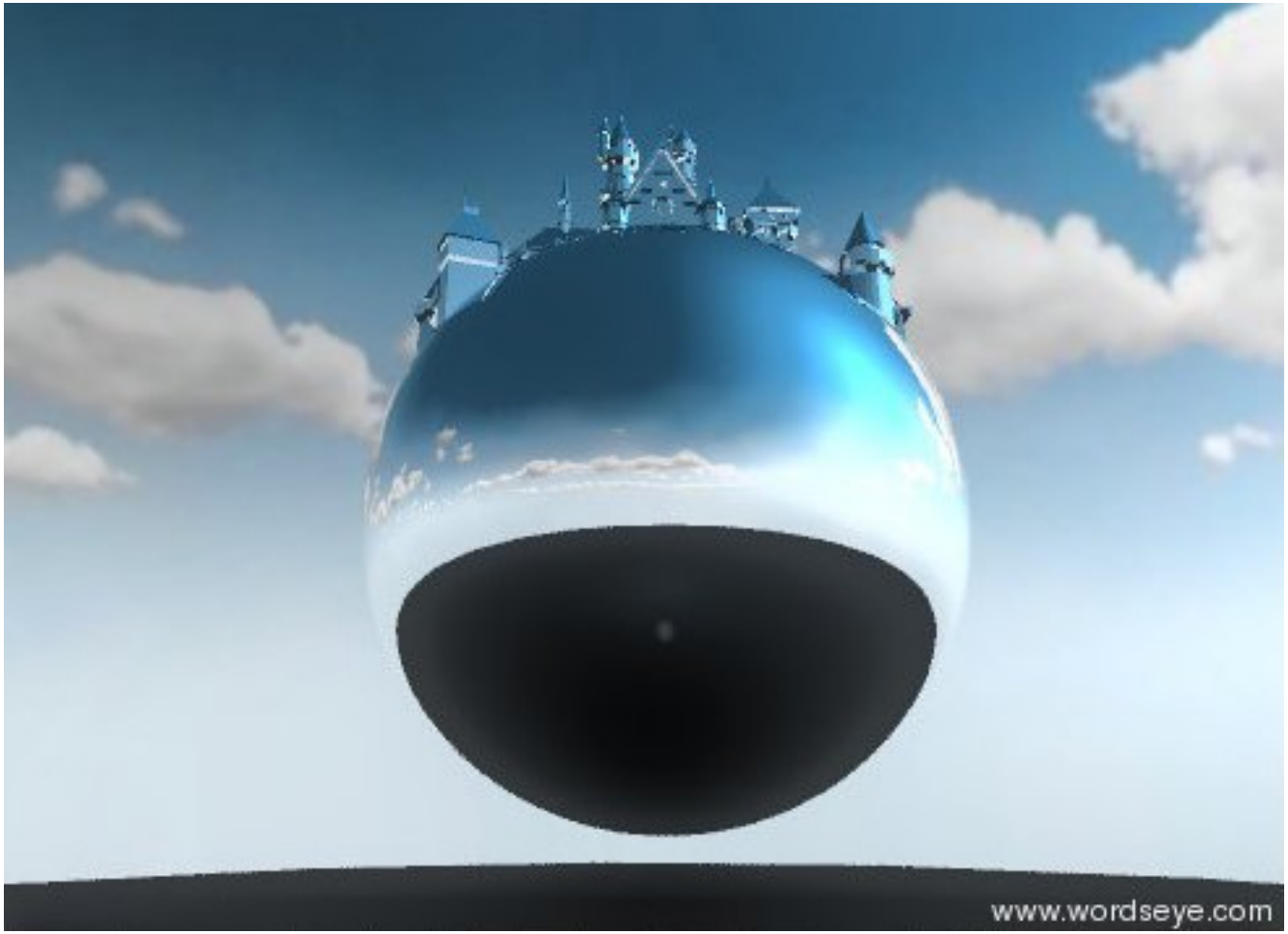
WordsEye: A system for converting natural language text to 3D scenes

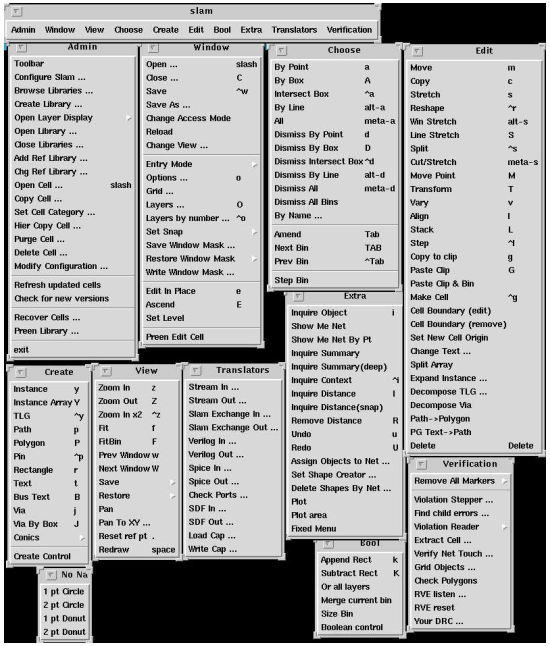


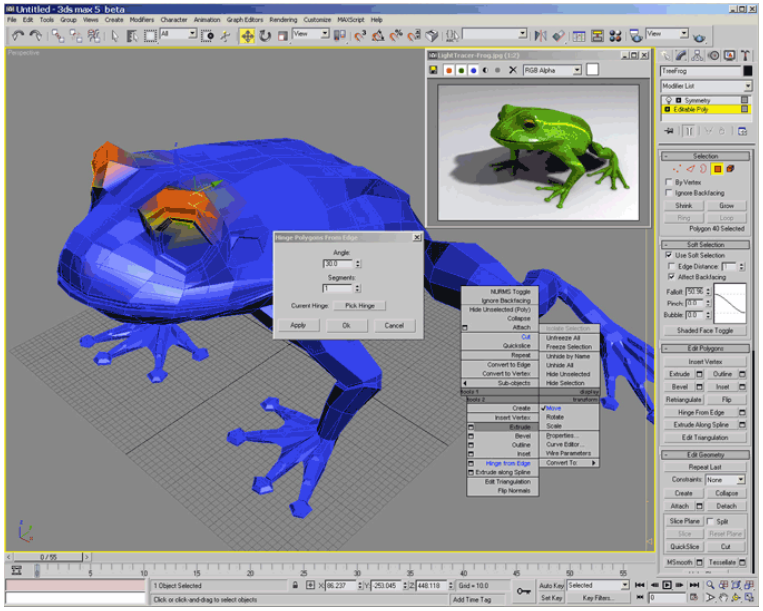
Bob Coyne (coyne@cs.columbia.edu)
Text2Scene group: Julia Hirschberg, Richard Sproat, Owen Rambow, Daniel Bauer, Masoud Rouhizadeh, Cecilia Schudel, Alex Klapheke

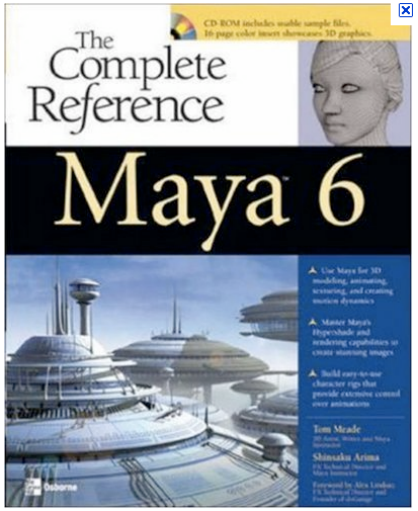
Outline

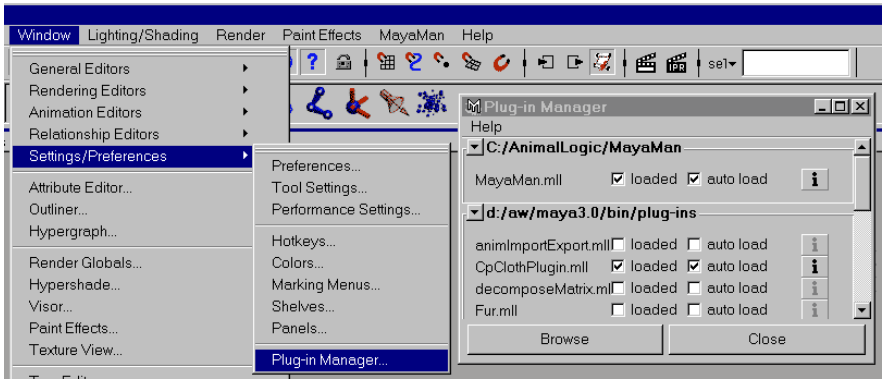
- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

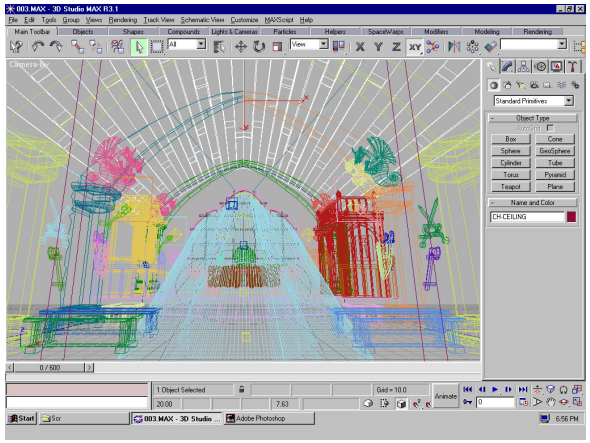
Untapped potential – 3D graphics is compelling but hard to create







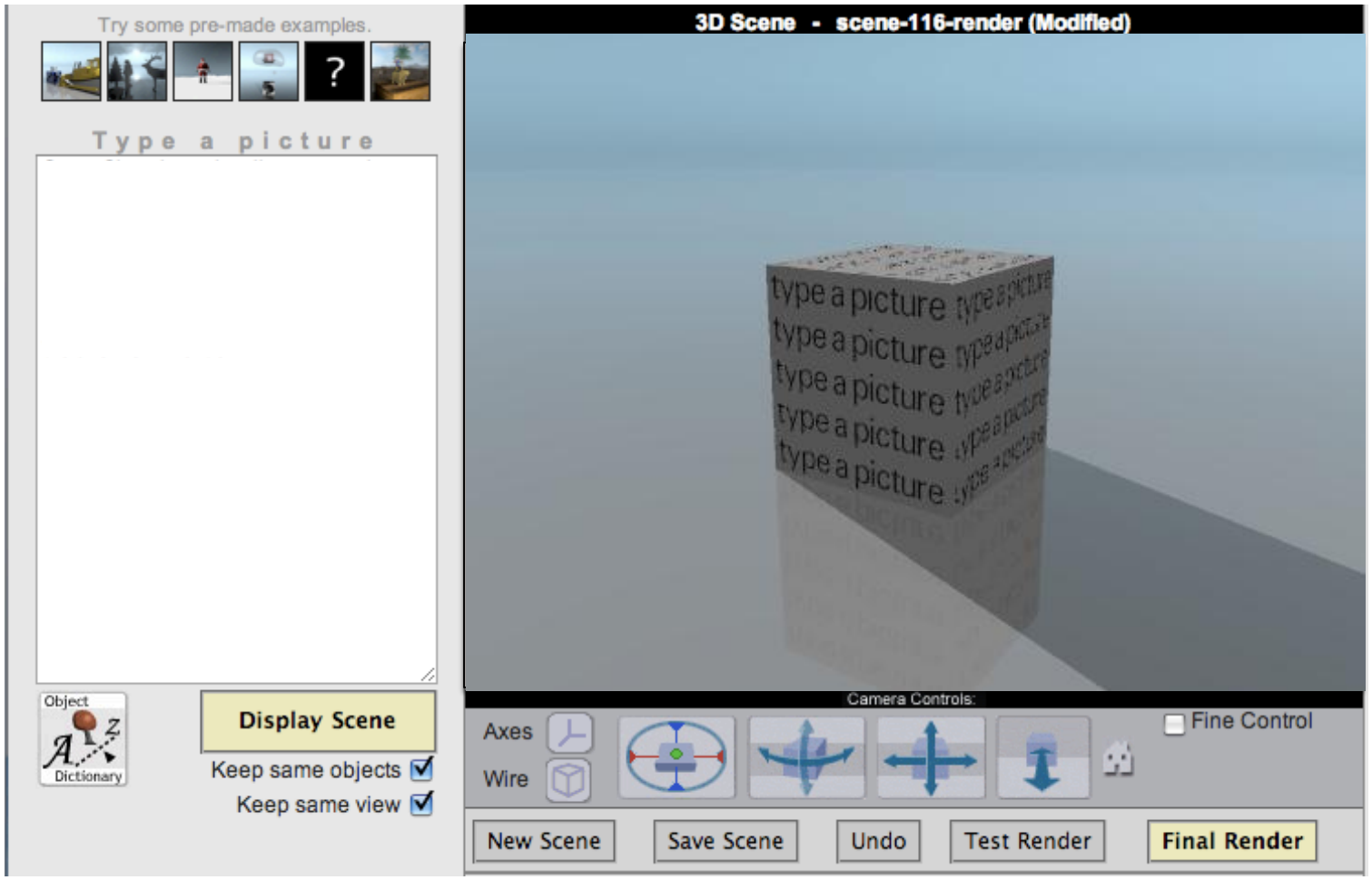




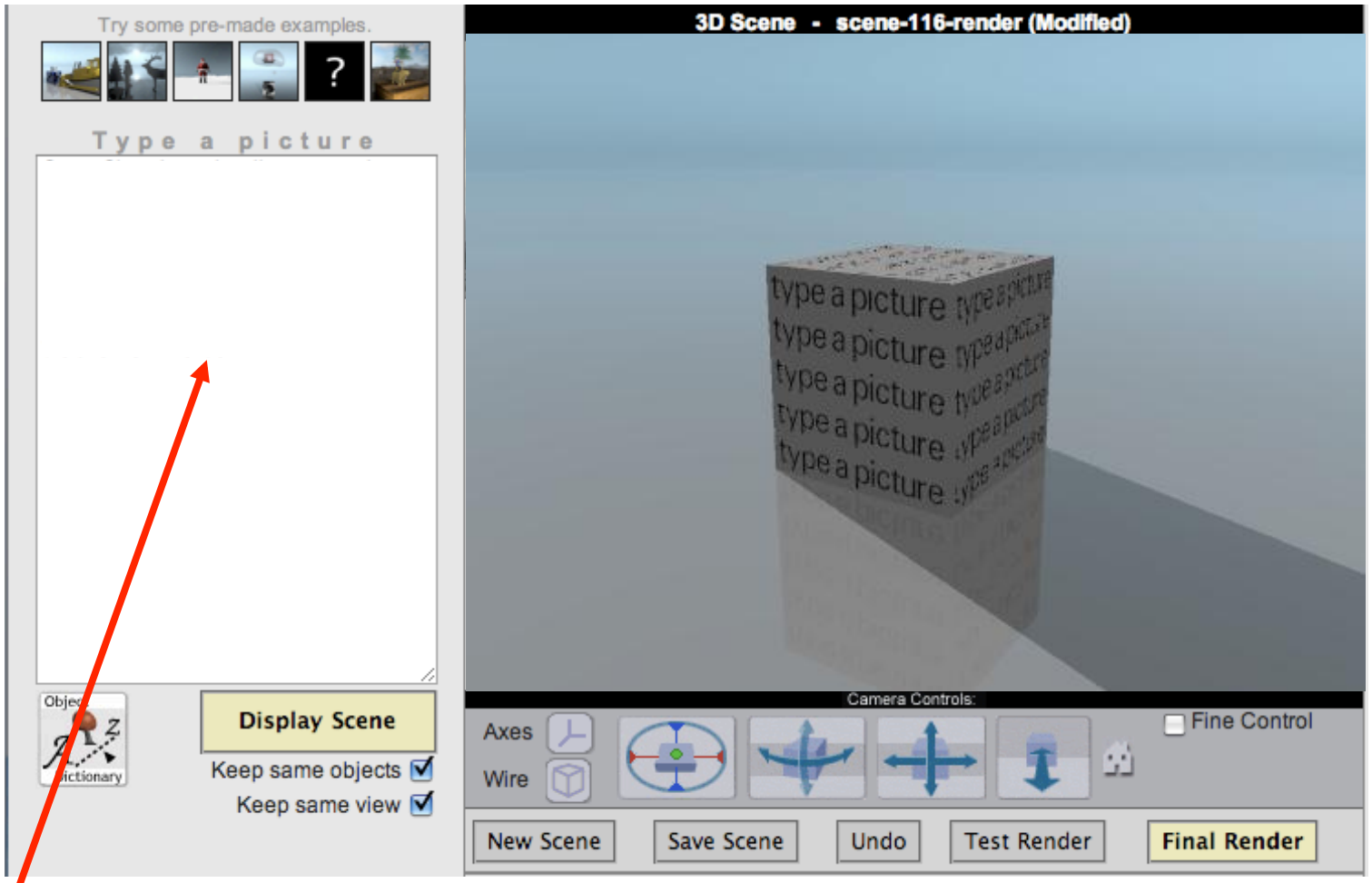
New Approach - Create 3D scenes with language

Santa Claus is on the white mountain range. he is blue. it is cloudy. the large yellow illuminator is in front of him. the alien is in front of him. the mountain range is silver.

WordsEye: Create 3D scenes with language

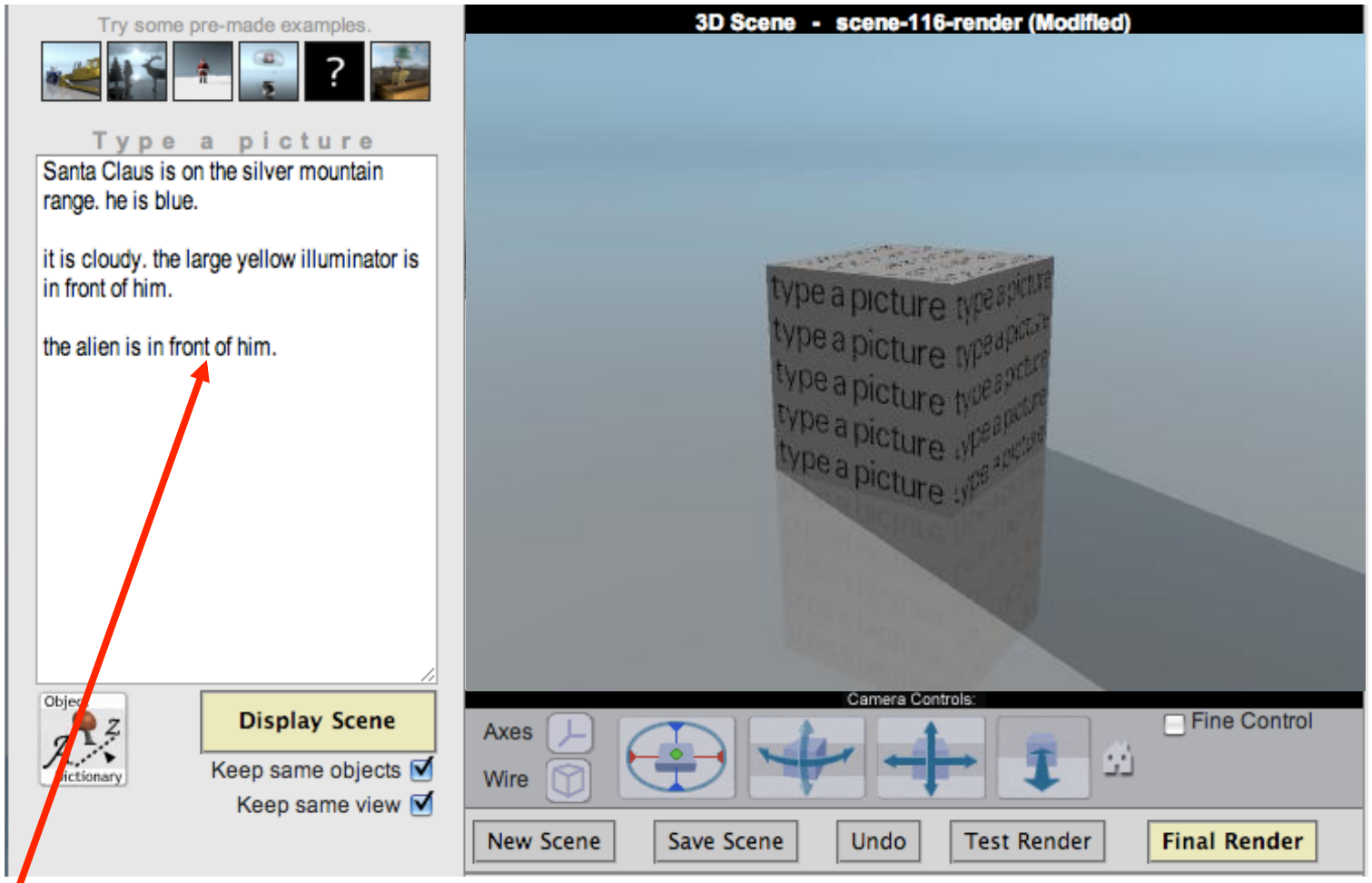


WordsEye: Create 3D scenes with language



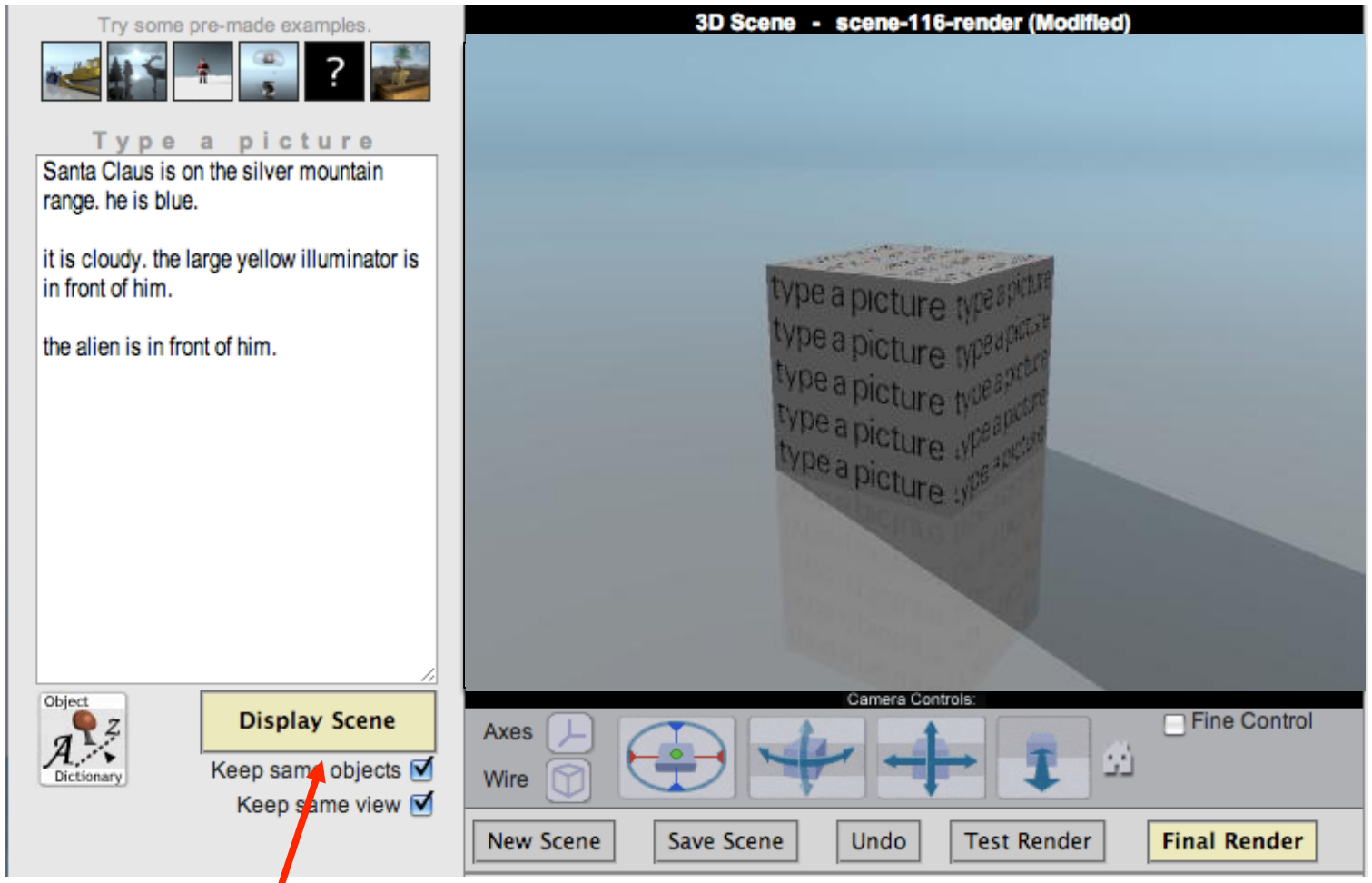
1 **Describe** a scene

WordsEye: Create 3D scenes with language



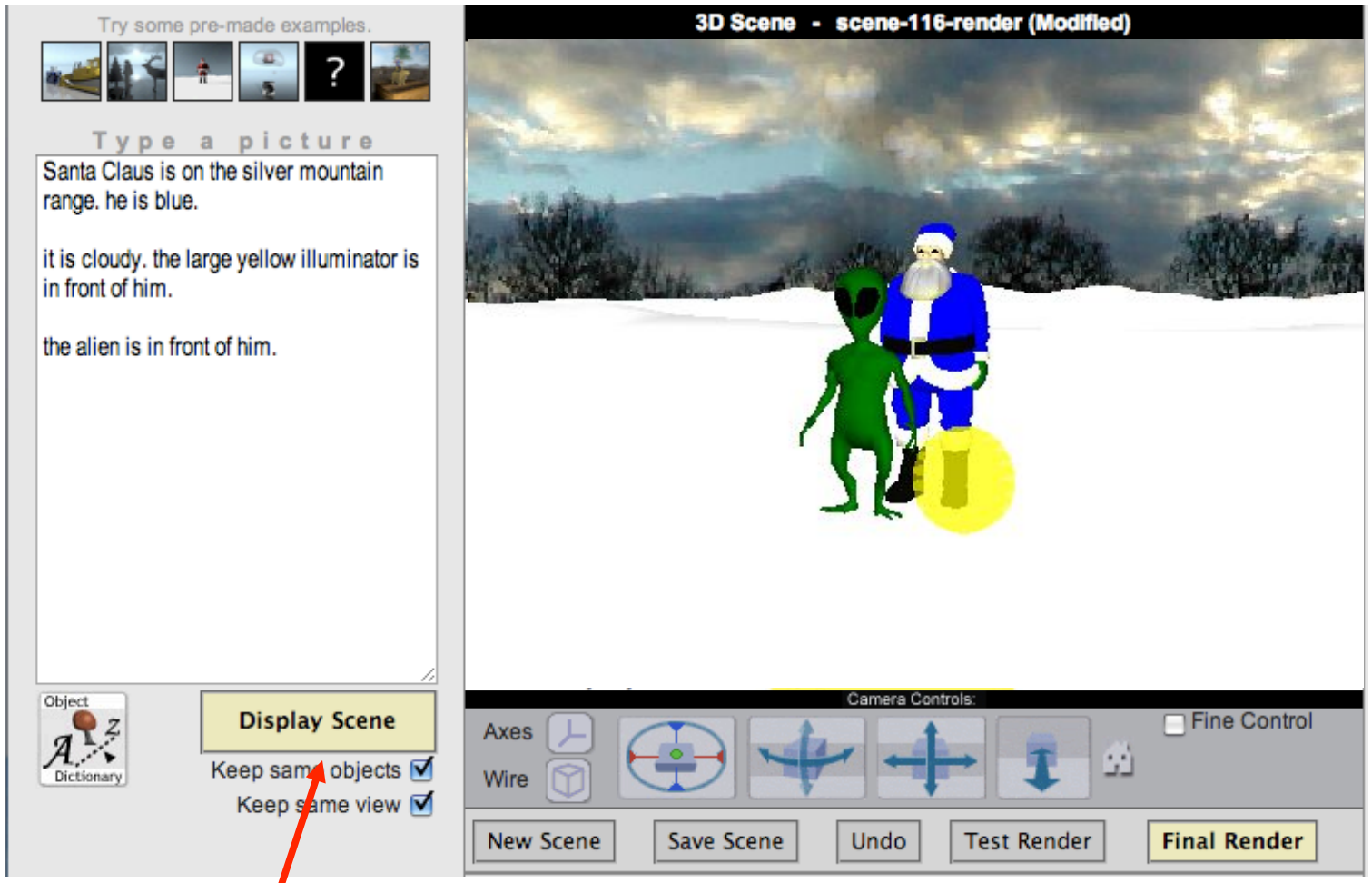
1 **Describe** a scene

WordsEye: Create 3D scenes with language



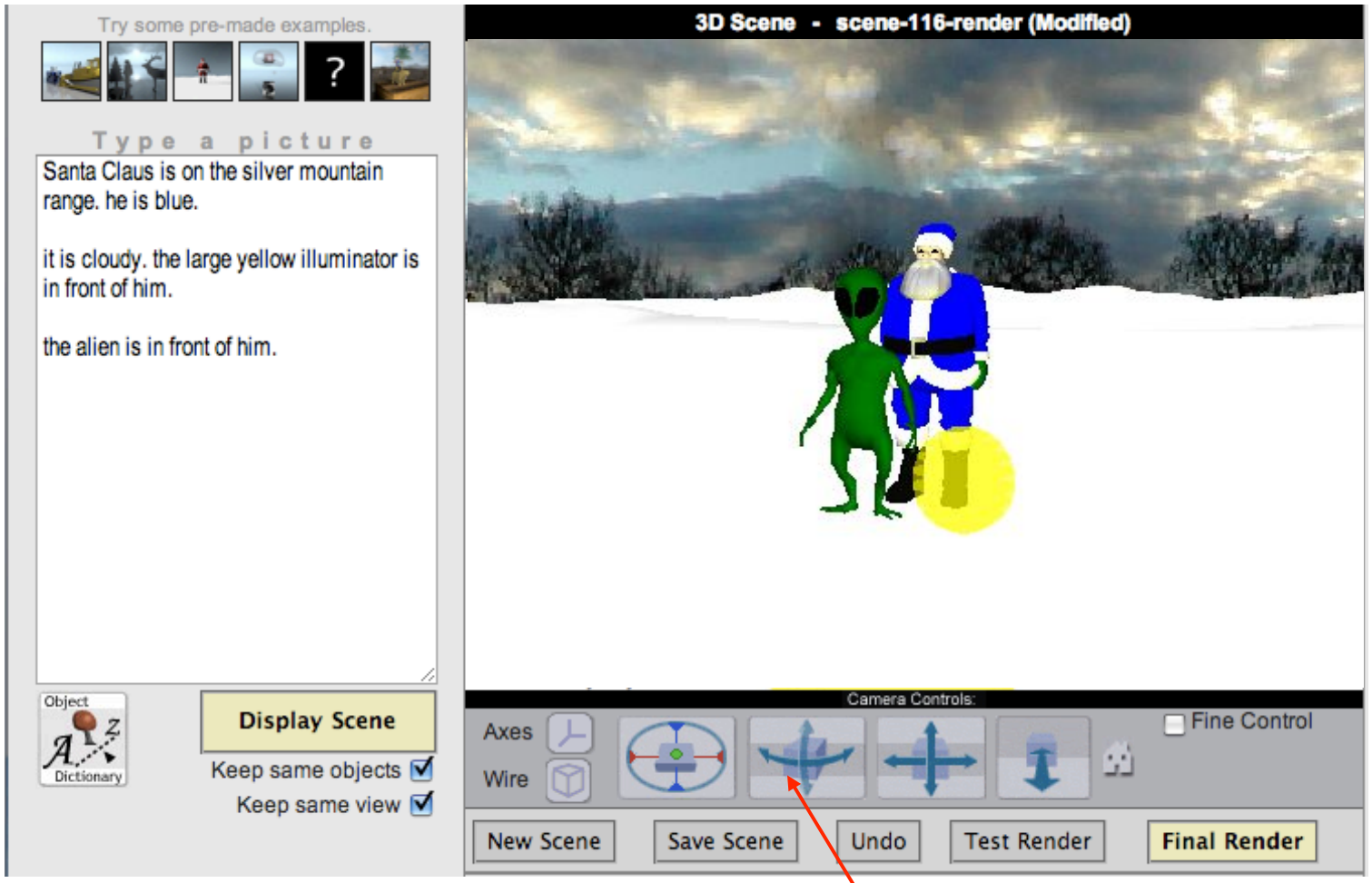
2 Click **Display**

WordsEye: Create 3D scenes with language



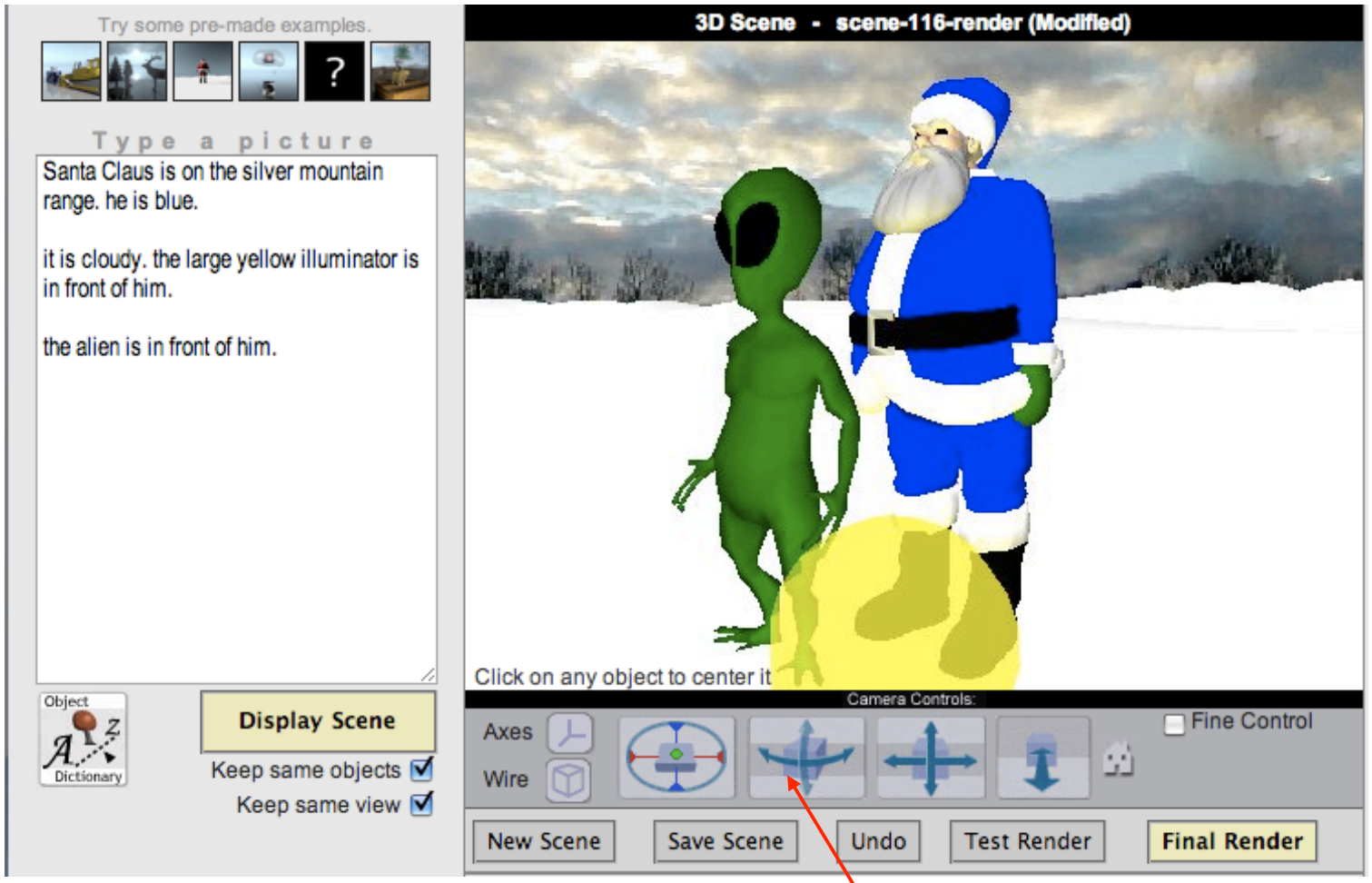
2 click **Display**

WordsEye: Create 3D scenes with language



3 Change 3D viewpoint using camera controls

WordsEye: Create 3D scenes with language

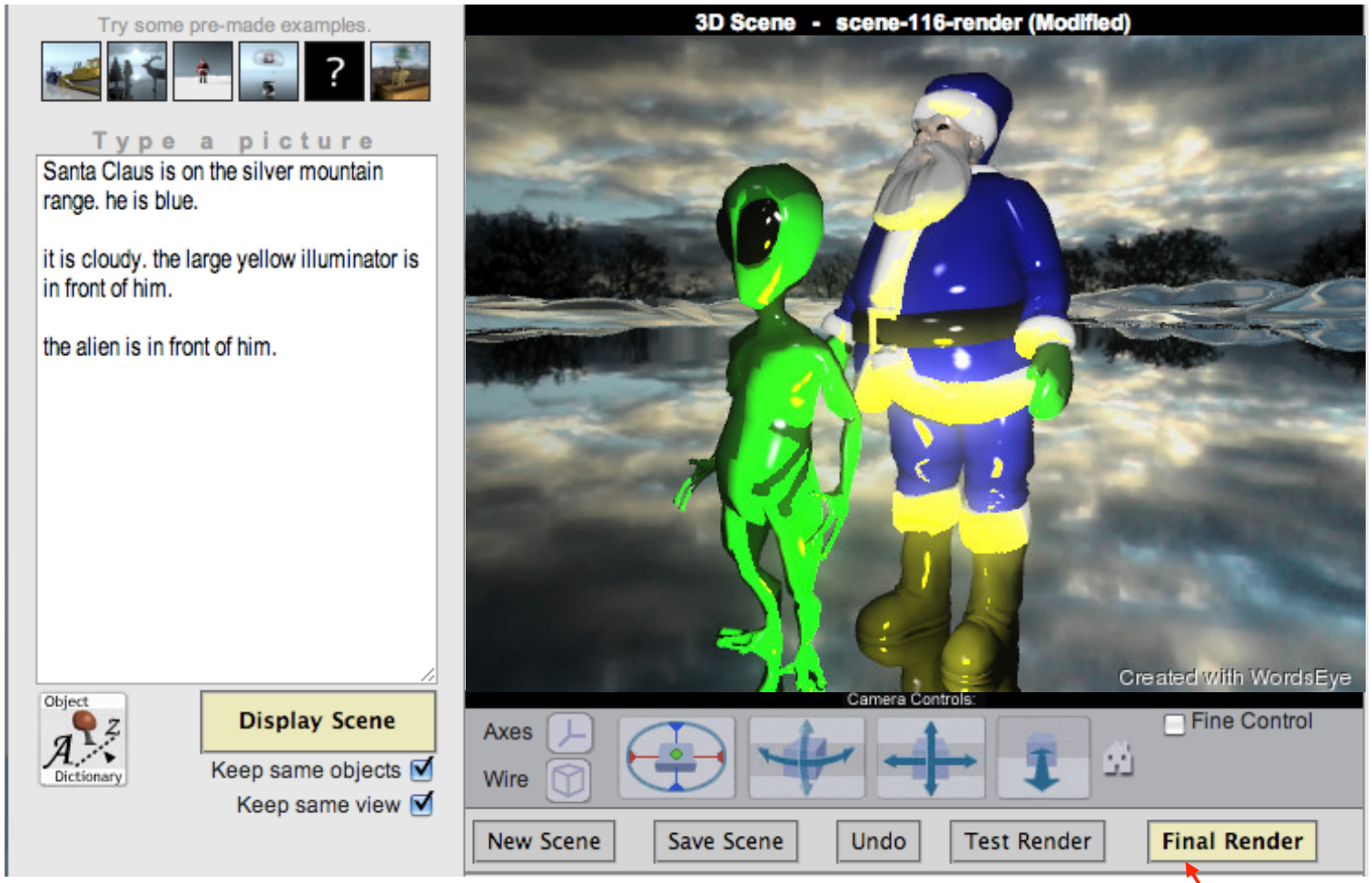


3 Change 3D viewpoint using camera controls

WordsEye: Create 3D scenes with language

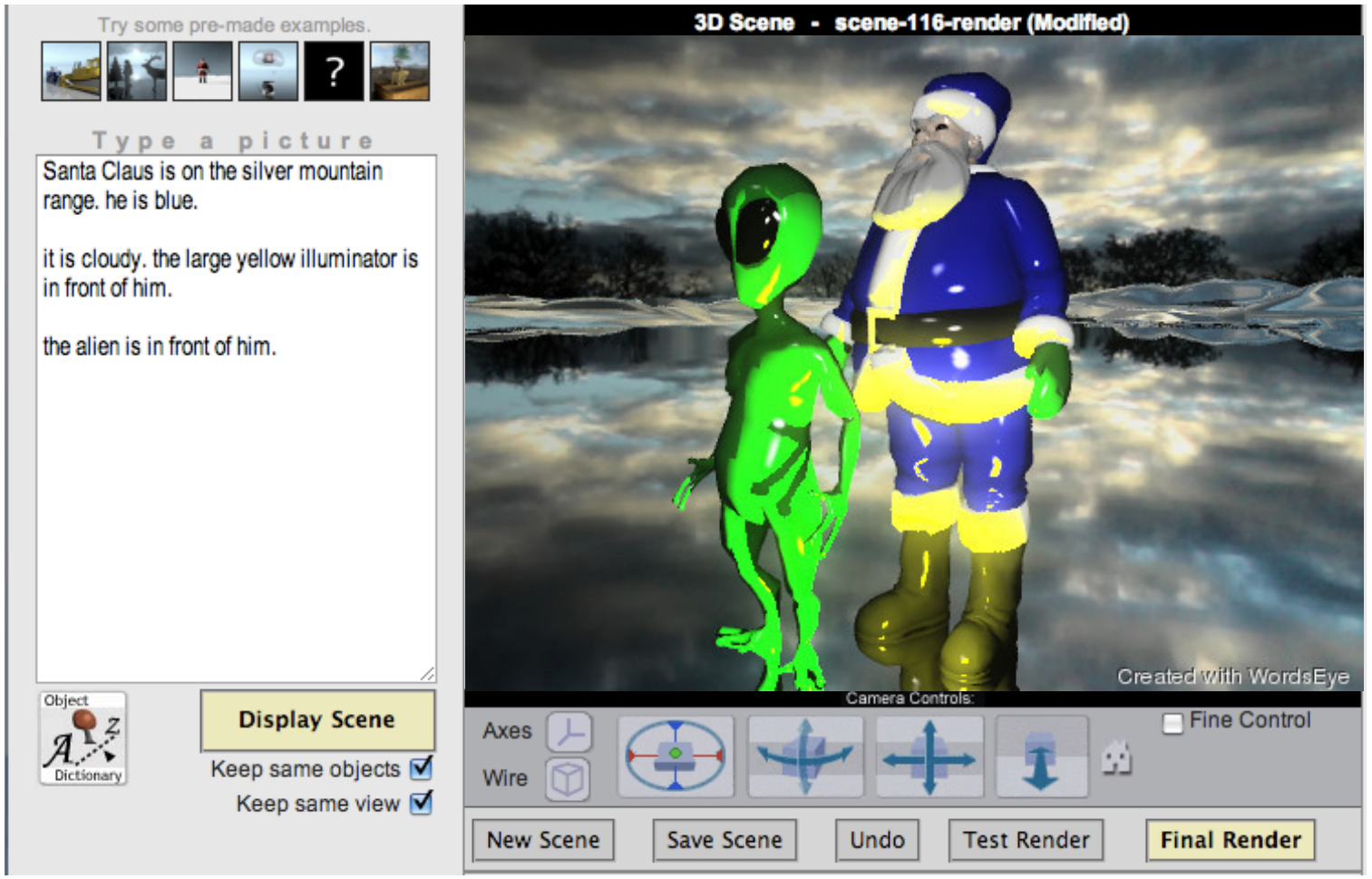
4 Perform final render to add reflections, shadows, etc.

WordsEye: Create 3D scenes with language



4 Perform final render to add reflections, shadows, etc.

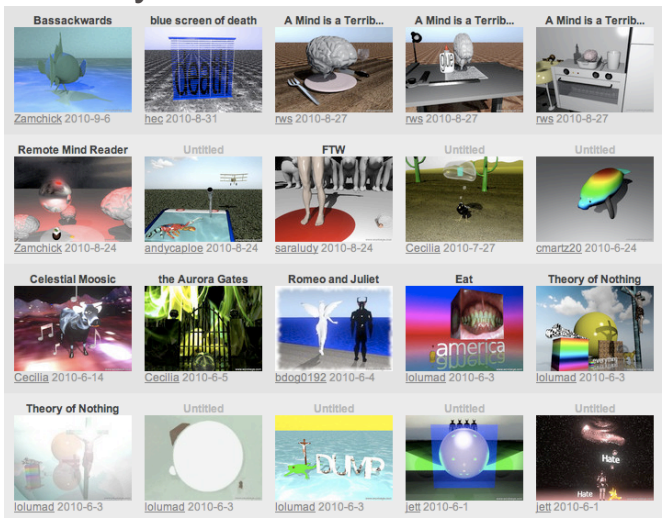
WordsEye: Create 3D scenes with language

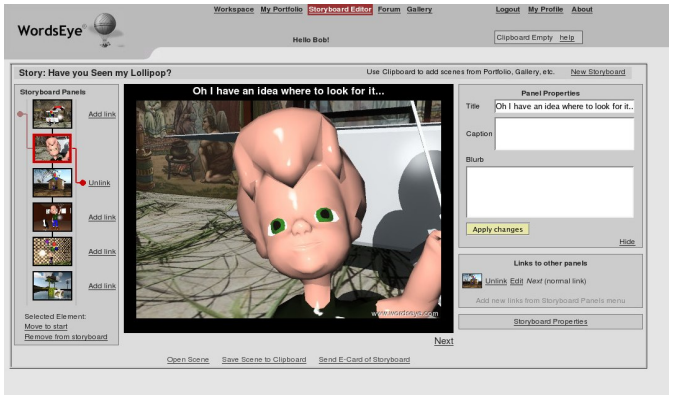


5 Final Rendered picture can be put in online Gallery
Or linked with other pictures to form a Picturebook

Online gallery and picturebooks

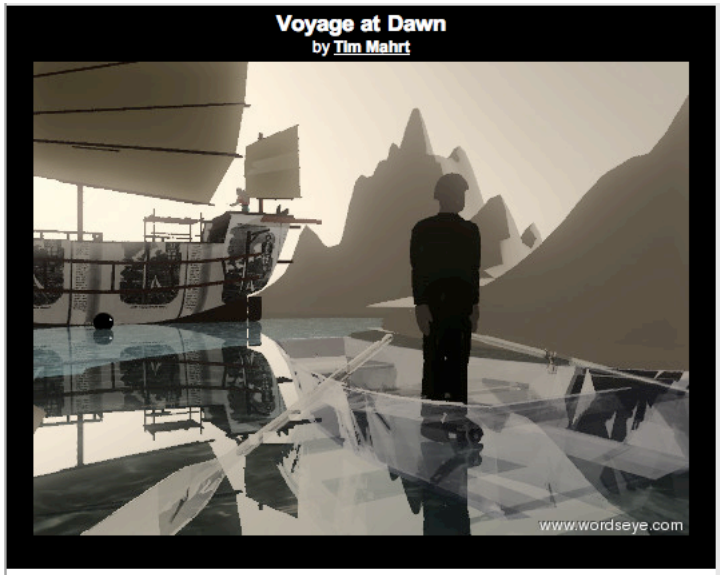
Gallery





Picturebook editor

User comments



Generated from text: It is morning.The sky is tan. The ground is water. A very large boat is in the ground. The boat is glass. A man is in the boat. The camera is in front of the man. A tall mountain is on the left of the boat. The mountain is tan A tall mountain is on the right of the boat. The mountain is tan. A tall mountain is fifty feet behind the boat. The mountain is tan. A large pirate boat is fifty feet behind the boat. It is facing West. A pirate is on the boat. A bird is above the boat. A large sphere is in front of the boat. The sphere is black

Open in Workspace Copy to Clipboard Send E-Postcard

coyne -- 1/27/2006

Wow...this is really cool!

Zamchick -- 1/27/2006

Is there a bird on the shoulder of that pirate :-)

rws -- 1/29/2006

Tim is a student in my intro computational linguistics class. He's interested in NLP and graphics.

This shows how quickly someone can get up to speed if they are motivated. We just have to write off the speech engineer types who try three things and conclude it doesn't work.

coyne -- 1/30/2006

right. It's great to see people do stuff like this right off the bat.

I agree motivation is the key factor (as it is with any creative tool). We might never be able to make the system do enough to satisfy the reluctant user who has little creative urge or motivation to learn.

rws -- 1/31/2006

Tim told me today that this picture took him 15 minutes or so to make. That's 15 minutes from his first ever access to the system, including figuring out what the system can and cannot do, and then producing the scene.

coyne -- 1/31/2006

Good datapoint.

Zamchick -- 1/31/2006

Where's Tim's second picture?

rws -- 1/31/2006

He's busy doing homework.

Zamchick -- 1/31/2006

Ease up on the wordartist, the wordseyeon, the literfitti artist, the Graffonic fellow!

Outline

- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

What is needed to depict a scene?

- 3D Objects and images
- Spatial relations
- Surface properties: color, texture, transparency, etc
- Poses, facial expressions, and other shape changes
- Compound objects (e.g. living room) composed of many objects

And then… tie everything to language!

Graphical library: 2,200 3D objects and 10,000 images – all tied to words

3D objects

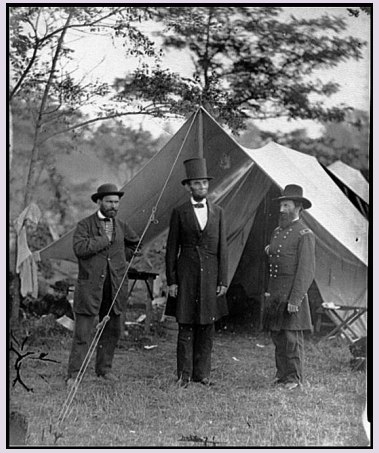
3D Object 	**Associated words:** weight bench, sportsware, bench **Size:** 6.0 feet	3D Object 	**Associated words:** furniture, television, furnishing, appliance, display device, communication device **Size:** 4.0 feet
3D Object 	**Associated words:** furniture, television, furnishing, appliance, display device, communication device **Size:** 3.0 feet	3D Object 	**Associated words:** end table, compound, compound object, table, group **Size:** 3.0 feet
3D Object 	**Associated words:** coffee table, table **Size:** 3.5 feet	3D Object 	**Associated words:** table, furniture **Size:** 3.0 feet
3D Object 	**Associated words:** kitchen table, patio table, table **Size:** 5.0 feet	3D Object 	**Associated words:** dining room table, table **Size:** 8.0 feet
3D Object 	**Associated words:** table, furniture **Size:** 3.0 feet	3D Object 	**Associated words:** bar stool, stool **Size:** 3.0 feet

2D Images and textures

B&W drawings

Texture Maps

Artwork



Photographs

Objects by another name (reference resolution)

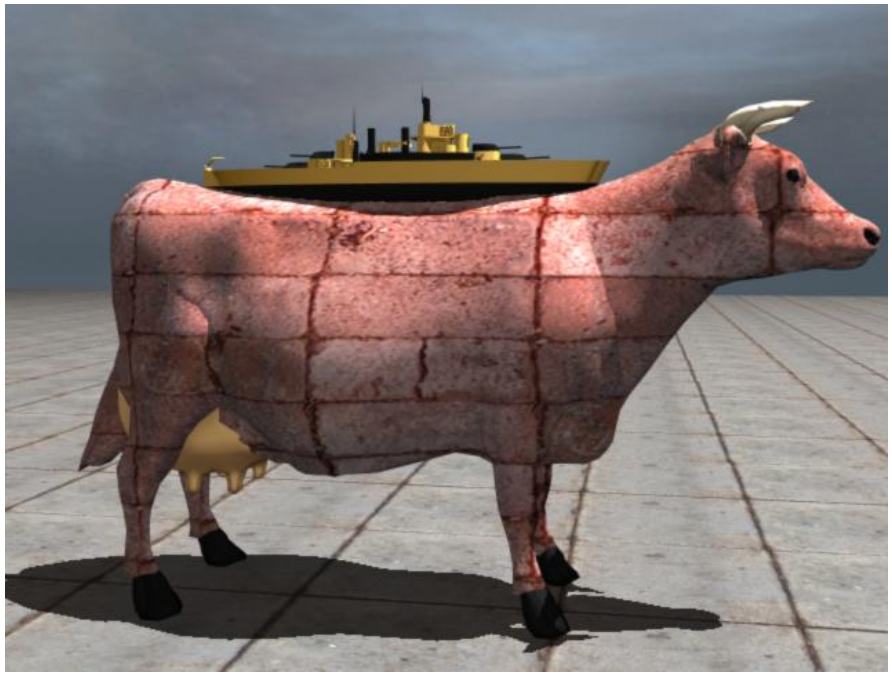
Anaphora and co-reference

*The duck is in the sea. **It** is upside down. The sea is shiny and transparent. The apple is 3 inches below the duck. **It** is in front of the duck. **It** is partly cloudy.*

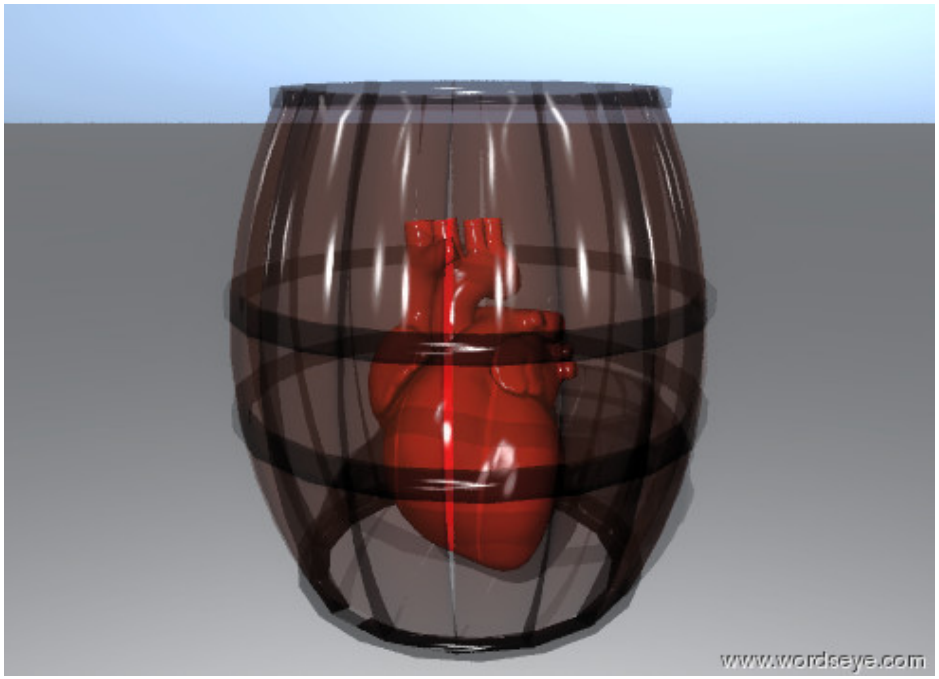
Attribute reference

*Three dogs are on the table. The **first** dog is blue. The **first** dog is 5 feet tall. The **second** dog is red. The **third** dog is purple.*

Attributes: size, color, transparency, texture



The orange battleship is on the brick cow. The battleship is 3 feet long



The red heart is in the tiny transparent barrel.

Poses and facial expressions

*The clown is **running**. …*

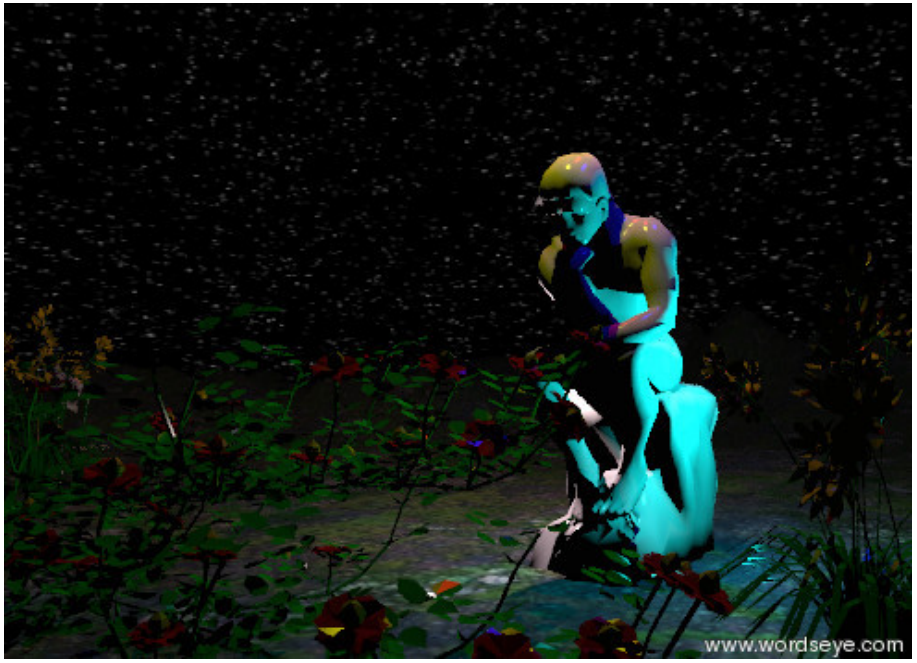
*Obama is **afraid** and **angry**….*

Put yourself in scenes

Obama is furious and surprised. He is to the right of Owen. Bob is behind Obama. The oil derrick is 10 feet behind Obama. It is in the ocean.

Environmental attributes: Time of day, cloudiness, lighting

the big palm tree is on the very large white sandy island. a palm tree is next to the big palm tree. the island is on the sea. The sun is pink. it is dawn. it is partly cloudy. The huge silver rocket is 20 feet above the sea…



The 7 enormous flowers are in front of the statue. It is midnight. The statue is 40 feet tall. The statue is on the mountain range. The 5 huge bushes are behind the mushroom. . . .

Full scene with several graphical constraints

*The **very large silver** ball is **on** the table. the ground is **shiny**. The table is **under** the **small** willow tree. The lion is **one foot in front of** the table. the lion is **facing** the ball. it is **cloudy**.*

Outline

- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

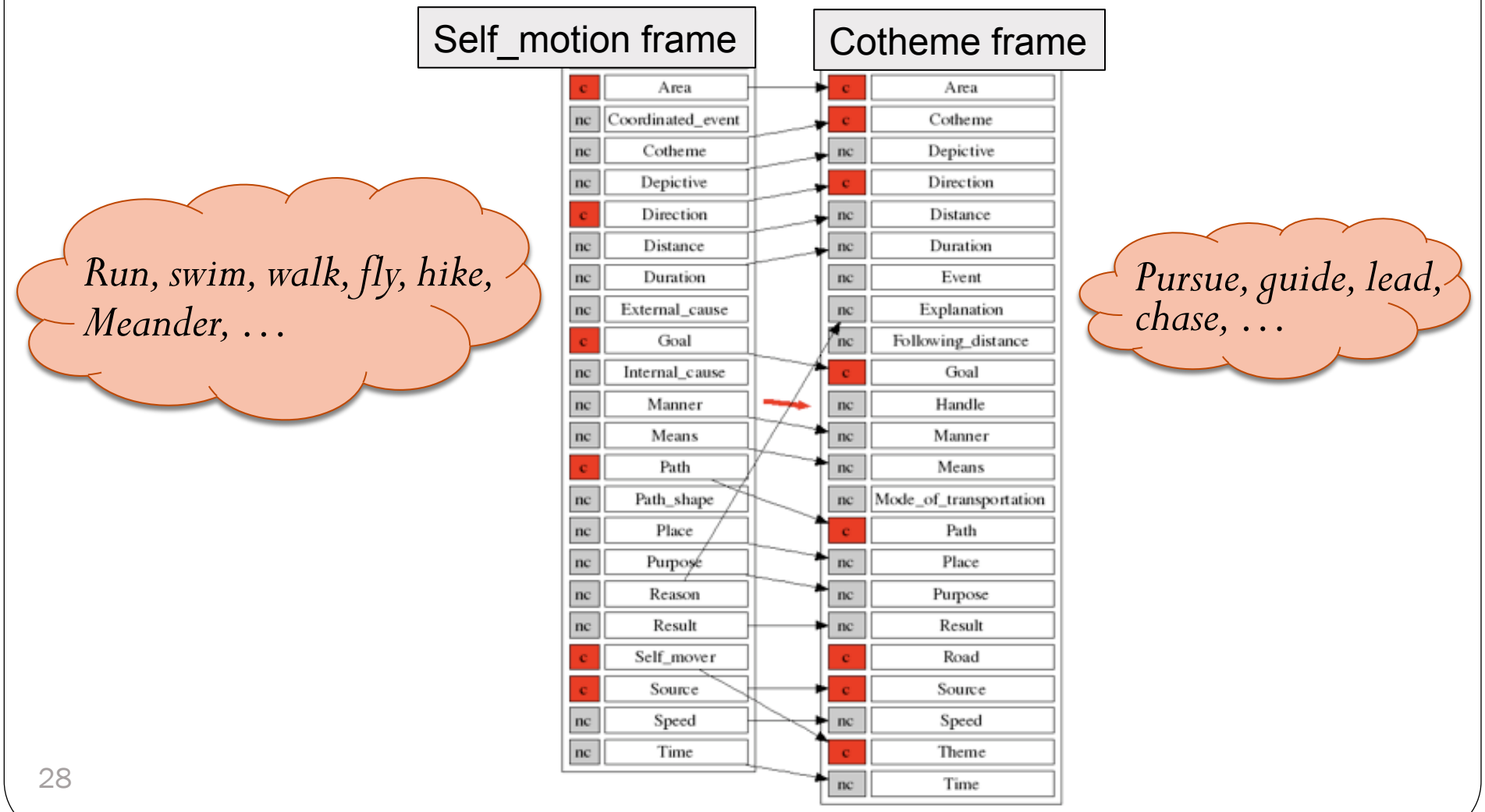
Lexicon – WordNet

- Online http://wordnetweb.princeton.edu/perl/webwn
- 120,000 Lexical items (synsets)
  - **hypernym/hyponym** *(Poodle* is a *Dog, Dog* is a *Canine)*
  - **Meronyms/holonyms** *(Birds* have *wings)*
- Multiple word senses. E.g. *Coffee*
  - <NOUN-12662772: ("coffee" "coffee tree") [hyper=("tree")]>
  - <NOUN-7929519: ("coffee" "java") [hyper=("beverage" "drink" "drinkable" "potable")]>
  - <NOUN-7929351: ("coffee bean" "coffee berry" "coffee") [hyper=("seed")]>
  - <NOUN-4972451: ("chocolate" "coffee" "deep brown" "umber" "burnt umber") [hyper=("brown" "brownness")]>
- Connect 3D objects to lexicon

Verb Semantics with FrameNet

- Online at http://framenet.icsi.berkeley.edu/
- **Frame**
  - Schematic representation of a situation, object, or event that provides the background and motivation for the existence and everyday use of words in a language. i.e. grouping of words with common semantics.
  - 1,000 frames and 10,000 words in those frames
  - **Frame Elements** (FEs): frame-based roles (agent, instrument, …)
- E.g. COMMERCE_SELL (*sell, vend*)
  - Core FEs (BUYER, GOODS, SELLER) ,
  - Peripheral FEs (TIME, LOCATION, MANNER, …)

Frames, roles, frame-to-frame relations



Syntax

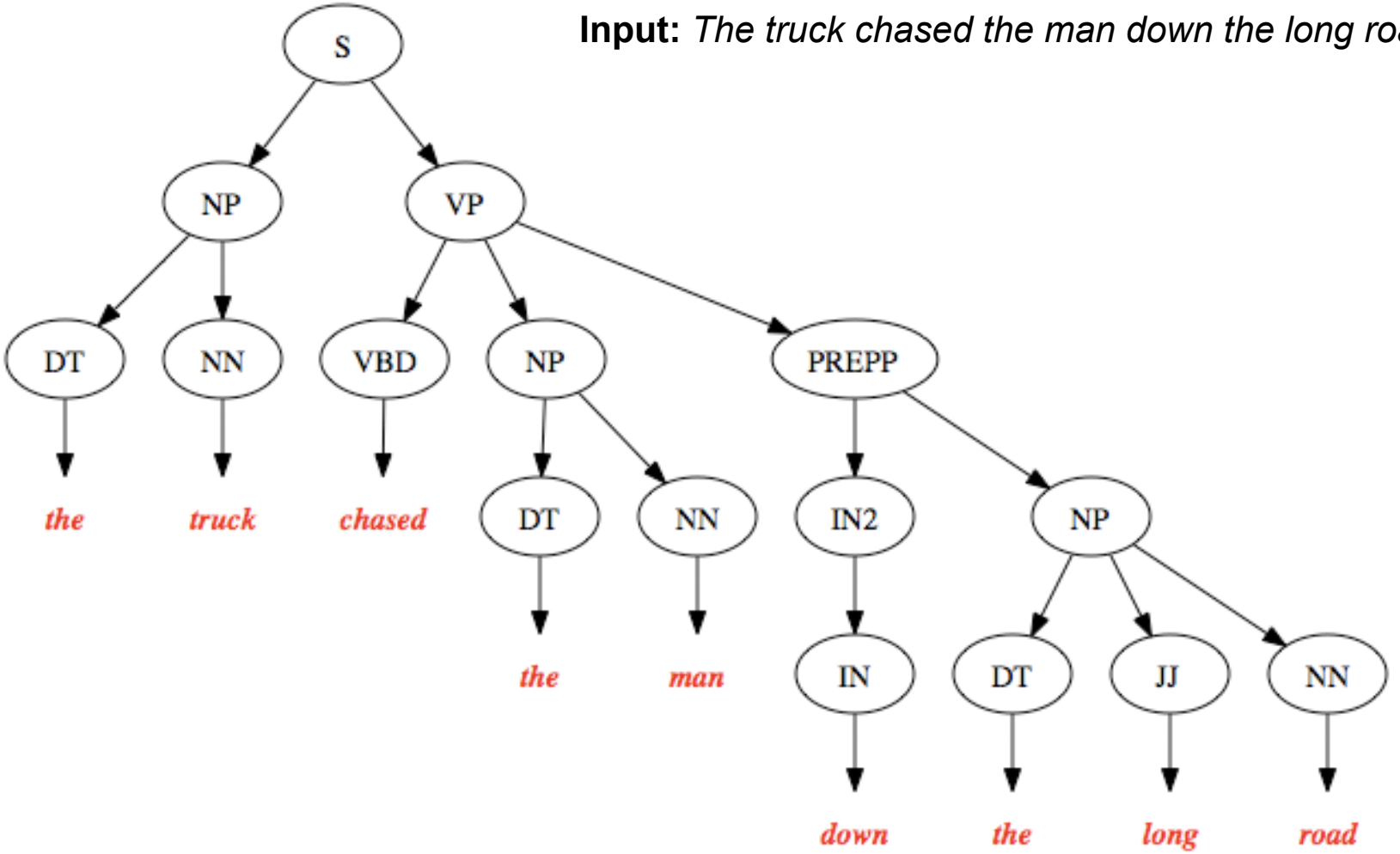
- The meaning of a sentence depends not only on the words in the sentence but their order and relation to each other. These relations are embedded In the **syntax**.

- **Part of speech** for each word in sentence:
  The det *truck* n *chased* v *the* det *man* n *down* p *the* det *long* adj *road* n

- **Grammar rules**:
  S -> NP VP
  VP -> V
  VP -> V PP
  VP -> V NP PP
  PP -> P NP
  NP -> N
  NP -> Det N
  NP -> Det Adj N

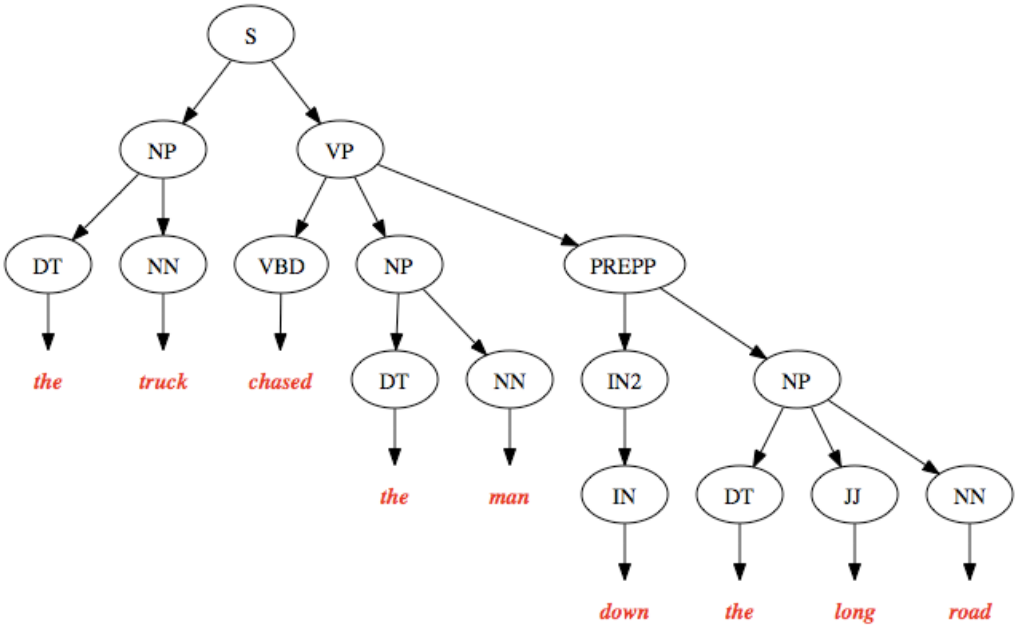
Parsing applies grammar rules to part-of-speech tokens to produce a representative tree structure

Input: *The truck chased the man down the long road*

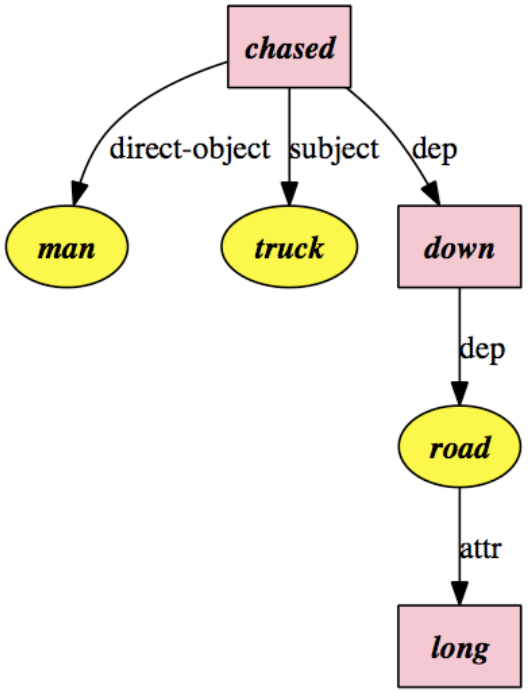


Parse tree ➔ dependency tree

Input: *The truck chased the man down the long road*



Parse tree



Dependency relations

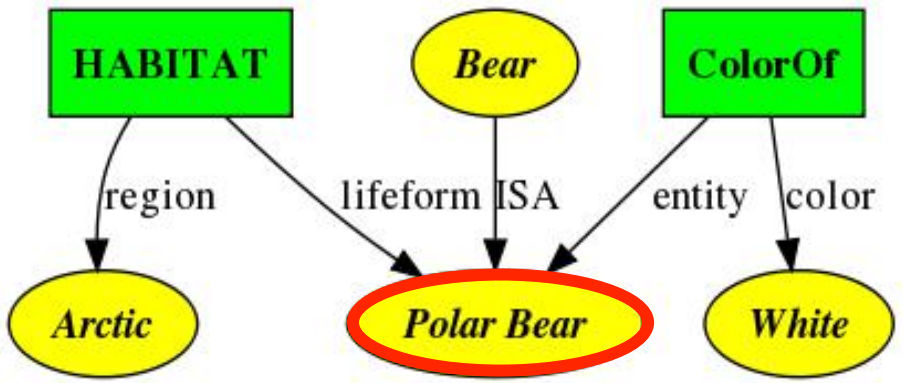
An online parser: http://nlp.stanford.edu:8080/parser/index.jsp

Representing Knowledge

- **Semantic nodes** represent entities and types:
  - Types *(House, teenager, polar bear, …)*
  - Elaborated types (*A tiny house with a metallic front door)*
  - Events (*John slowly washing the apple)*
  - Individuals (*Barack Obama, Homer Simpson, my house)*
- **Semantic relations** are applied to semantic nodes to give them meaning. (Semantic nodes have no structure of their own)
  - HABITAT-OF(polar-bear, Arctic)
  - SIZE-OF(house-1, tiny)
  - Need inventory of semantic relations
- **Meta-relations** translate between semantic relations.
  - E.g. *buy* and *sell* are different perspectives on same action
  - Used to decompose/ground meaning

Representing Word Meaning

polar bear: a white colored bear that lives in arctic

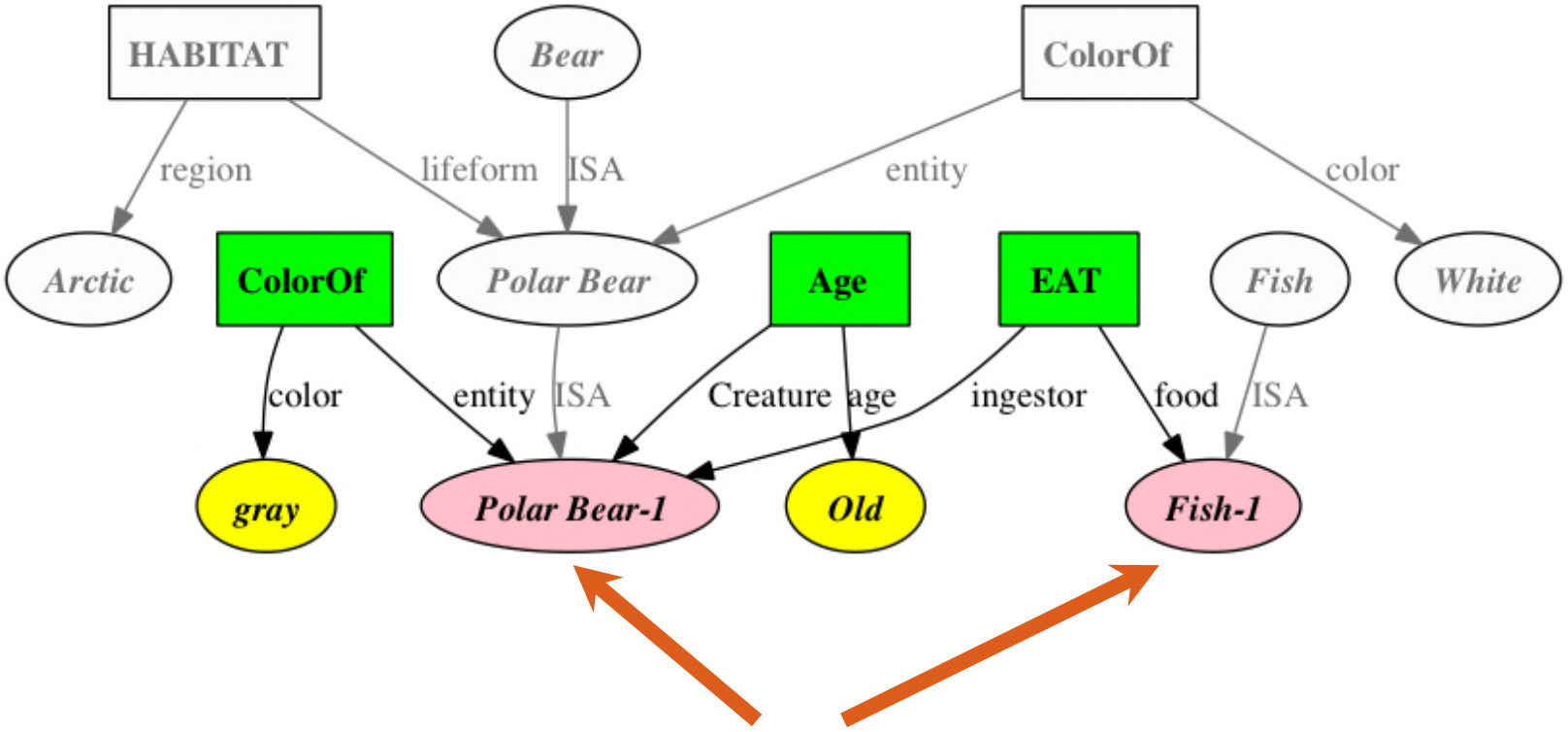


Semantic relation giving meaning to semantic node for polar bear

Semantic node for “polar bear”

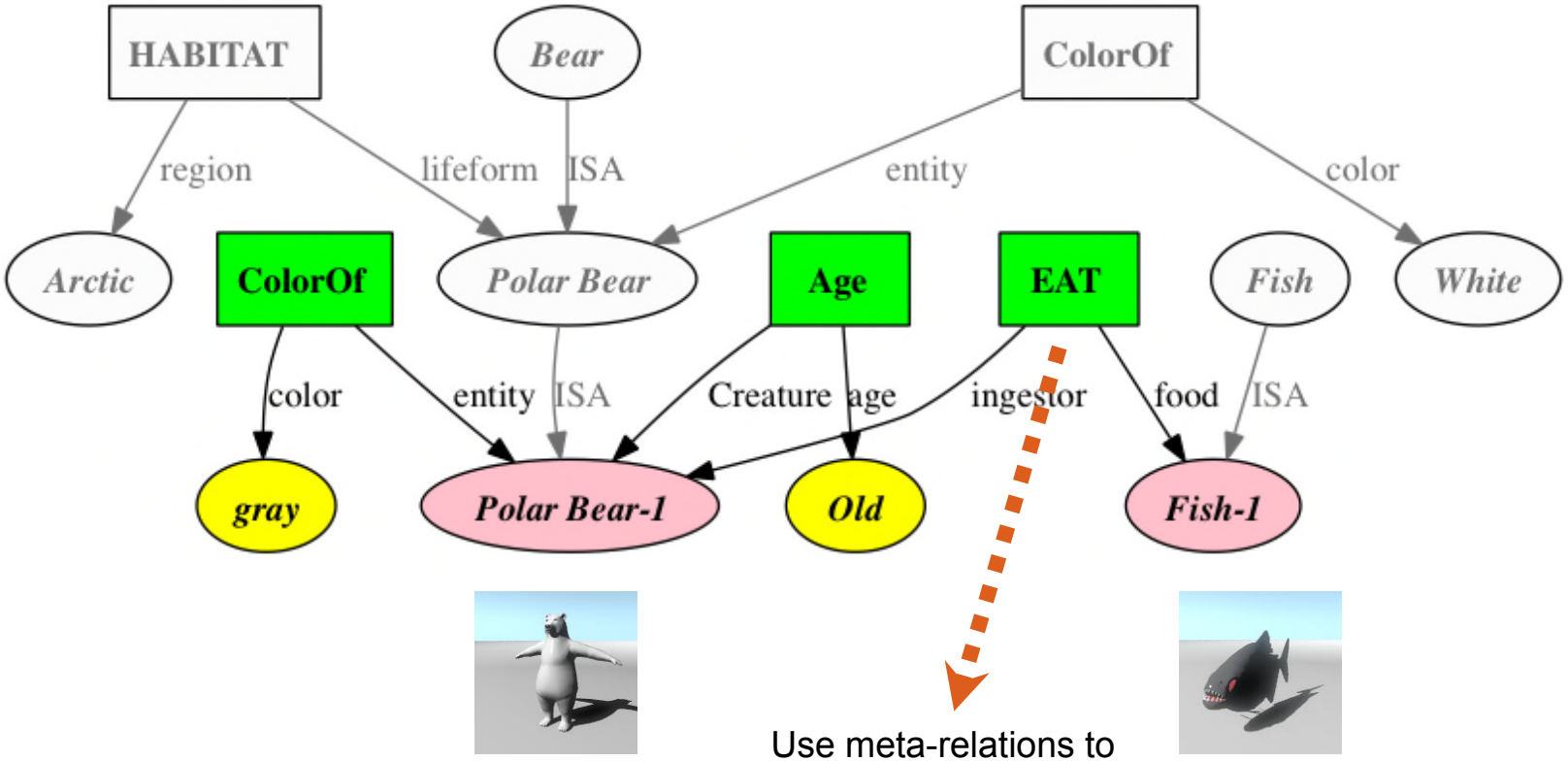
...and sentence meaning

The old gray polar bear ate the fish



Grounding sentence meaning

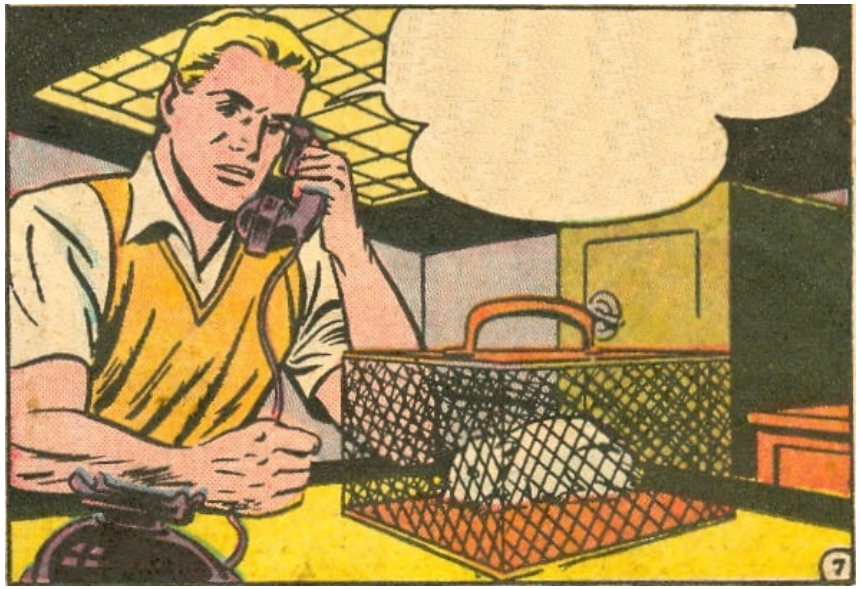
The old gray polar bear ate the fish



Outline

- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

Examples of high/low level descriptions (via Mechanical Turk)



- Acquire typical language (hi/low) for 100 comic book scenes
- Each scene described by 5 different Turkers

Low-level:
A man is using the telephone.
The man is wearing a yellow vest.
The man has blonde hair.
The man has white skin.
A white rodent is inside a cage.
The cage is on a table.
The phone is on the table.
The cage has a handle.
A safe is in the background of the room.

#High-level:
The man is a scientist working with white rodents.

#High-level:
The man is talking to another scientist.

#High-level:
The man feels guilt at imprisoning a white rodent.

Problem: Very different grounding for similar high-level semantics (Man wash floor/apple)

The man washed the floor.

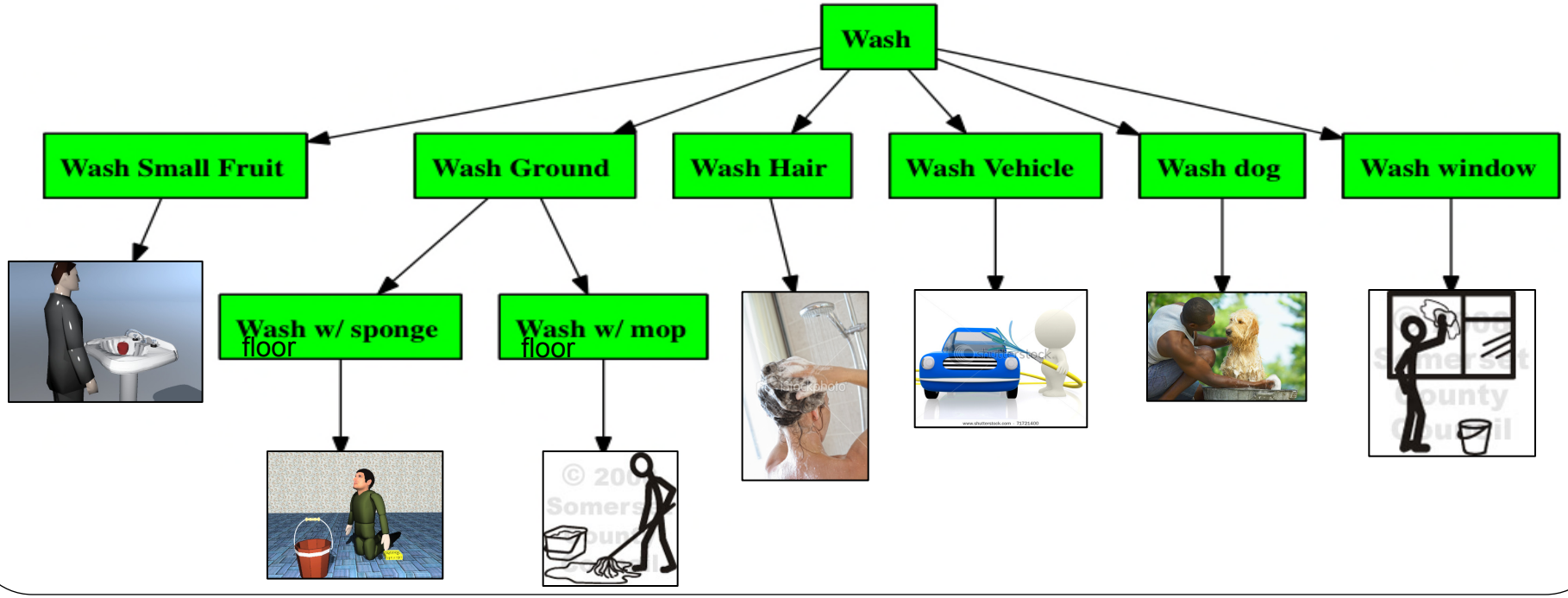
Low-level:
HOLD(man, sponge)
NEAR(man, bucket)
POSE(man, kneel)

The man washed the apple.

Low-level:
FACING(man, sink)
HOLD(man, apple) – not shown
IN-FRONT-OF(man, sink)

Proposed solution: Vignettes

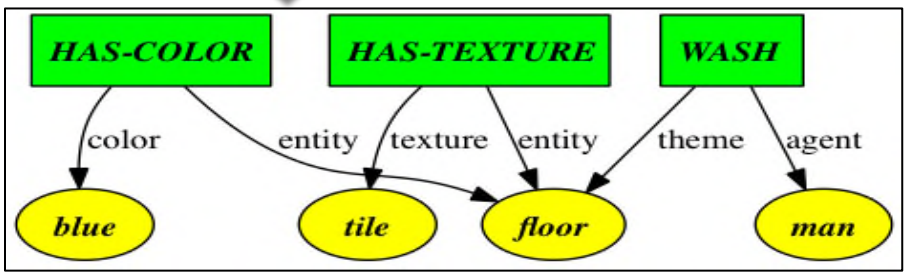
- ***Vignettes*** represent the ways meaning is realized in the world – they link **form** and **function**
  - Actions: freeze the action in time as in a comic panel
  - Locations: typical configurations of constituent objects
    - *Living room* ➔ *{Couch IN-FRONT-OF wall. Coffee table IN-FRONT-OF couch….}*



Example: Representing and grounding meaning

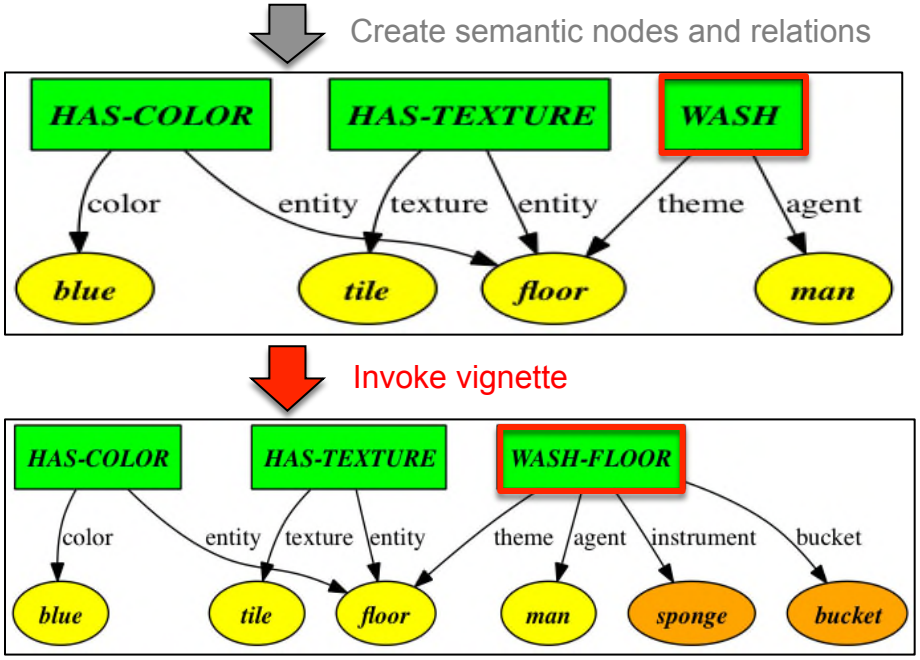
Input: *The man washed the blue tile floor*

Create semantic nodes and relations



Example: Representing and grounding meaning

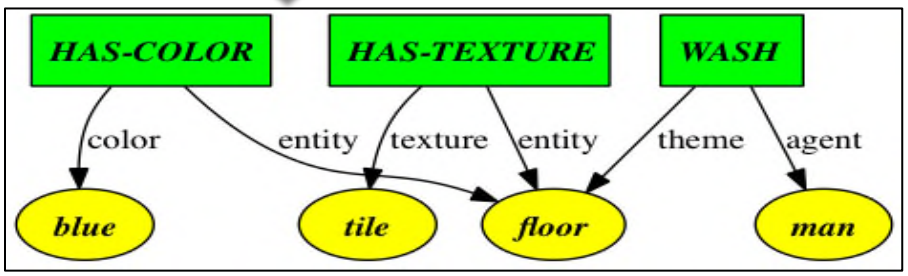
Input: *The man washed the blue tile floor*



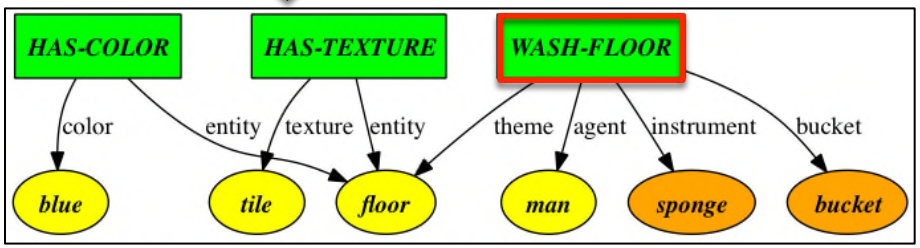
Example: Representing and grounding meaning

Input: *The man washed the blue tile floor*

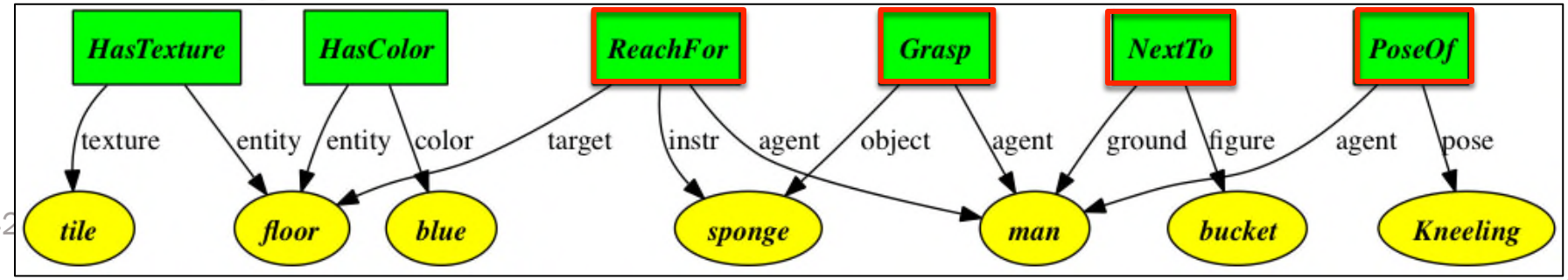
Create semantic nodes and relations



Invoke vignette

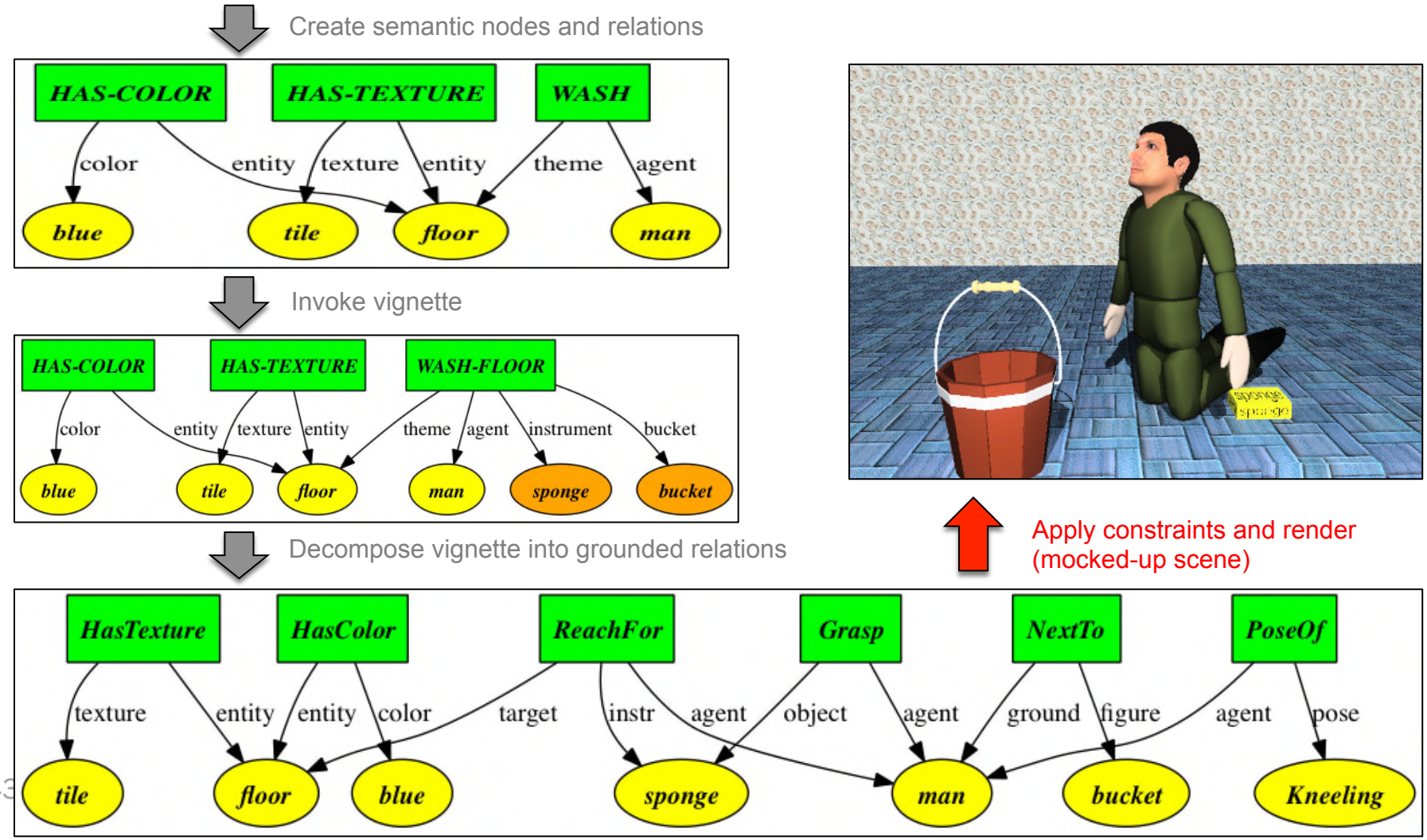


Decompose vignette into grounded relations



Example: Representing and grounding meaning

Input: *The man washed the blue tile floor*



Where do Vignettes come from? Use Crowd-sourcing with Amazon's Mechanical Turk

- https://www.mturk.com/mturk/welcome
- Location Vignettes
  - Show picture of different types of rooms. Turker does:
    1. Identifies main objects
    2. Specifies spatial relations for those objects
    3. Identifies subobject types using 3D library (e.g. kitchen table versus picnic table)
- Action Vignettes (given location and action)
  - Provide action sentence and location vignette picture/objects. Turker does:
    1. Identifies additional objects and participants
    2. Identifies subobject types using 3D library
    3. Specifies spatial relations, facial expressions, and poses for those objects

General World Knowledge with ConceptNet

- Acquired using crowd-sourcing
- http://openmind.media.mit.edu/en/concept/
  - (LocationOf "computer" "at apartment")
  - (LocationOf "computer" "at desk")
  - (LocationOf "computer" "at fraternity house")
  - (LocationOf "computer" "at home")
  - (LocationOf "computer" "at internet cafe")
  - (LocationOf "computer" "at library")
  - (LocationOf "piano" "at concert")
  - (LocationOf "piano" "at neighbor 's house")
  - (LocationOf "piano" "in church")
  - (LocationOf "piano" "in concert")
  - (LocationOf "piano" "in living room")
  - (LocationOf "piano" "in music room")

Outline

- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

Problem: spatial relation meaning depends on object properties

- Resolve "in" to more specific spatial relations
  - *Boat in water (*EMBEDDED-IN)
  - *Dog in boat (*IN CUPPED REGION)
- Depends on object shape and function

The boat is in the ocean. The dog is in the boat.

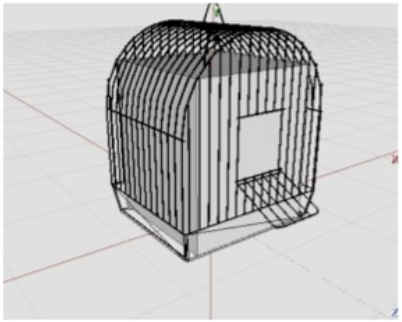
Solution: Spatial Tags assigned to 3D objects

BASE

CUP

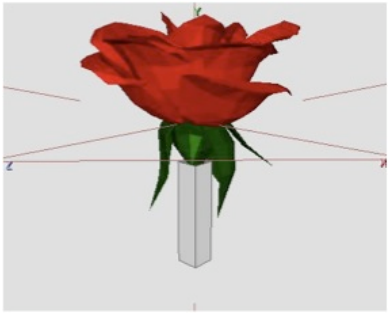
ON

CANOPY

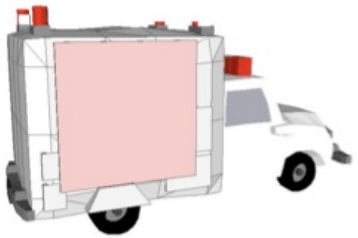


ENCLOSURE

HANDLE

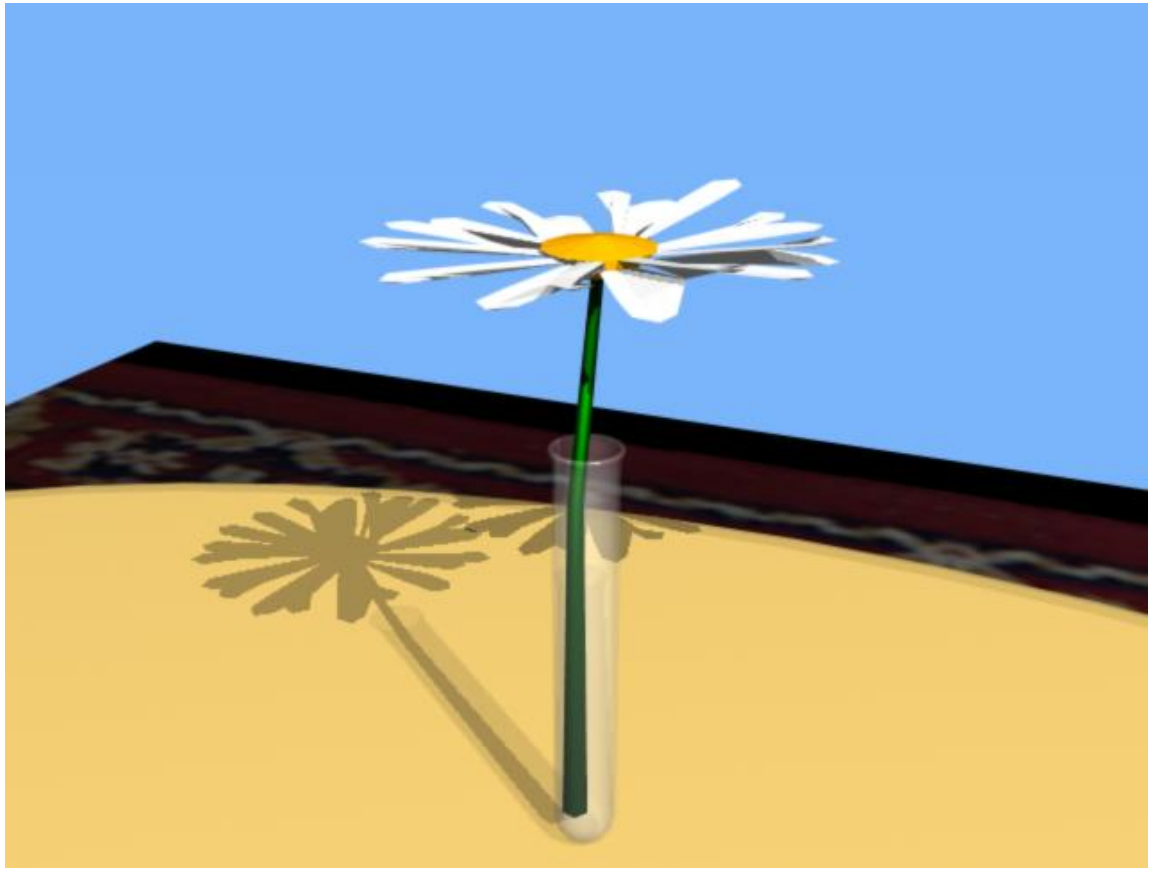


STEM



WALL

Using STEM and CUP tags



The daisy is in the test tube.

Using ENCLOSURE and ON tags

The bird is in the bird cage. The bird cage is on the chair.

Grounding of spatial relations

Spatial relation	Scene elements
ENCLOSED-IN	***Chicken in cage***
EMBEDDED-IN	***Horse in ground***
IN-CUP	***Chicken in bowl***
ON-TOP-SURFACE	*Apple on wall*
ON-VERTICAL-SURFACE	*Picture on wall*
PATTERN-ON	*Brick-texture on wall*
UNDER-CANOPY	*Vase under umbrella*
UNDER-BASE	*Rug under table*
STEM-IN-CUP	***Flower in vase***
LATERALLY RELATED	*Wall behind table*
LENGTH AXIS	*Wall*
DEFAULT SIZE/DIRECTION	All objects
REGION	*Right side of*
DISTANCE	*2 feet behind*
SIZE	*Small* and *16 ft long*
ORIENTATION	*facing*

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

Grounding of spatial relations

Spatial relation	Scene elements
ENCLOSED-IN	*Chicken in cage*
EMBEDDED-IN	*Horse in ground*
IN-CUP	*Chicken in bowl*
ON-TOP-SURFACE	***Apple on wall***
ON-VERTICAL-SURFACE	***Picture on wall***
PATTERN-ON	***Brick-texture on wall***
UNDER-CANOPY	*Vase under umbrella*
UNDER-BASE	*Rug under table*
STEM-IN-CUP	*Flower in vase*
LATERALLY RELATED	*Wall behind table*
LENGTH AXIS	*Wall*
DEFAULT SIZE/DIRECTION	All objects
REGION	*Right side of*
DISTANCE	*2 feet behind*
SIZE	*Small* and *16 ft long*
ORIENTATION	*facing*

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

Grounding of spatial relations

Spatial relation	Scene elements
ENCLOSED-IN	*Chicken in cage*
EMBEDDED-IN	*Horse in ground*
IN-CUP	*Chicken in bowl*
ON-TOP-SURFACE	*Apple on wall*
ON-VERTICAL-SURFACE	*Picture on wall*
PATTERN-ON	*Brick-texture on wall*
UNDER-CANOPY	***Vase under umbrella***
UNDER-BASE	***Rug under table***
STEM-IN-CUP	*Flower in vase*
LATERALLY RELATED	*Wall behind table*
LENGTH AXIS	*Wall*
DEFAULT SIZE/DIRECTION	All objects
REGION	*Right side of*
DISTANCE	*2 feet behind*
SIZE	*Small* and *16 ft long*
ORIENTATION	*facing*

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

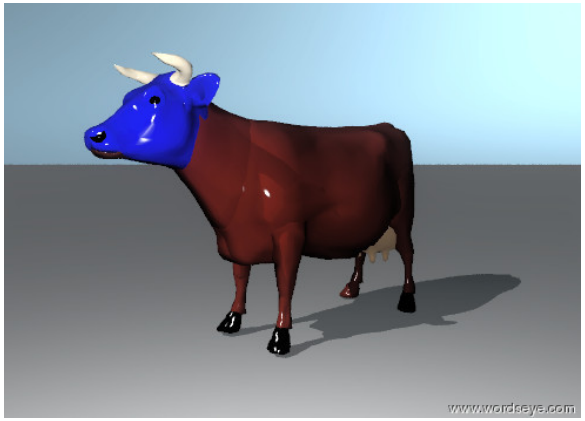
Ambiguity of "*Of*"

Text (A of B)	Conditions	Resulting Semantic Relation
Bowl of cherries	A=container, B=plurality-or-mass	CONTAINER-OF (bowl, cherries)
Slab of concrete	A=entity, B=substance	MADE-OF (slab, concrete)
picture of girl	A=representing-entity, B=entity	REPRESENTS (picture, girl)
Arm of the chair	A=part-of(B), B=entity	PART-OF (chair, arm)
Height of the tree	A=size-property, B=physical-entity	DIMENSION-OF (height, tree)
Stack of plates	A=arrangement, B=plurality	GROUPING-OF (stack,plates)

Semantic types, functional properties, and spatial tags used to resolve semantic relation for "*of*"

Grounding of “of”

Containment: *bowl of cats*



Part: *head of the cow*

Dimension*: height of horse is..*

Grouping: *stack of cats*

Substance: *horse of stone*

Representation: *Picture of girl*

Outline

- Motivation
- Background and system overview
- Under the hood
  - Lexicon, syntax, semantics, world knowledge
  - Vignettes
  - Spatial relations and ambiguity
- Applications and Conclusion

Potential Applications

Social Media

- Speed enables social interaction with pictures and promotes "visual banter".
- Automatically depict Twitter tweets

3D Games

- Trend toward malleable environments and interfaces (e.g. Spore, Tom Clancy's End War, Scribblenauts)
- Use WordsEye to modify game environment with language as part of the gameplay

Education (language skills, special needs learning)

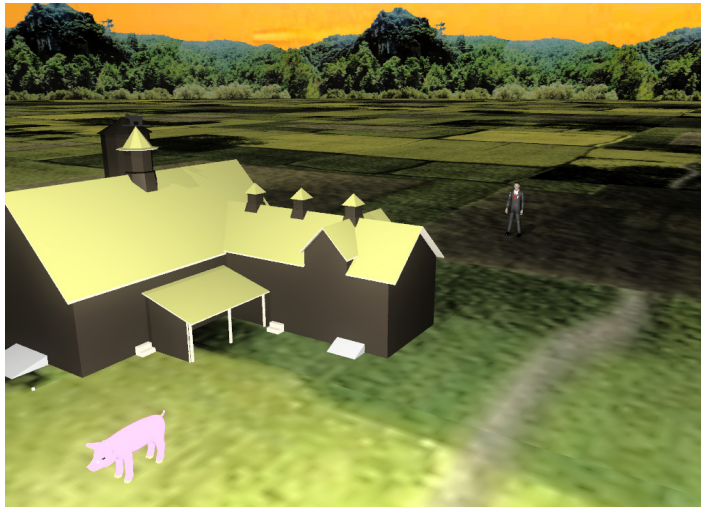
- Pilot study in a Harlem summer enrichment program

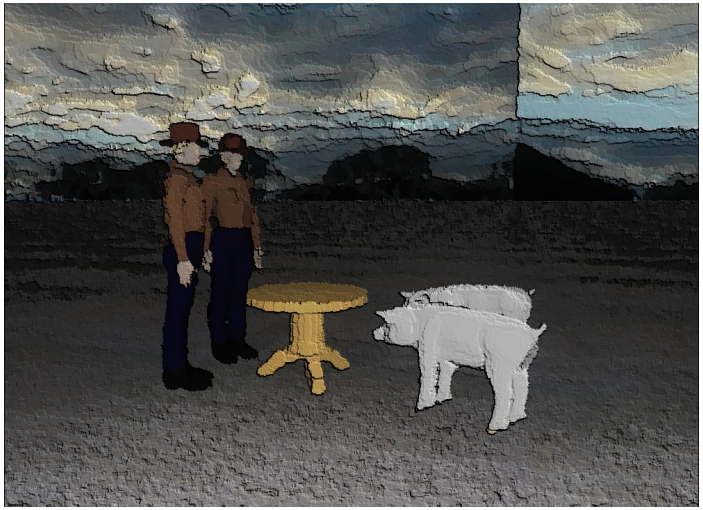
Using WordsEye in Education

- Tested with 6th graders at HEAF (Harlem Educational Activities Fund) over 5 weeks
- Students made storyboards for scenes in Animal Farm and Aesop's Fables
- Significantly higher literacy growth for WordsEye group in controlled experiment
- System helped imagine and visualize stories. They loved putting each other's 3D faces in scenes.

	Pre-test	Post-test	Growth
Group 1 (WordsEye)	15.82	23.17	**7.35**
Group 2 (control)	18.05	20.59	**2.54**

HEAF pictures from Aesop's Fables and Animal Farm





Conclusion

- Key Concepts
  - Lexical, syntactic, and semantic processing
  - Annotated 3D and 2D object library
  - World knowledge to translate from high to low level semantics
- Application in education, gaming, and social media
- Online tools and resources students can explore:
  - WordNet: http://wordnetweb.princeton.edu/perl/webwn
  - Mechanical Turk: https://www.mturk.com/mturk/welcome
  - FrameNet: http://framenet.icsi.berkeley.edu/
  - Stanford parser: http://nlp.stanford.edu:8080/parser/index.jsp
  - graphviz: http://www.graphviz.org/
  - WordsEye http://bit.ly/wordseye

Thank You

This work was supported in part by the NSF IIS- 0904361

http://bit.ly/wordseye (research system)
www.wordseye.com (old system)