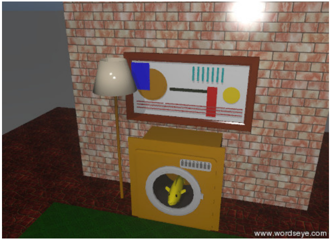
Frame-Based Representation of Lexical, Graphical, and Factual Knowledge for Text-to-Scene Generation

Daniel Bauer, Bob Coyne, Owen Rambow

Columbia University



This work is supported by the National Science Foundation under Grant No.IIS-0904361.

Contents

Text-to-Scene Generation

Knowledge in Text-to-3D-Scene Generation
- Lexical Knowledge
- Graphical Knowledge
- Factual Knowledge

Inference in Text-to-Scene Generation

Text-to-Scene Generation

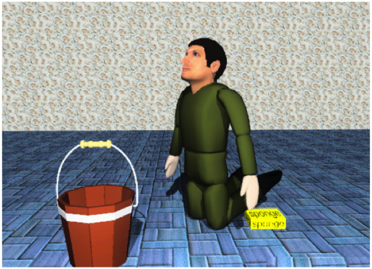
Knowledge in Text-to-3D-Scene Generation

- Lexical Knowledge
- Graphical Knowledge
- Factual Knowledge

Inference in Text-to-Scene Generation

Text-to-Scene Generation

- Generate a graphical scene depicting a textual description.
- Types of descriptions:



- **Low-level (primitive spatial relations)**:

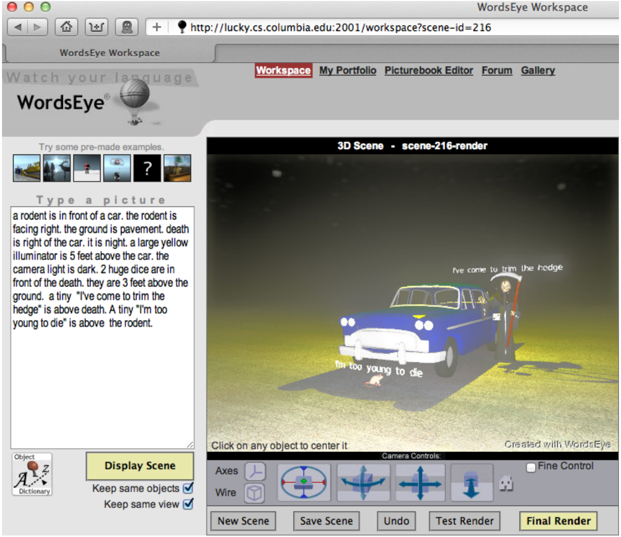
  The man is on the floor. He is kneeling. He is holding the sponge. The bucket is near the man.

- **High-level**:

  The man washed the floor.

WordsEye

[Coyne and Sproat, 2001]



http://lucky.cs.columbia.edu:2001/workspace?scene-id=216

a rodent is in front of a car. the rodent is facing right. the ground is pavement. death is right of the car. it is night. a large yellow illuminator is 5 feet above the car. the camera light is dark. 2 huge dice are in front of the death. they are 3 feet above the ground. a tiny "I've come to trim the hedge" is above death. A tiny "I'm too young to die" is above the rodent.

More eye-candy: http://www.wordseye.com/

Levels of Scene Description

- **High-Level:**
  - Functional view: Who does what to whom?
  - **Wash**(washer:x_1, theme:x_2)
  - Descriptions involves action/event verbs, complex entities...
- **Low-level:**
  - Realization view: How is it done?
    (graphical: what does it look like)
  - **On**(figure:x_1, ground:x_2), **Grasp**(grasper:x_1, theme:x_3), **Reach**(reacher: x_1, ground:x_2), **Kneel**(kneeler:$x1$)
- One high-level description → many low-level descriptions.

Text-to-Scene Generation

Knowledge in Text-to-3D-Scene Generation

- Lexical Knowledge
- Graphical Knowledge
- Factual Knowledge

Inference in Text-to-Scene Generation

Translating from High-Level Descriptions to Low-level Graphical Representations

- Requires three sources of knowledge:
    - **Lexical Knowledge**
        - Textual description to high-level semantic representation.
    - **Graphical Knowledge**
        - Translate high-level semantics into low-level graphical relations.
    - **Factual Knowledge**
        - Guide translation, rule out impossible/unlikely graphical representations.
- Use a common frame-based representation to bridge between language, functional and graphical meaning.
- Starting point: Frame Semantics[Fillmore, 1982].

Lexical Knowledge: Frame Semantics

[Fillmore, 1982]

- Word meaning can only be understood by referring to conceptual structure evoked by it.
- Frames are cognitive schemas describing relations between state/event participants.
- Semantic roles are specific to each frame (vs. universal roles or predicate-specific roles).
- Role of syntax and lexicon:
  - Valence patterns of a lexical item map syntactic arguments to frame elements.

Lexical Knowledge: FrameNet

- Bridge language and high-level semantic representation.
- Can build on FrameNet [Ruppenhofer et al., 2010]
  - High-level semantics / functional view.
  - $>$ 1000 frames for verbs, nouns, adjectives, prepositions.
  - Mapping from syntax / lexicon to frame semantics by providing example annotations for each frame.
  - Frame-to-Frame relations.

[Mary]$_{buyer}$	**bought**$_{Commerce\_buy}$	[an apple]$_{goods}$	[for \$1]$_{money}$	.
Subj		Obj	PP(for)	.

Problem with FrameNet

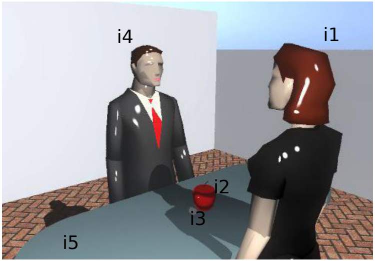
$[\text{Mary}]_{\textit{buyer}}$ $\textbf{bought}_{\textit{Commerce\_buy}}$ $[\text{an apple}]_{\textit{goods}}$ $[\text{for \$1}]_{\textit{money}}$.

- FrameNet annotations are ‘shallow’ (no semantic objects as arguments, just text spans).
- Cannot represent full sentence meaning
  - coreference, compositionality, $\cdots$.
  - Want graph structure semantics.

Instantiating Frames: Types and Instances

- Frames describe concept types.
- When lexical items evoke a frame in a description, the frame is instantiated.
- All frames carry a ‘self’ frame element, which is bound to the instance of the frame.
- When instantiating a frame, bind all the frame elements to instances (which may be defined by another frame).

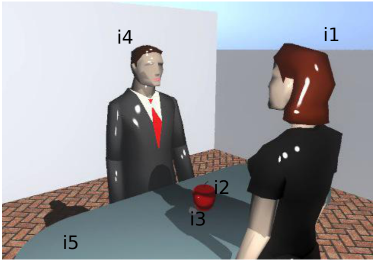
i6:



‘ Mary bought an apple for $1.‘

Commerce_buy(self: i6,
buyer: i4,
seller: i1,
goods: i2,
money: i3)

Instantiating Frames: Sentence meaning as AVMs / DAGs

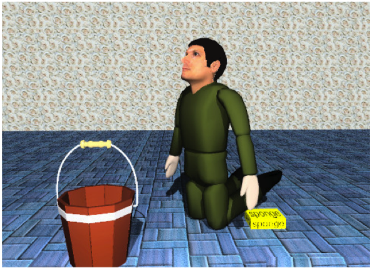


i6:

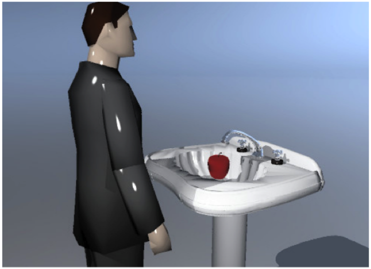
```
(i6 / Commerce_buy
    :buyer (i4 / Female
                :name "Mary")
    :seller (i5 / Human)
    :goods (i2 / Apple)
    :money (i3 / Money
                :currency "$"
                :value 1))
```

Graphical Knowledge

- Need knowledge about arrangement of 3D models to depict a situation/event.
- Low-level semantics, realization view.
- Non-compositionality of verb meaning:
  - Correct visualization of verb depends on verb *and* its arguments.



'The man washed the floor'

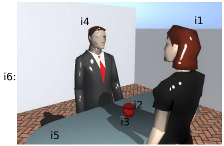


'The man washed the apple'

Graphical Knowledge: Vignettes

[Coyne et al., 2011]

- Frames with decomposition, grounded in graphical primitives.
- Represent different realizations for lexical frames.
- Vignettes extend frames by
    - optionally introducing new frame elements that participate in the visualization.
    - decomposition into sub-frames:
        - link to specific 3D model types (frames describing entities).
        - describe graphical structure of a scene (frames describing events/situations).

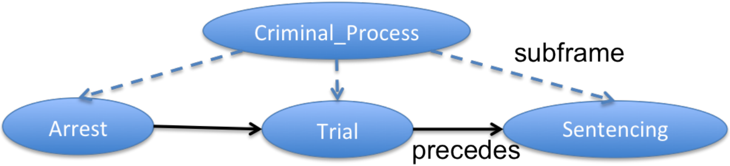


Commerce_counter(buyer, goods, money, seller)

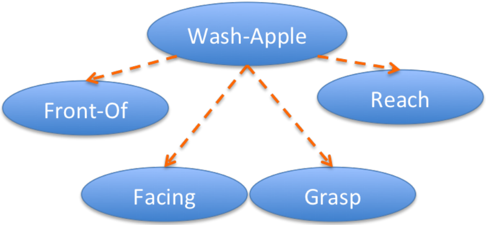
(ISA) Commerce_buy
at_counter(partcpt1:buyer, partcpt2:seller, counter:c)
on(figure:goods, ground:c)
on(figure:money, ground:c)

Graphical Knowledge: Vignette Decomposition

- (temporal) subframe relation in FrameNet:



- New frame-to-frame relation **subframe_parallel**.



Factual Knowledge

- Vignettes / Frames should define selectional restrictions on their frame elements to select the appropriate vignette.
- Some ontological information already incoded in frame-to-frame relations (ISA).
- In addition frame definitions for entity types need:
  - non-graphical properties of objects / attributes.
  - information about parts.
  - world knowledge (‘apples grow on trees‘).

Factual Knowledge

- Frame decompositions are declarative.
  - Simultaneously define properties of frame element fillers and restrict fillers to instances of frame elements that define this property.
- Can create frame elements for properties.
- Or use ‘self’ frame element to define properties of frames for entity types.

commerce_counter(buyer, goods, money, seller)

size(figure:goods, size:**small**)
animate(self:seller)
animate(self:buyer)

(ISA) commerce_buy
at_counter(partcpt1:buyer, partcpt2:seller, counter:c)
on(figure:goods, ground:c)
on(figure:money, ground:c)

apple( )

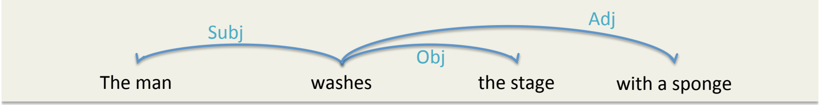
(ISA) fruit
size(figure:self, size:**small**)
shape(figure:self, shape:**round**)

Status of the VigNet Resource

VigNet currently contains:

- a small set of primitive spatial relations (on, next-to (direction and distance), in, direction..)
- small set (about 30) ‘abstract‘ vignettes
  - holding/touching target or patient, using handheld instruments, using stationary machine, human poses...
- several hundred verbal vignettes inheriting from and parametrizing abstract vignettes (ongoing...).
- about 2000 nominal vignettes mapping to about 3000 3D models (with physical attributes, parts, affordances).
- about 80 location vignettes (all rooms, including fixtures/affordances).

Example Analysis

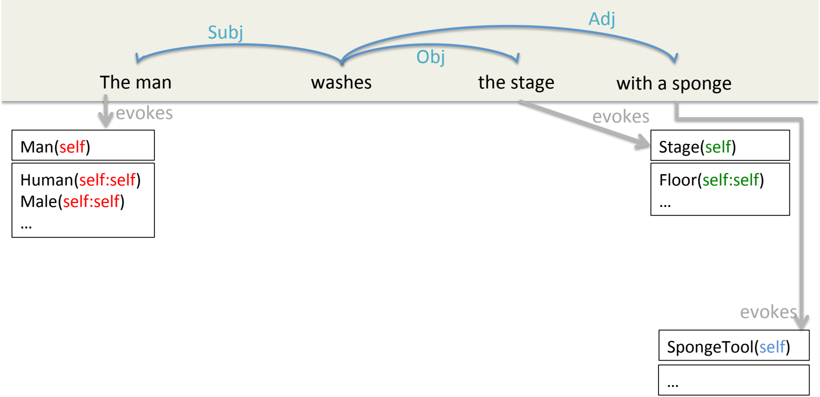

Subj

Adj

Obj

The man washes the stage with a sponge

Example Analysis



The man washes the stage with a sponge

Man(self)

Human(self:self)
Male(self:self)
...

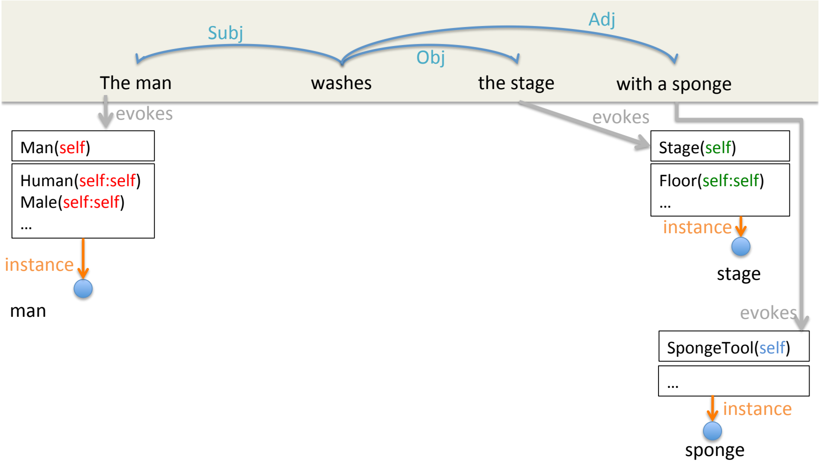
Stage(self)

Floor(self:self)
...

SpongeTool(self)

...

Example Analysis



Man(self)

Human(self:self)
Male(self:self)
…

man

Stage(self)

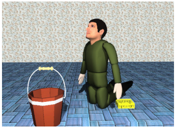
Floor(self:self)
…

stage

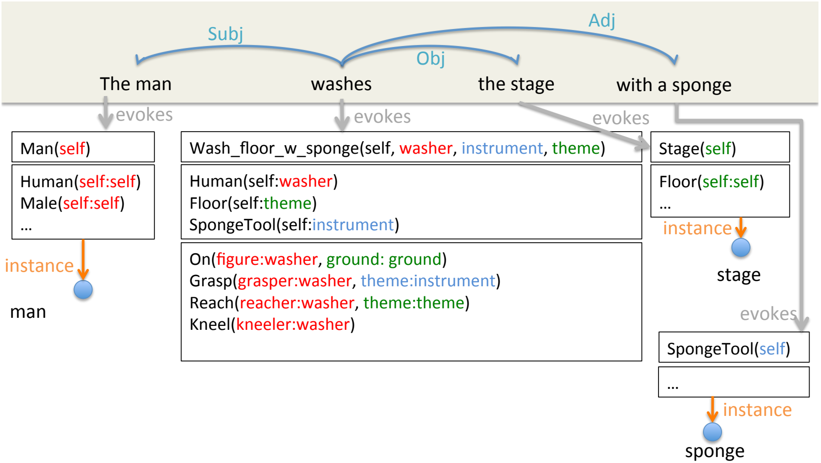
SpongeTool(self)

…

sponge



Example Analysis



Subj Obj Adj

The man washes the stage with a sponge

evokes evokes evokes evokes

Man(self)

Human(self:self)
Male(self:self)
…

instance

man

Wash_floor_w_sponge(self, washer, instrument, theme)

Human(self:washer)
Floor(self:theme)
SpongeTool(self:instrument)

On(figure:washer, ground: ground)
Grasp(grasper:washer, theme:instrument)
Reach(reacher:washer, theme:theme)
Kneel(kneeler:washer)

Stage(self)

Floor(self:self)
…

instance

stage

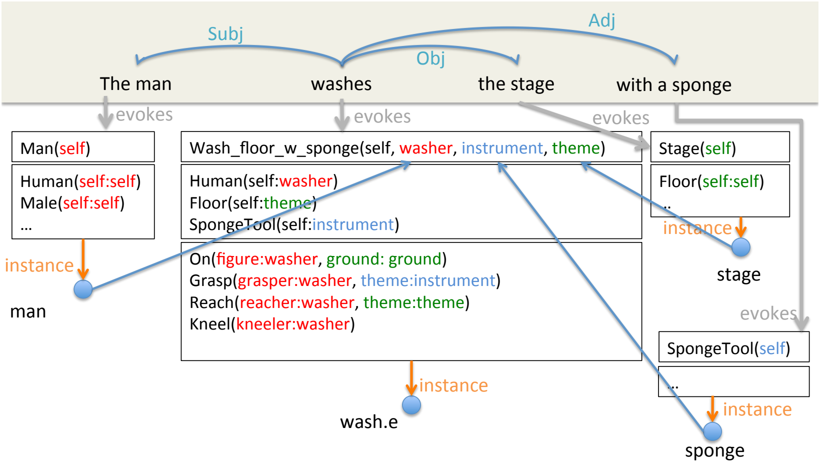
SpongeTool(self)

…

instance

sponge

Example Analysis



Subj

Obj

Adj

The man

washes

the stage

with a sponge

evokes

evokes

evokes

evokes

Man(self)

Human(self:self)
Male(self:self)
...

instance

man

Wash_floor_w_sponge(self, washer, instrument, theme)

Human(self:washer)
Floor(self:theme)
SpongeTool(self:instrument)

On(figure:washer, ground: ground)
Grasp(grasper:washer, theme:instrument)
Reach(reacher:washer, theme:theme)
Kneel(kneeler:washer)

instance

wash.e

Stage(self)

Floor(self:self)
..

instance

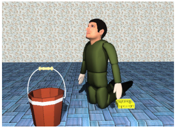
stage

SpongeTool(self)

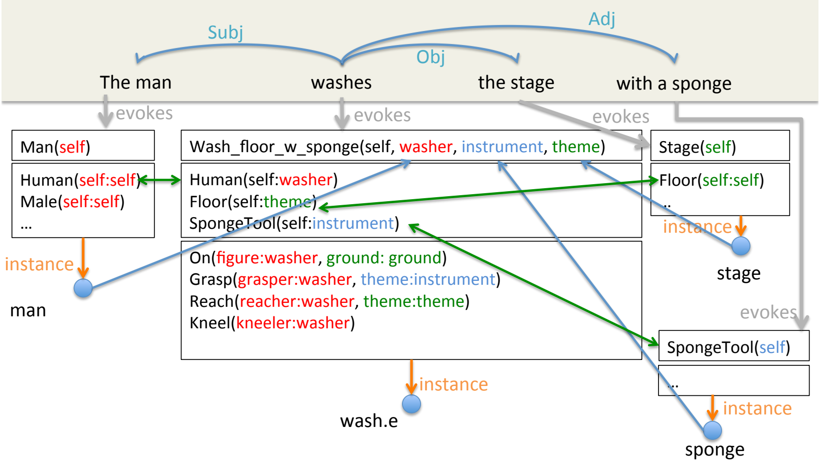
..

instance

sponge



Example Analysis



Subj

Adj

Obj

The man

washes

the stage

with a sponge

evokes

evokes

evokes

evokes

Man(self)

Human(self:self)
Male(self:self)
...

instance

man

Wash_floor_w_sponge(self, washer, instrument, theme)

Human(self:washer)
Floor(self:theme)
SpongeTool(self:instrument)

On(figure:washer, ground: ground)
Grasp(grasper:washer, theme:instrument)
Reach(reacher:washer, theme:theme)
Kneel(kneeler:washer)

instance

wash.e

Stage(self)

Floor(self:self)
..

instance

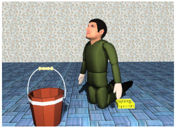
stage

SpongeTool(self)

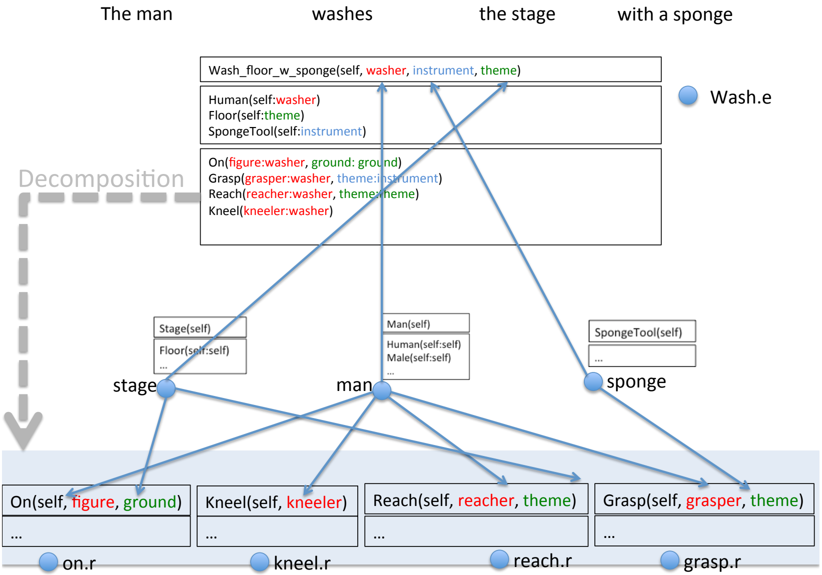
..

instance

sponge



Example Analysis



The man washes the stage with a sponge

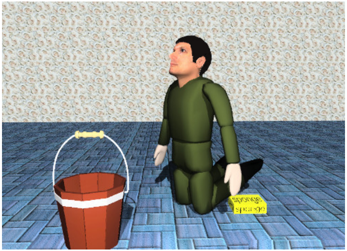
Wash_floor_w_sponge(self, washer, instrument, theme)

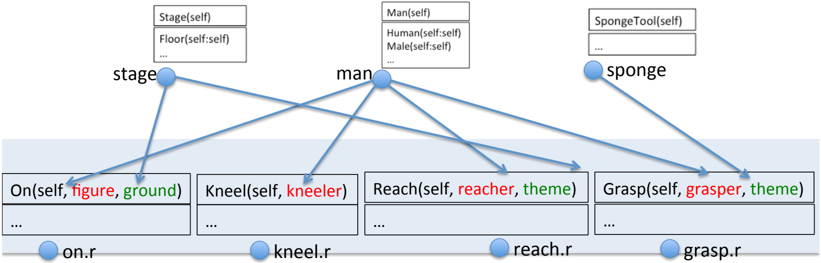
Human(self:washer)
Floor(self:theme)
SpongeTool(self:instrument)

On(figure:washer, ground: ground)
Grasp(grasper:washer, theme:instrument)
Reach(reacher:washer, theme:theme)
Kneel(kneeler:washer)

Example Analysis

The man washes the stage with a sponge





Stage(self)

Floor(self:self)
...

stage

Man(self)

Human(self:self)
Male(self:self)
...

man

SpongeTool(self)

...

sponge

On(self, figure, ground)

...

on.r

Kneel(self, kneeler)

...

kneel.r

Reach(self, reacher, theme)

...

reach.r

Grasp(self, grasper, theme)

...

grasp.r

Text-to-Scene Generation

Knowledge in Text-to-3D-Scene Generation

- Lexical Knowledge
- Graphical Knowledge
- Factual Knowledge

Inference in Text-to-Scene Generation

Inference in Text-to-Scene Generation

Two levels of inference:

- Interpret primitive spatial relations (On, Near...) using spatial reasoning to create an actual 3D scene.
  - Already supported in WordsEye.
  - Currently adding support for more elaborate reasoning (in rooms etc, ...)
- Resolve high-level frame semantics into vignettes (**this talk**)
  - Two step approach:
    1. Parse into high-level functional frame semantics.
    2. Resolve to low-level vignettes.

Semantic Parsing

- n-best dependency analysis for input sentence.
- create n-best FrameNet-style semantic parses for each frame evoking element and its frame elements in isolation:
  - Map input parse to previously observed frame annotations using statistical alignment model between syntactic dependency structures in FrameNet annotations and input.
  - Rank annotation hypotheses using semantic information.
- Construct forest of possible analyses for the entire sentence.

Decoding the Hypothesis Forest / Finding Vignettes

- Find all possible vignettes for analyzed frames, find consistent sets of vignettes.
- Checking if a vignette assignment is consistent is easy:
  - Subsumption is easy: there is no negation or quantification.
  - Just check if selectioal restrictions are met.
- Finding a consistent vignette annotation is hard:
  - Too many vignettes to check.
  - Methods from constraint solving?

Greedy Vignette Annotation

- Greedy heuristics:
  - Moving down the parse tree, assign the most restrictive vignette first.
  - This limits the choice of other vignettes.
  - Could produce to more interesting realizations.
- Disadvantages:
  - May not be the right strategy, can miss solutions.
  - Not sure if syntactic structure should guide semantic composition.

Thank you!

www.wordseye.com

‘the world is ROUND‘, by Gary Zamchick:

> *the humongous white shiny bear is on the american mountain range. the mountain range is 100 feet tall. the ground is water. the sky is partly cloudy. the airplane is 90 feet in front of the nose of the bear. the airplane is facing right.*

Coyne, B., Bauer, D., and Rambow, O. (2011).
Vignet: Grounding language in graphics using frame semantics.
In *ACL Workshop on Relational Models of Semantics (RELMS 2011)*, Portland, OR.

Coyne, B. and Sproat, R. (2001).
WordsEye: An automatic text-to-scene conversion system.
In *Proceedings of the Annual Conference on Computer Graphics*, pages 487–496, Los Angeles, CA.

Fillmore, C. J. (1982).
Frame semantics.
In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. R., and Scheffczyk, J. (2010).
Framenet II: Extended Theory and Practice.
ICSI Berkeley.