The habitual listener: a usage-based view of lexical meaning, and a matching computational model of utterance understanding

Katrin Erk (University of Texas at Austin)

joint work with:
Aurélie Herbelot (University of Trento),

Gabriella Chronis (University of Texas at Austin)

What is a usage-based approach to lexical semantics?

- Analyzing word meaning based on observed uses of a word in a large corpus

- Used in lexicography since the Collins-Cobuild dictionary, which drew on a corpus to determine dictionary senses
- Can be done to…
    - Identify word senses: Collins-Cobuild, and Zeevat et al 2014 (78 senses for the verb "fall" across English, Dutch, French, Russian, German)
    - Identify usage clusters (Glynn 2009)
- Based on different theories:
    - Zeevat et al 2014: decomposition into "moderately universal" semantic features that seem to be truth-conditional
    - Glynn 2009: Cognitive linguistics, "encyclopedic semantics", that is, word meaning includes cultural and social traces as well as some world knowledge
      This is the view that is most closely associated with the term "usage-based semantics"

What is an encyclopedic view of lexical semantics? Let's look at Fillmore's frame semantics (1982, 1985)

Words evoke "**frames**",
chunks of background knowledge that involves concepts, stories, cultural influences

- Breakfast: Assumes that we have one long period of sleep each day. It is the first meal eaten after that long period of sleep, and involves certain foods that are typically eaten
  - They met for breakfast: infer time of meeting
  - breakfast for dinner: infer foods
- coast, shore: boundary between land and sea, but different perspectives.
  - Compare: coast to coast, shore to shore
- decedent: dead person, but only in the legal context of a will

- Understanding an utterance:
  Integrating all these backgrounds evoked by the words into a larger whole

Usage-based lexical semantics with large language models

One way to look at language models / word embeddings:

as a **compacted collection of utterances from many speakers**

Using language models / word embeddings in lexical semantics:
as a record of the patterns and regularities occurring in the use of a word
(caveat: at least the ones that the model was able to pick up on)

Particularly useful with a **usage-based view of lexical semantics**, if we assume

- that word meanings are learned from usages,
- and that stored meanings match usage patterns

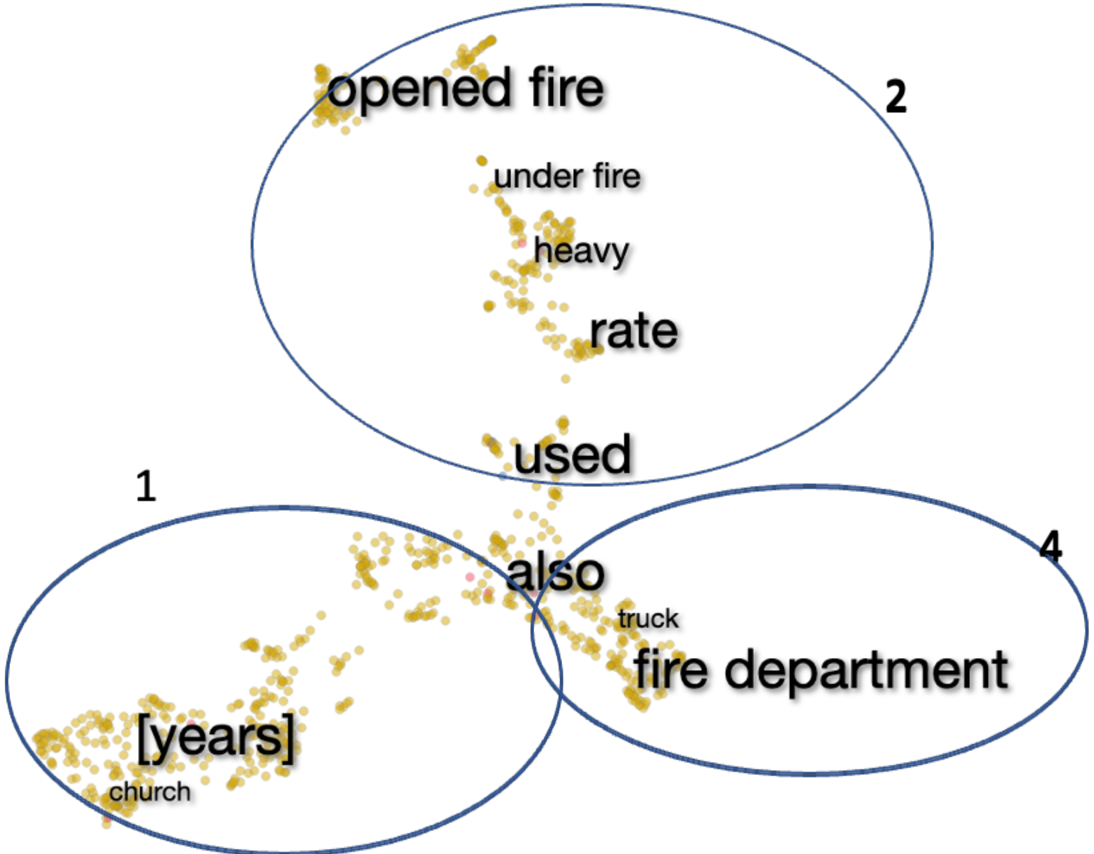
What patterns do we find in word embeddings? Do we see social, cultural traces?

Inspecting clusters of BERT token embeddings (Erk/Chronis 2023)

- Manual inspection of **multi-prototype embeddings** of words: BERT-base token clusters
- 5 clusters per lemma, k-means
- Layer 8, because best at taxonomic similarity prediction (as opposed to relatedness = topical similarity prediction)
- 45 nouns, sampled from different polysemy and concreteness bins
- For each, 100 tokens from the BNC
    - (also inspected Context Atlas, Wikipedia data, layer 8)

Inspecting multi-prototype embeddings: the noun “fire”

- Cluster 0: emotion, transformative fire
  - Changez said nothing, but shuffled backwards , away from the fire of Anwar’s blazing contempt
  - Never again, … would the ceremonies be performed; gone were the offerings, the blood-shedding , the fire and incense
- Cluster 1: destructive fire
  - There was a fire at Mr’s store and they called it arson.
  - An electrical short circuit started the fire, they think.
- Cluster 2: artillery
  - small-arms fire
- Cluster 3: hearth
  - or reading in the shadow of a fire;
  - The bar is warm and cosy, with an open *fire* and oak beams.
- Cluster 4: compound nouns/fire control
  - half a layer of fire cement
  - fire alarms were installed



Example of multi-prototype embeddings: the noun “fire”

- Cluster 0: emotion, transformative fire
  - Changez said nothing, but shuffled backwards , away from the fire of Anwar’s blazing contempt
  - Never again, … would the ceremonies be performed; gone were the offerings, the blood-shedding , the fire and incense
- Cluster 1: destructive fire
  - There was a fire at Mr’s store and they called it arson.
  - An electrical short circuit started the fire, they think.
- Cluster 2: artillery
  - small-arms fire
- Cluster 3: hearth
  - or reading in the shadow of a fire;
  - The bar is warm and cosy, with an open *fire* and oak beams.
- Cluster 4: compound nouns/fire control
  - half a layer of fire cement
  - fire alarms were installed

- Not just one sense fire=flame,
  but separate senses for dangerous fire and cozy fire
- A cluster that could be described as just syntactic (noun compounds) but that turns out to have a meaning
- “Story traces” and emotions in the embeddings

"Story traces," judgments, social context in all embeddings, from count-based to predicted to contextualized

- Clustering count-based vectors (Reisinger/Mooney 2010): separate senses of "wizard" for Merlin, Harry Potter
- Count-based vectors with interpretable dimensions (Baroni et al 2010 STRUDEL, Baroni/Lenci 2010 TypeDM): dimensions "prefer actional and situational properties (riding, parking, colliding, […]) over parts (such as wheels and engines)"
- In diachronic distributional semantics, the same models are used to detect changes
  - in word sense over time (Rosenfeld 2019, del Tredici 2020)
  - in social context over time (Kutuzov et al 2017)
  - One researcher's signal is the other's noise!
- Social dimensions in recent embeddings:
  - … as a problem: Detecting and removing bias in embeddings (Bolukbasi et al 2016, Webson et al 2020, in contextualized embeddings: Bommasani and Cardie 2020, Kaneko and Bollegala 2021)
  - … as good data: Kozlowski et al 2019: "the geometry of culture"

Why do we find “story traces”, judgments, social/cultural traces in embeddings?

- Embeddings aggregate the usage contexts of a word
- Human use of a word reflects…
  - what they find important, useful, or useless
  - what judgments they make
  - what stories they often tell using the words

What does this mean for the human lexicon?

- Doesn’t mean that the human lexicon has to include story effects.

- Human lexicon could still be “thin” and highly abstracted.
- Or rich, even consisting just of exemplars.
  Or everything inbetween

- But: story signal is there for humans to pick up
  - Mechanism for humans to synchronize on “word story representations” through the stories that they hear others tell
  - Also, there are lexico-syntactic signals that group word tokens along story lines, like “fire X” for fire control words

Outline

- A usage-based view on lexical meaning, with “story traces”
- **Which theories of utterance meaning go well with this?**
- Situation Description Systems: a framework for utterance meaning that allows for “story influences” on word meaning in context
- Not a “literal listener” but a “habitual listener”: Less work to do for pragmatic reasoning
- Situation Description Systems with corpus-derived parameters

Dividing up utterance meaning: Recanati 2003, Literal Meaning

- Recanati characterizes **Minimalism** as distinguishing
  - **Literal meaning**: sentence meaning, truth-conditional, independent of speakers, may have "holes" to be filled in by context
  - **Speaker meaning** includes pragmatics
- and characterizes **Contextualism** as distinguishing:
  - **What is said**: what a naive speaker would say the sentence means. Determining this already involves some pragmatics (**primary pragmatic processes**)
  - **What is implied**: conscious reasoning over what the speaker implies (**secondary pragmatic processes**)

Contextualism, and an encyclopedic view of word meaning

- Contextualism:
  - **What is said**: determined through **primary pragmatic processes**
  - **What is implied**: conscious reasoning, **secondary pragmatic processes**
- **Word meanings that contain chunks of background knowledge** including social/cultural and emotional traces:
  - **Pragmatic** in nature
  - But **habitual, conventionalized** knowledge rather than something that needs to be inferred consciously
  - Determining word meaning in context as part of primary pragmatic process (Recanati 2017)
  - Involves primary pragmatic influences on word meaning in context

My current favorite example sentence

The astronomer married the star

Why is the “celestial object” sense of “star” salient?

Must be because of the word “astronomer”,

but this is a script/scene type influence, as they are not semantic neighbors (not linked by a semantic role)

Why is there even a pun?

If disambiguation because of local semantic neighborhood was first, then “star” would be disambiguated to “person” by “marry"

Script/Scene knowledge in utterance understanding

- Fillmore 1985: Utterance understanding involves script knowledge
  - He pushed against the door. The room was empty.
  - likely goal of pushing door: opening it.
  - likely postcondition of open door: look into room behind it
- Recanati 2003: Knowledge about wider scenario as part of primary pragmatic processes

Outline

- A usage-based view on lexical meaning, with “story traces”
- Which theories of utterance meaning go well with this?
- **Situation Description Systems: a framework for utterance meaning that allows for “story influences” on word meaning in context**
- Not a “literal listener” but a “habitual listener”: Less work to do for pragmatic reasoning
- Situation Description Systems with corpus-derived parameters

Situation Description Systems:
Modeling interacting constraints on word meaning in context

Model interacting constraints on word meaning in context:

- Local context, as in She caught a ball / She organized a ball : selectional constraints
- Global context, as in The astronomer married the star : scenario influences

Constraints can "pull in different directions":

Probabilistic graphical model, constraints as edges in a graph'

Represent sentence meaning through

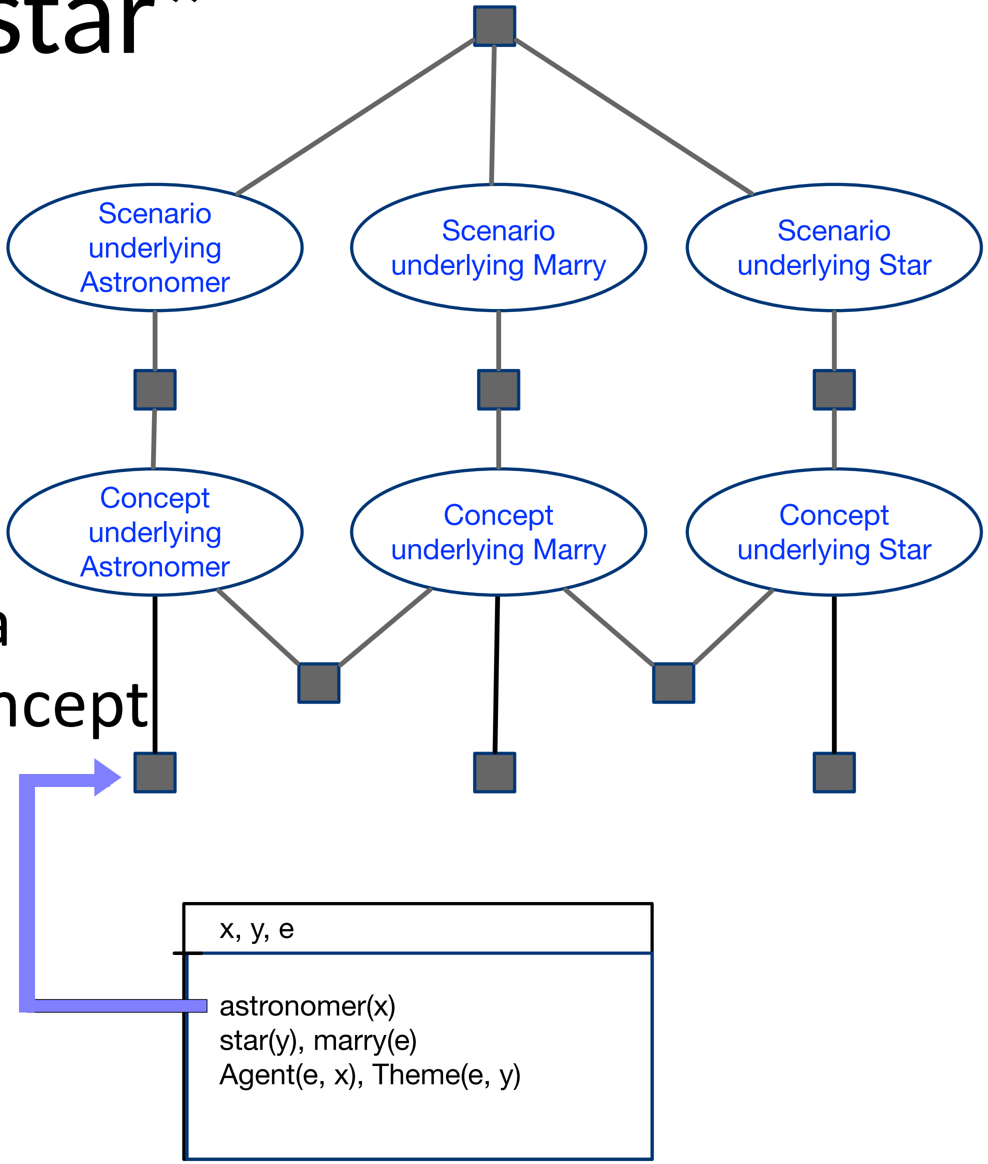
- Logical form: Discourse Representation Structure
- Conceptual representation: probabilistic graphical model to represent constraints on lexicalized concepts underlying words

The two representations are connected:

Literals from the Discourse Representation Structure as observations in the probabilistic graphical model

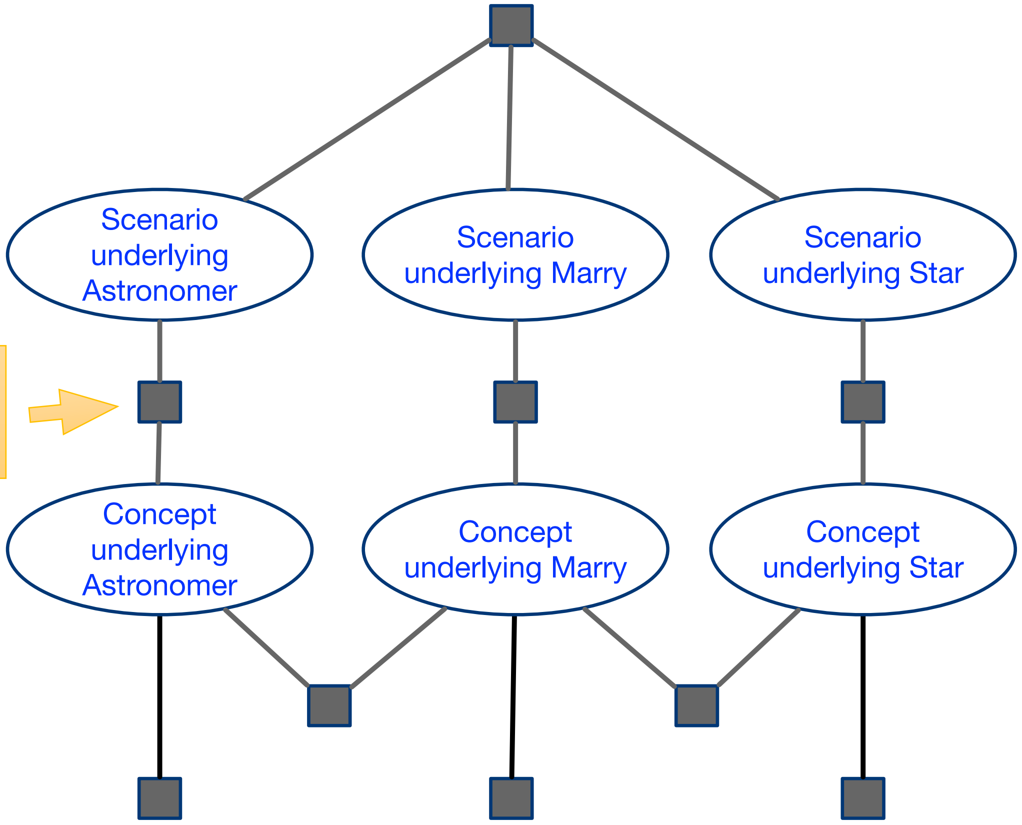
An example of a situation description system: "an astronomer married a star"

- Meaning represented through logical form and conceptual graph
  - Logical form: Discourse Representation Structure (DRS)
  - Conceptual structure: For each lexical item in the utterance, a probabilistically inferred underlying concept
  - Concept is constrained by different factors, shown as black boxes in a factor graph



New formalization with factor graphs, simplified wrt. Erk/Herbelot 2023

Factor graphs



factor node a, function f_a
often represented as matrix

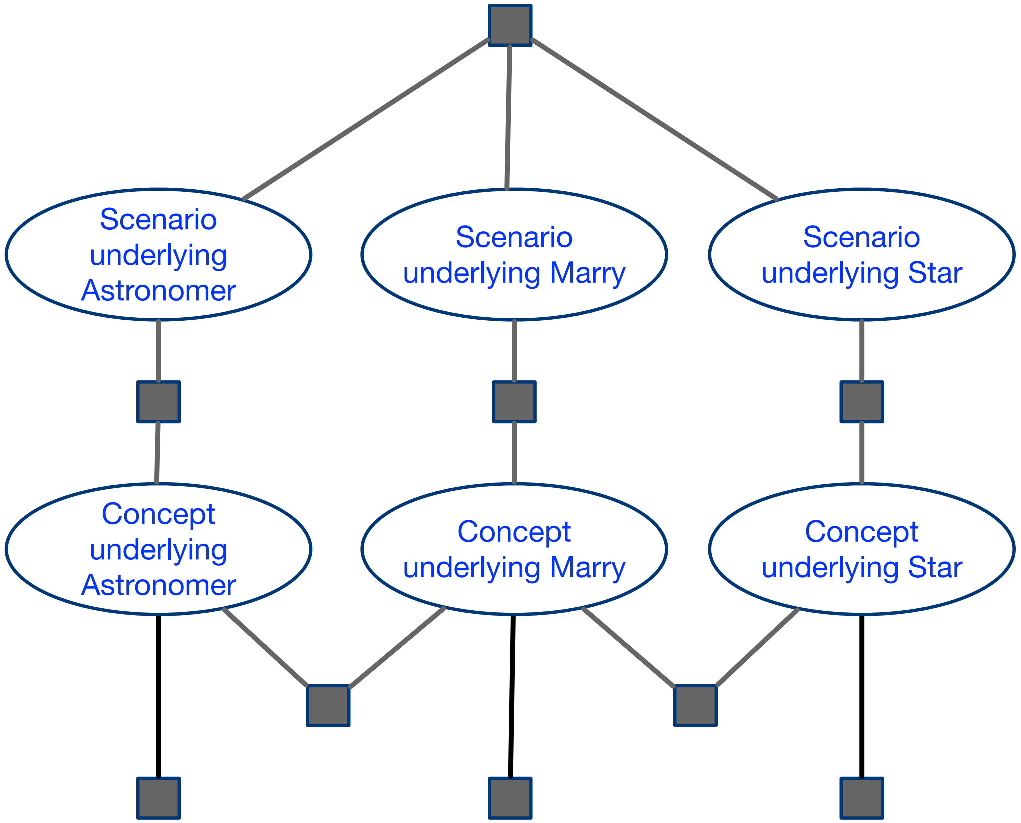
- A probabilistic graphical model with two types of nodes:
    - Variable nodes are random variables
    - Factor nodes express constraints on adjacent variable nodes
- Factor node a has associated function f_a: values for all adjacent variables -> weight
    - Example for the factor node f_a from above:

scenarios

concepts

f_a: P(concept|scenario)	**Movies**	**Space**
Astronomer	0	0.3
Marry	0.3	0.3
Star-person	0.3	0
Star-sun	0	0.3
Director	0.3	0

Factor graphs



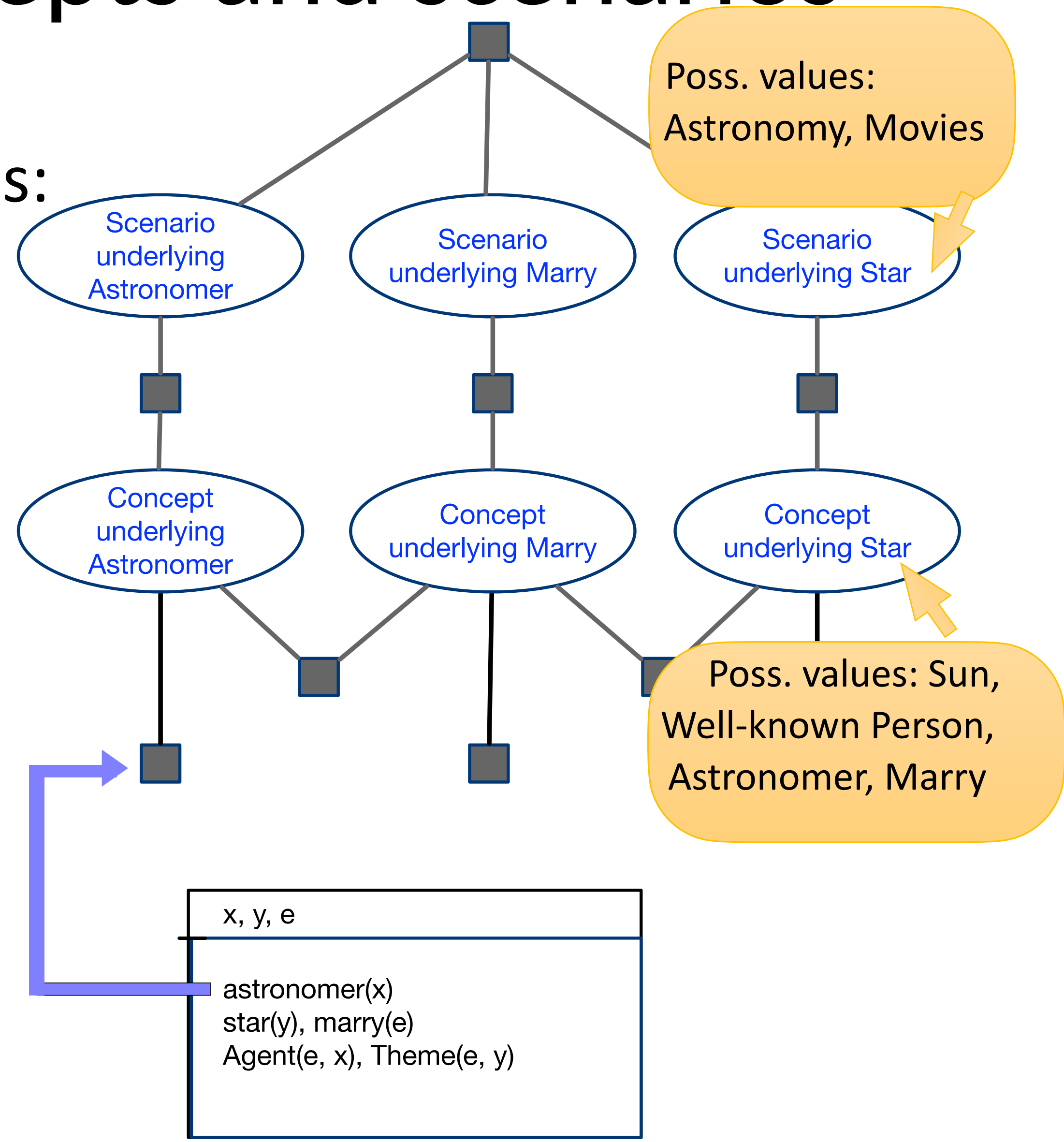
(Non-normalized) probability of assignment $\mathbf{x} = x_0..x_n$ to all variables $X_0..X_n$:

$$p(\mathbf{x}) = \prod_{a \in F} f_a(x_a)$$

where F set of factors, x_a assignments to variables adjacent to factor a

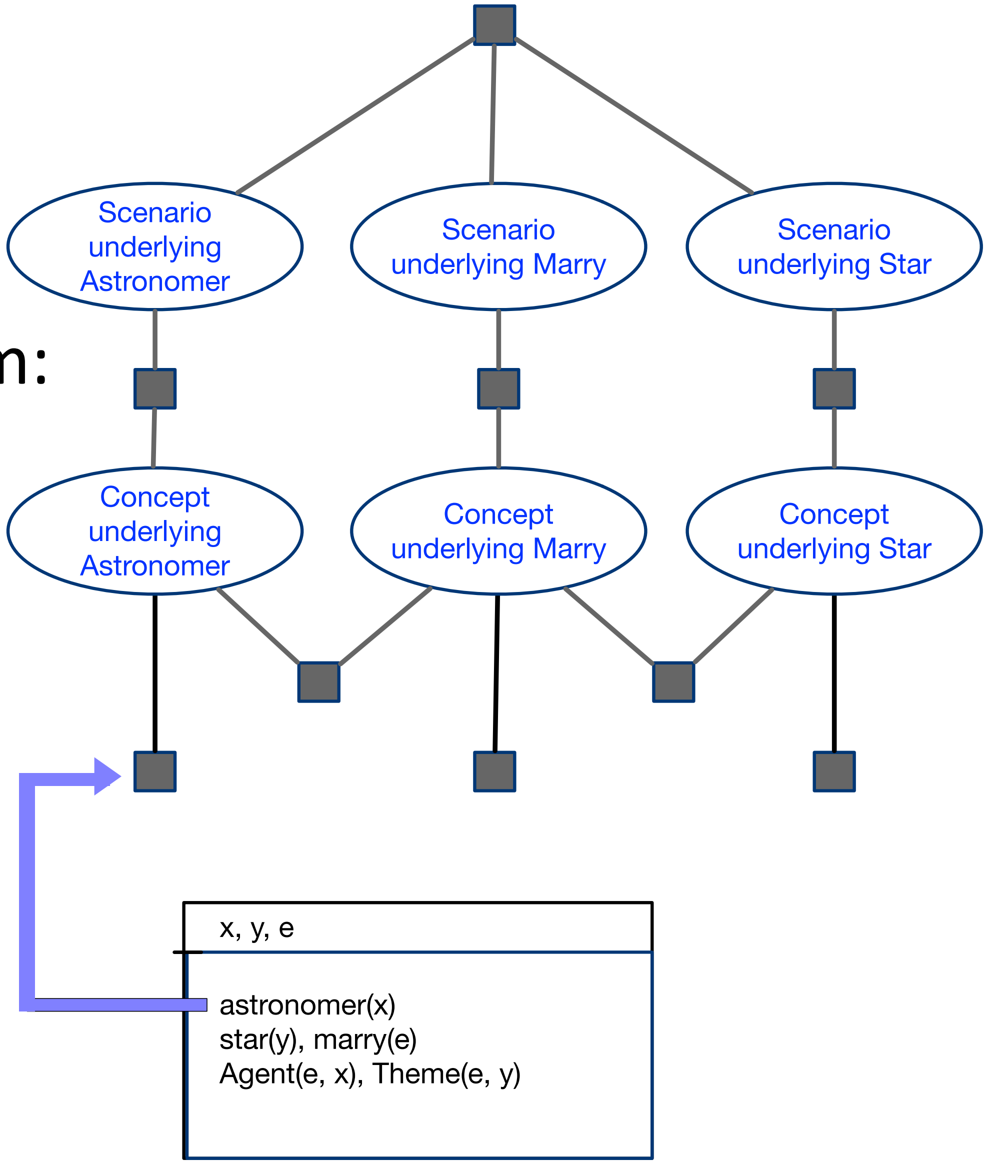
Random variables for concepts and scenarios

- Random variables that stand for concepts: Possible values are labels of lexicalized concepts, like Sun, Well-known Person for "star"
- Random variables that stand for scenarios: Possible values are labels for scenario types, like Astronomy, Movies
- Probabilistic assignment of values



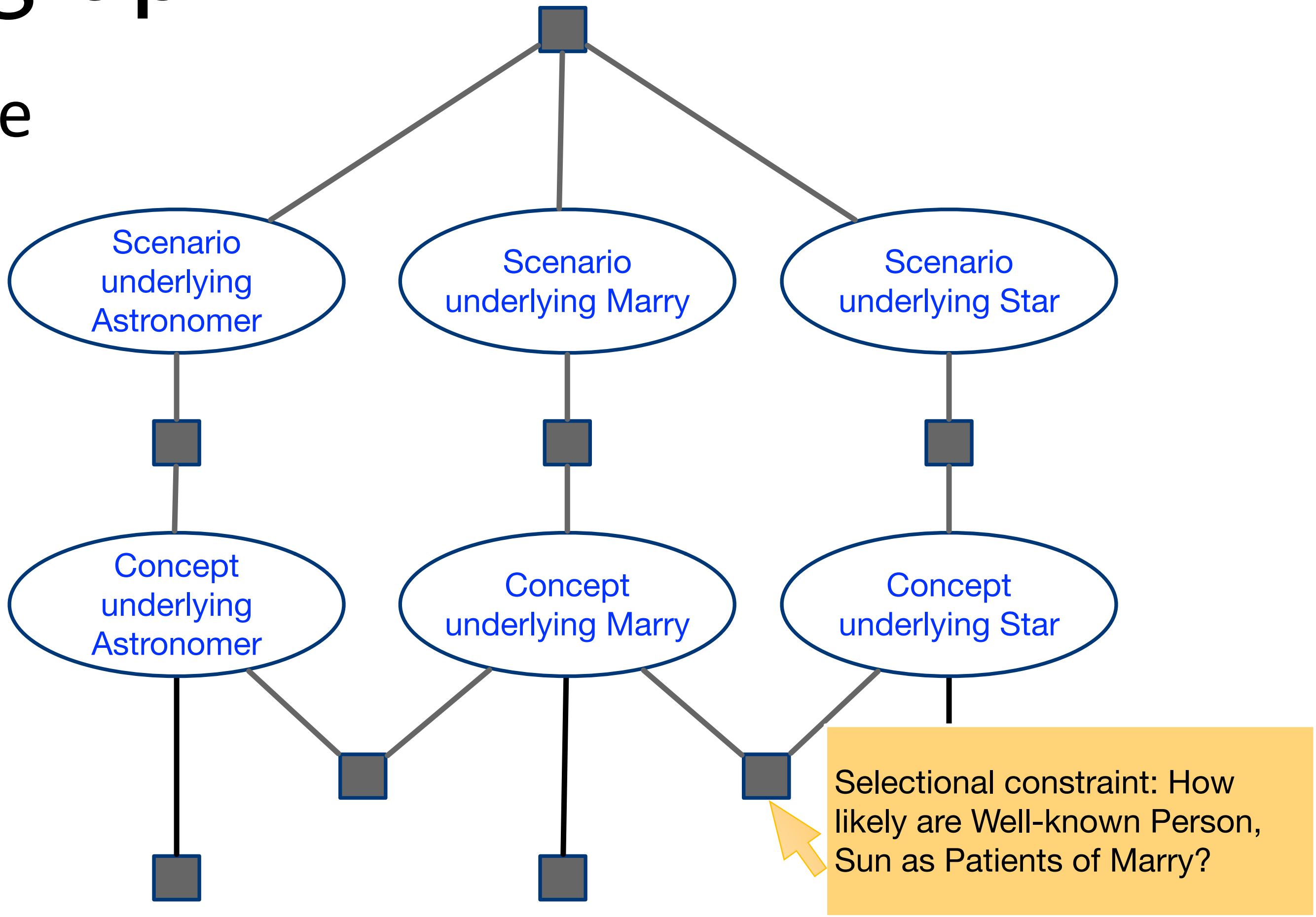
Constraints on word meaning in context as factors in the factor graph

- For every unary literal in the DRS, say astronomer(x):
  - Underlying concept for the lexical item: variable node
  - Factor constraining the concept node matching the observed predicate of the literal
- Observations here become unary factors



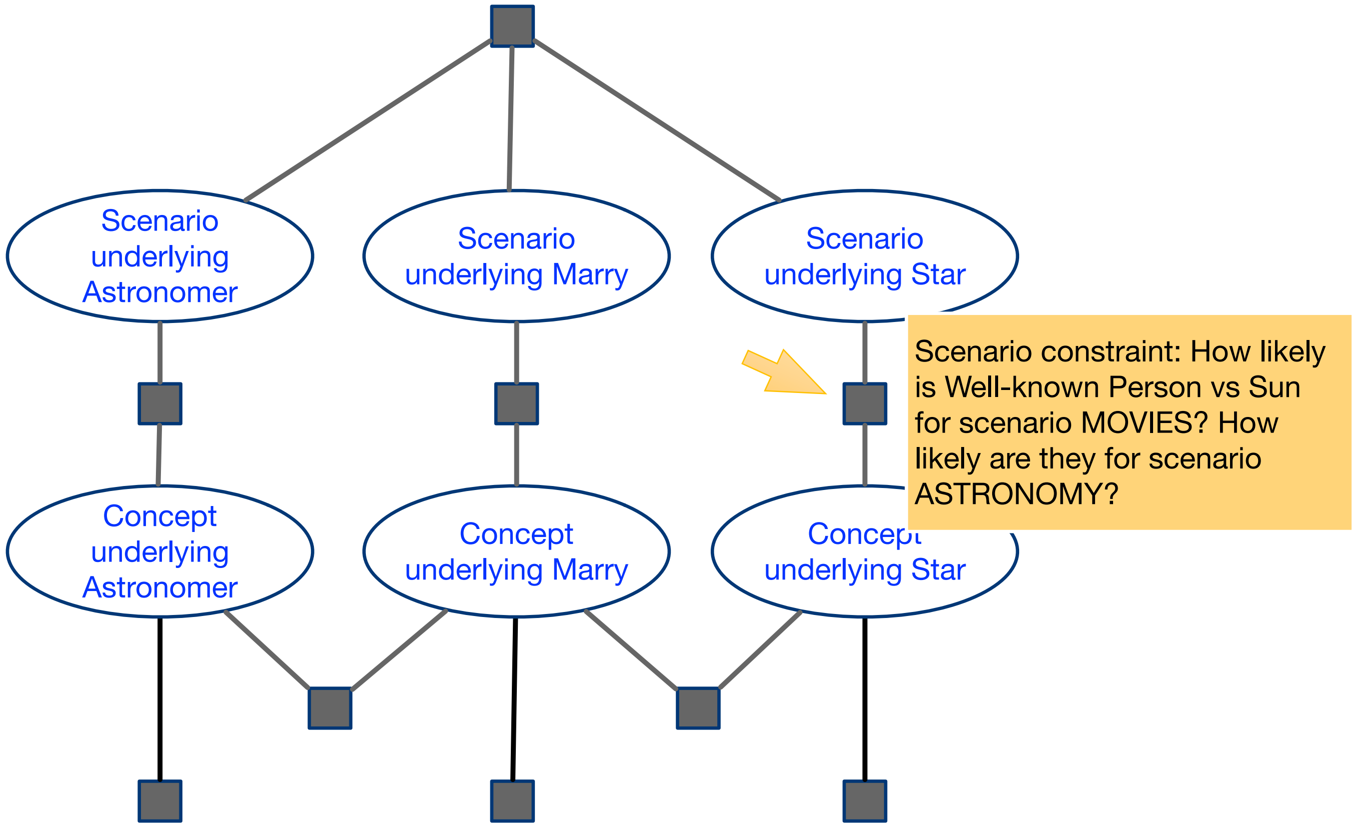
Constraints on word meaning in context as factors in the factor graph

- For every binary literal from the DRS, for example Theme(e, y): a binary factor implementing a selectional preference



Constraints on word meaning in context as factors in the factor graph

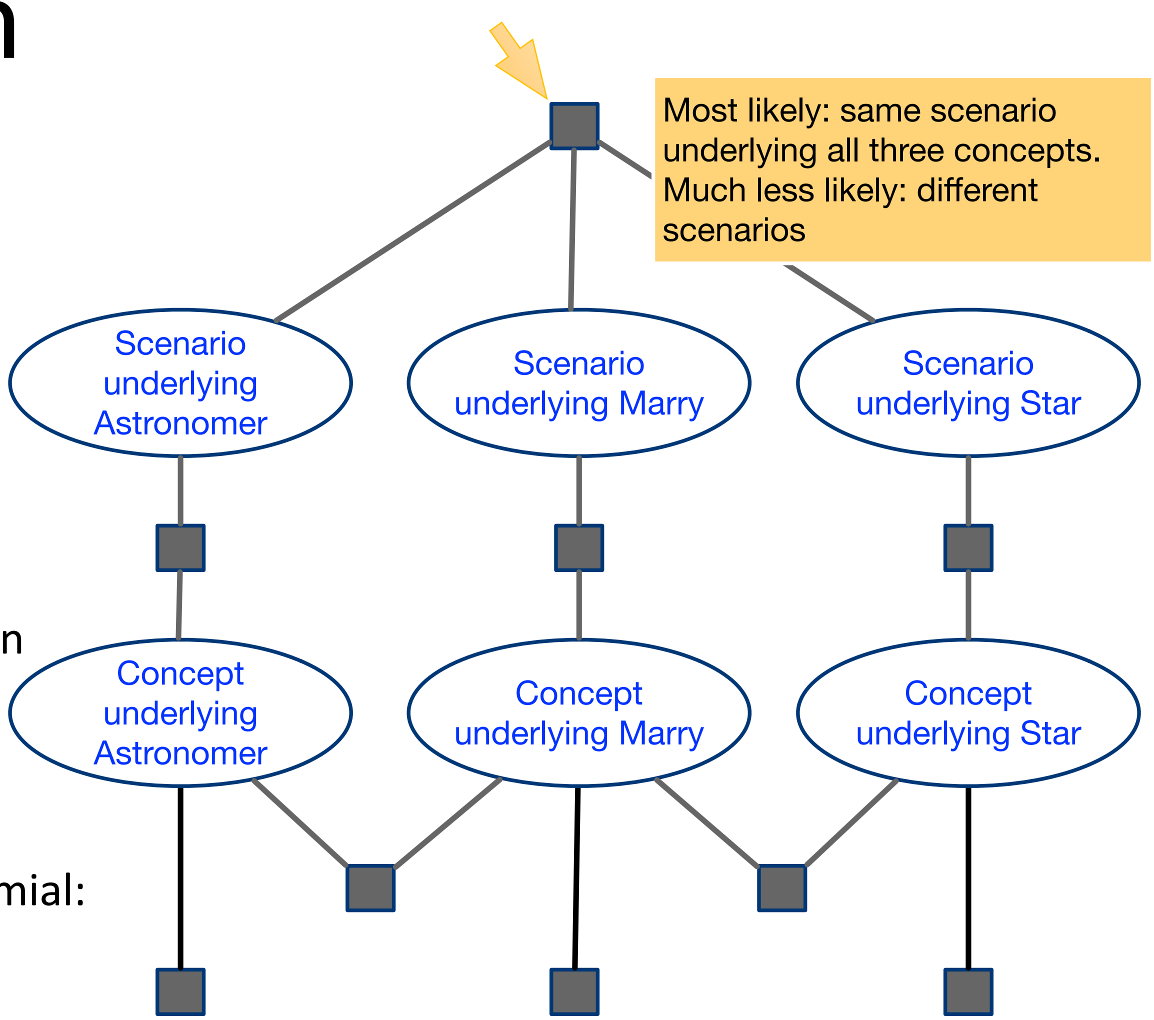
- For every variable node that stands for a concept, a variable node that stands for a scenario
- Factor constraining the two: Sun concepts are more likely to appear in an Astronomy scenario, Well-known Person concepts are more likely to appear in a Movies scenario



Constraints on word meaning in context as factors in the factor graph

- One factor connecting all scenarios:
  More likely to have the same scenario underlying all three concepts, less likely to have different scenarios
- Modeling:
  (adapting Latent Dirichlet Allocation, Blei 2003):
  - Multinomial distribution over scenarios underlying the whole sentence
  - The multinomial is drawn from a Dirichlet, a distribution over multinomials. Dirichlet parameter alpha:
    When alpha < 1, prefer to sample sparse multinomials
- Marginalizing over the multinomial distribution,
  we describe all scenarios as drawn from a Dirichlet-Multinomial:
  Probability of scenario vector **s**, with n scenarios overall, is:

$$P(\mathbf{s}, n, \alpha) = \int_p \mathrm{Mult}(\mathbf{s} \mid n, \mathbf{p})\ \mathrm{Dir}(\mathbf{p}|\alpha)\ d\mathbf{p}$$



Situation Description Systems: some main points

- Interacting constraints on meaning in context
  - selectional constraints: “marry” imposes a constraint on meaning of “star”, and vice versa
  - scenario: “astronomer” evokes something like a “space” scenario, which influences the sense of “star”
- Probabilistic inference over lexicalized concepts underlying words
  - Formally: factor graph
- Scenarios/frames/generalized event knowledge
  - Recanati, Fillmore: scripts/frames influence utterance understanding
  - McRae and colleagues, Elman: influence of generalized event knowledge on expected upcoming words

Outline

- A usage-based view on lexical meaning, with “story traces”
- Which theories of utterance meaning go well with this?
- Situation Description Systems: a framework for utterance meaning that allows for “story influences” on word meaning in context
- **Not a “literal listener” but a “habitual listener”: Less work to do for pragmatic reasoning**
- Situation Description Systems with corpus-derived parameters

Rational Speech Acts theory (RSA)

- Pragmatic reasoning: Speaker and Listener as mutually reasoning about each other
- Which state of the world w is conveyed with utterance u?
- Rational speaker (soft-)maximizes expected utility of an utterance: weighs
  - utterance cost of u versus
  - likelihood of listener of understanding w
- Rational listener weighs
  - probability of state of the world w against
  - probability of the speaker choosing u to convey w

- Grounding out the back-and-forth reasoning: Literal Listener as the starting point
- Literal listener: considers probability of possible w's that make u true

Rational Speech Acts theory: The trouble with the literal listener

- RSA is important because it opened the door to new experimental approaches in pragmatics:
    - Straightforward probabilistic formulation
    - Use participant results to estimate parameters
- **But RSA is being overused for phenomena that don't need *"me reasoning over you reasoning over me"*:**
    - Conventional metaphor (Kao et al 2014)
    - Generics and habituals (Tessler et al 2016, 2019)
- **The problem is the Literal Listener:**
If it is literal, then even conventionalize pragmatics has to go the Mutual Reasoning route
    - Also includes cases that should be viewed as primary pragmatic processes in Recanati's terminology
- White et al 2020, "Learning to refer informatively by amortizing pragmatic reasoning":
    - Amortized reasoning = learning a shortcut to mutual reasoning
    - That is basically admitting that habitual pragmatic reasoning doesn't go the Mutual Reasoning route!
- **What we need is to replace the Literal Listener by a Habitual Listener**

RSA: original formulation. Reasoning over utterance u, state of affairs w

$$\text{Pragmatic listener: } P_{L_1}(w|u) \propto P_{S_1}(u|w) \quad P(w)$$

Balance speaker's preference for u with likelihood of world state w

$$\text{Pragmatic speaker: } P_{S_1}(u|w) \propto \exp\Big(\alpha\big(\log P_{L_0}(w|u) + \text{Cost}(u)\big)\Big)$$

Balance listener's probability of "getting" w from u with utterance cost of u

$$\text{Literal Listener: } P_{L_0}(w|u) \propto P(w) \quad I(u \text{ true in } w)$$

How likely is w, given my world knowledge? And is u true in w?

The Habitual Listener

$$\text{Pragmatic listener: } P_{L_1}(w|u) \propto P_{S_1}(u|w) \quad P(w)$$

Balance speaker's preference for u with likelihood of world state w

$$\text{Pragmatic speaker: } P_{S_1}(u|w) \propto \exp\Big(\alpha\big(\log P_{L_0}(w|u) + \text{Cost}(u)\big)\Big)$$

Balance listener's probability of "getting" w from u with utterance cost of u

$$\text{Habitual Listener: } P_{L_0}(w|u) \propto P(c|u) \quad P(w|c)$$

Infer conceptual representation c underlying utterance u,
for example infer SDS graph, includes habitual pragmatic inference

How likely are we to imagine world state w from this conceptual representation?

Imagining the world

- Goodman/Lassiter:
  - **World knowledge as generative**, able to imagine situations
  - **Linguistic knowledge as discriminative**, checking truth conditions: given a situation, is this utterance true in it?
  - You can see this in the literal listener: $P_{L_0}(w|u) \propto P(w) \;\; I(u \text{ true in } w)$
- **Frame semantics point of view: Lexicon should be generative**, able to imagine situations
  - Words "evoke" frames, they imagine/generate chunks of world knowledge
- Situation Description Systems:
  - Probabilistic graphical model is generative, can estimate P(w|c)
  - Can generate additional entities and events given the scenarios in the sentence
  - Can generate entity properties given their underlying concepts: if it's a bat, it's likely furry

$$P_{L_0}(w|u) \propto P(c|u) \;\; P(w|c)$$

Outline

- A usage-based view on lexical meaning, with “story traces”
- Which theories of utterance meaning go well with this?
- Situation Description Systems: a framework for utterance meaning that allows for “story influences” on word meaning in context
- Not a “literal listener” but a “habitual listener”: Less work to do for pragmatic reasoning
- **Situation Description Systems with corpus-derived parameters**

Situation Description Systems and fine-grained representations of word meaning

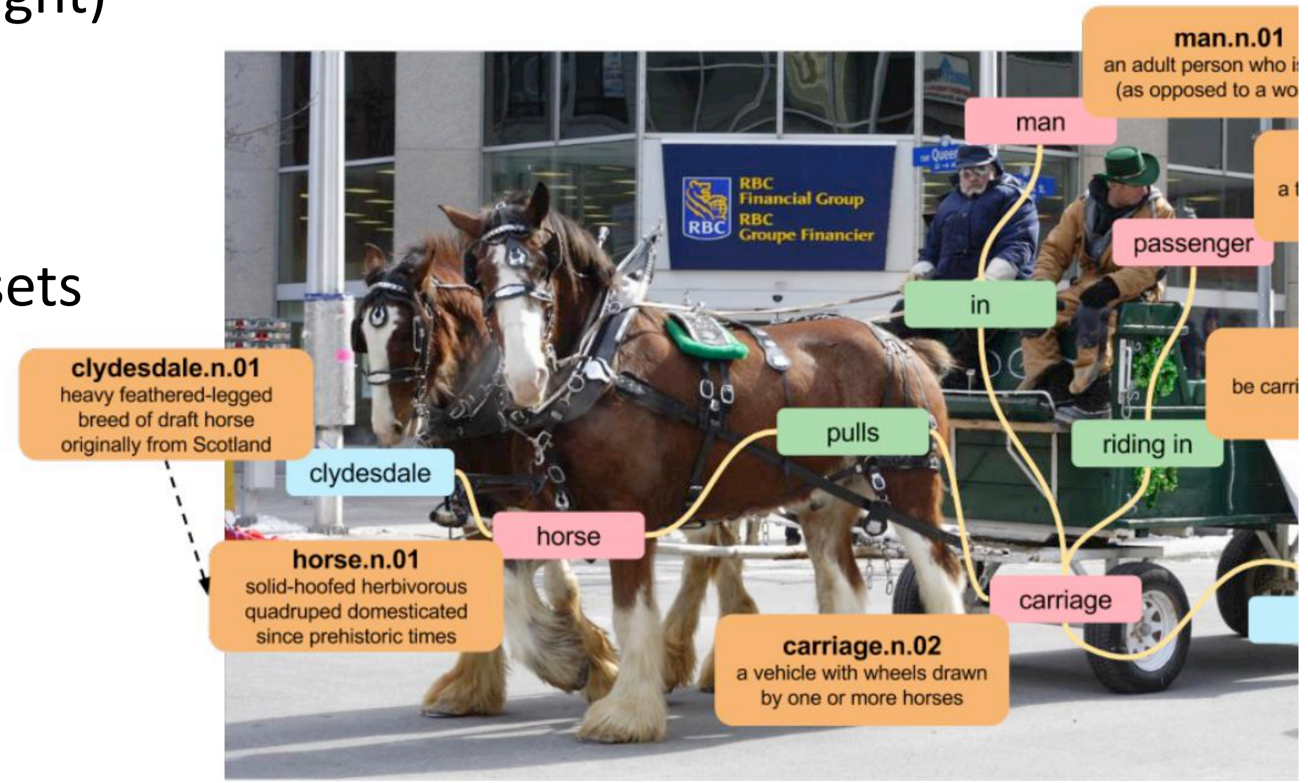
- **Fine-grained representations of lexical meaning in formal semantics**
  - Asher 2011: intensional and conceptual meaning representations
  - Zeevat et al 2014: Probabilistic feature matrices
  - Larsson 2020: Machine learning to model lexical acquisition
- **Specifically, embeddings for lexical meaning representations in formal semantics**
  - Baroni et al 2014: **the vastness of the lexicon**
  - Baroni/Zamparelli 2011, Coecke/Sadrzadeh/Clark 2010, …: Composition in semantic spaces
  - McNally 2017: Word vectors as kinds
  - Bernardy et al 2018: Inference in semantic spaces
  - Emerson 2018, 2020: Autoencoder for interactions between word meaning embeddings
- **Situation Description Systems:**
  - Learn concept parameters, scenario parameters from corpus data

Learning parameters for Situation Description Systems

- Lexicalized concepts: categorical, for inspectability
- Concepts are information-rich:
  - Selectional constraints:
    predict from centroid of seen role fillers (Erk/Padó/Padó 2010)
  - Typical properties of instances:
    predict from embeddings for concepts (Fagarasan et al 2015, Herbelot/Vecchi 2015, Rosenfeld/Erk 2022, Turton et al 2020, Chersoni et al 2021)
- Scenarios: learn using topic modeling
- For now, not end to end:
  - Individually tested, exchangeable components for implementing linguistic theories

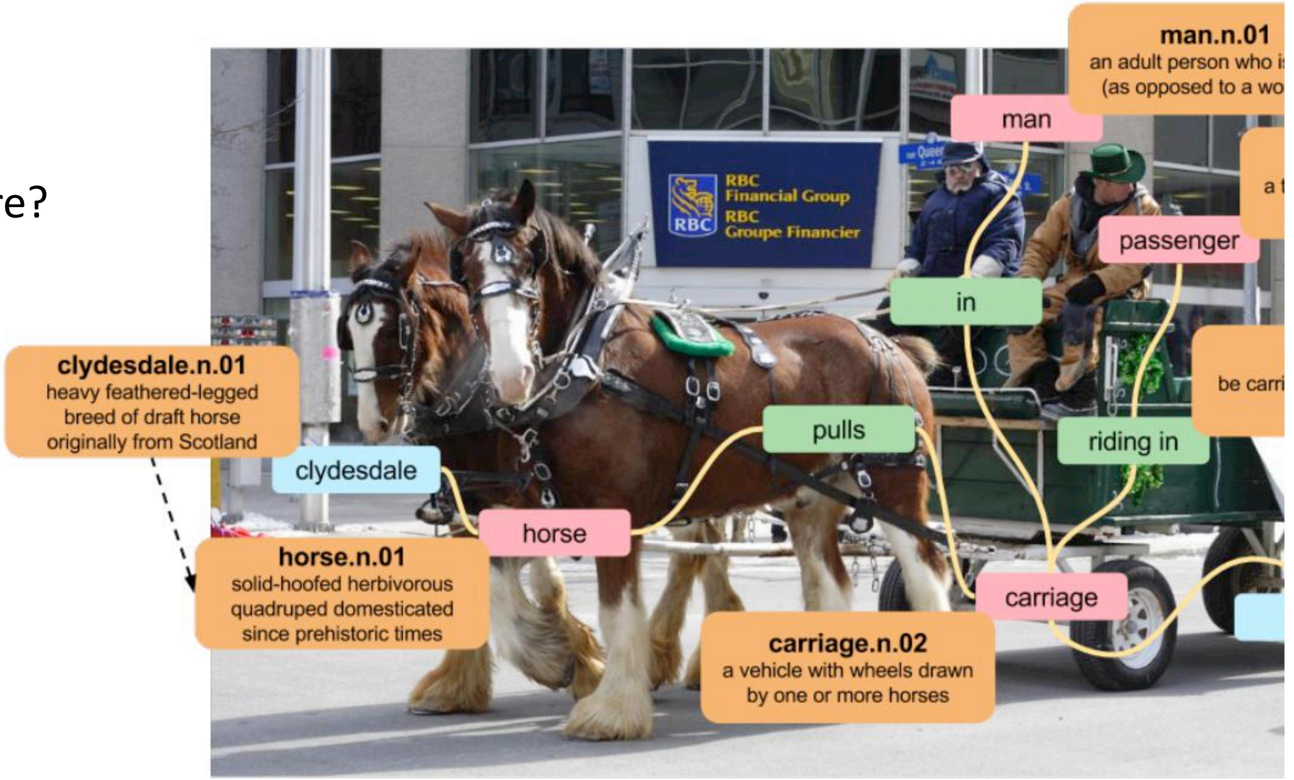
First testbed: Visual Genome

- Krishna et al, International Journal of Computer Vision, 2017
- Images hand-annotated with labels:
  - Objects (pink in the picture to the right)
  - Attributes (blue)
  - Relations (red)
- Word labels are linked to WordNet synsets (orange)
- Over 100k images
- Typically many objects per image, on average 21



First testbed: Visual Genome

- Compute word embeddings from Visual Genome labels (Herbelot 2020):
  - High-quality word embeddings from small data
- Visual Genome as grounded, situational data, good fit for model-theoretic semantics
  - Disambiguation task
  - Reference task: retrieve referent for "the Clydesdale horse"
  - Imagination task: what else could be in the picture?
- Objects have attributes: predict properties for concept instances based on embeddings
- Selectional constraints of relations, attributes: predict based on embeddings
- Scenarios based on co-occurrence in visual scene: LDA topic modeling



Scalable inference with categorical values

- Categorical concepts, scenarios
  - Recent machine learning methods: large improvements for working with continuous values
  - Categorical values, not so much
- We use pgmax: discrete probabilistic graphical models. efficient, scalable loopy belief propagation in JAX
- However:
  - Focus on scalability research for factor graphs: many nodes, few values
  - SDS: few nodes, many values, especially scenario-valued nodes
- Solutions:
  - For now, heuristics to reduce table sizes in factor graph
  - In the future:
    - Particle Belief Propagation (Ihler/McAllester 2009), sampling from messages in sum-product algorithm
    - Neural Factor Graph (Malaviya/Gormley/Neubig 2018) with topic/scenario factors predicted from neural model

The end — here’s a short summary of this talk

- (Contextualized and other) embeddings for a usage-based analysis of word meaning
  - Language models as compacted corpora
  - Social/cultural/emotional traces in embeddings
- Situation Description Systems
  - Intensional and conceptual representation of utterance meaning
  - Interacting constraints on word meaning in context
    modeled in a probabilistic graphical model
  - Including constraints from overall scenario: pragmatic, but conventionalized
- The “habitual listener”: We should model conventionalized pragmatic influences as directly, unconsciously applied, without mutual reasoning between speaker and listener
- Situation Description Systems and using machine learning to address the “vastness of the lexicon”:
  - Parameters learned from Visual Genome
  - Scaling up while retaining interpretability