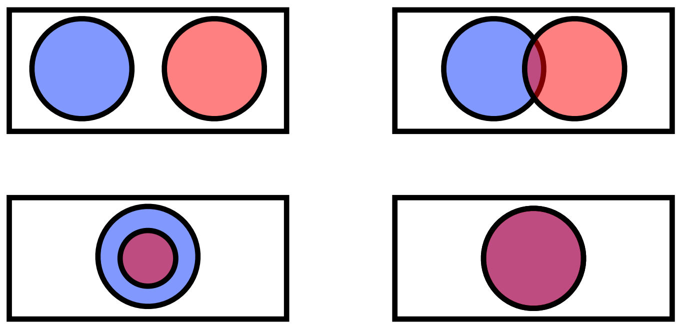
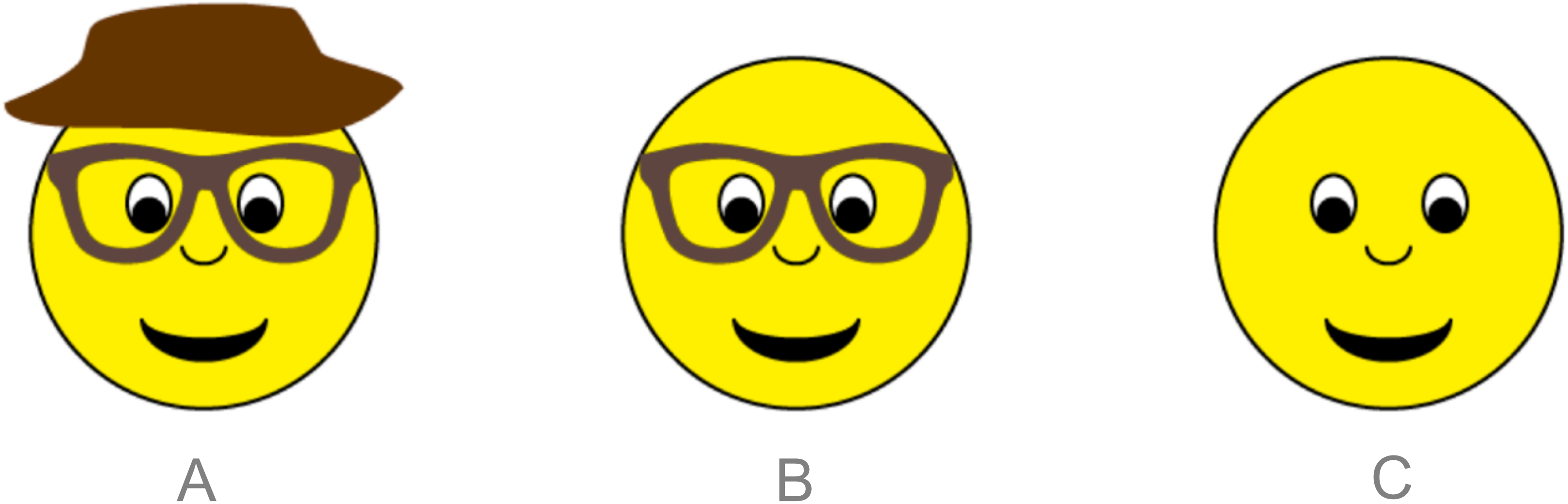
Introductory Bayesian pragmatics



9.19: Computational Psycholinguistics
22 November 2023
Roger Levy

Ad-hoc scalar inference



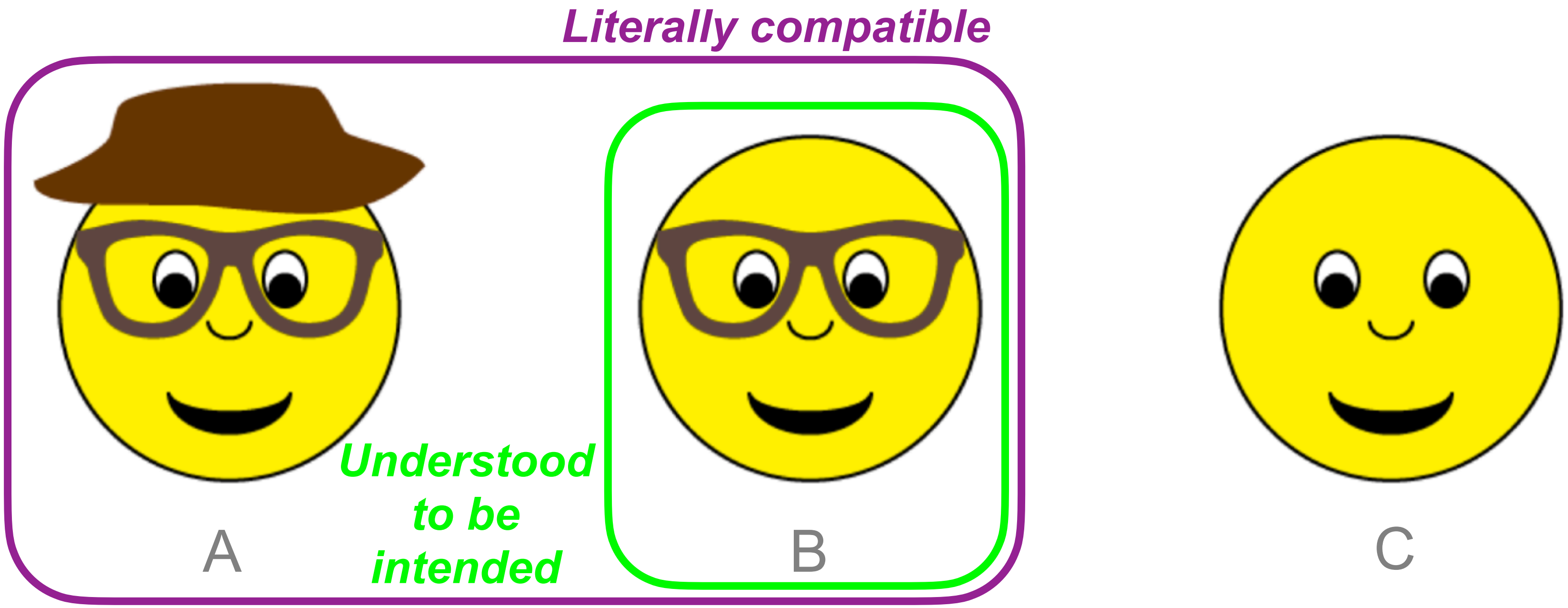
Bob can only say one word to communicate with you and he says: ***"glasses"***

Empirical finding: >75% of experimental participants choose character **B**!

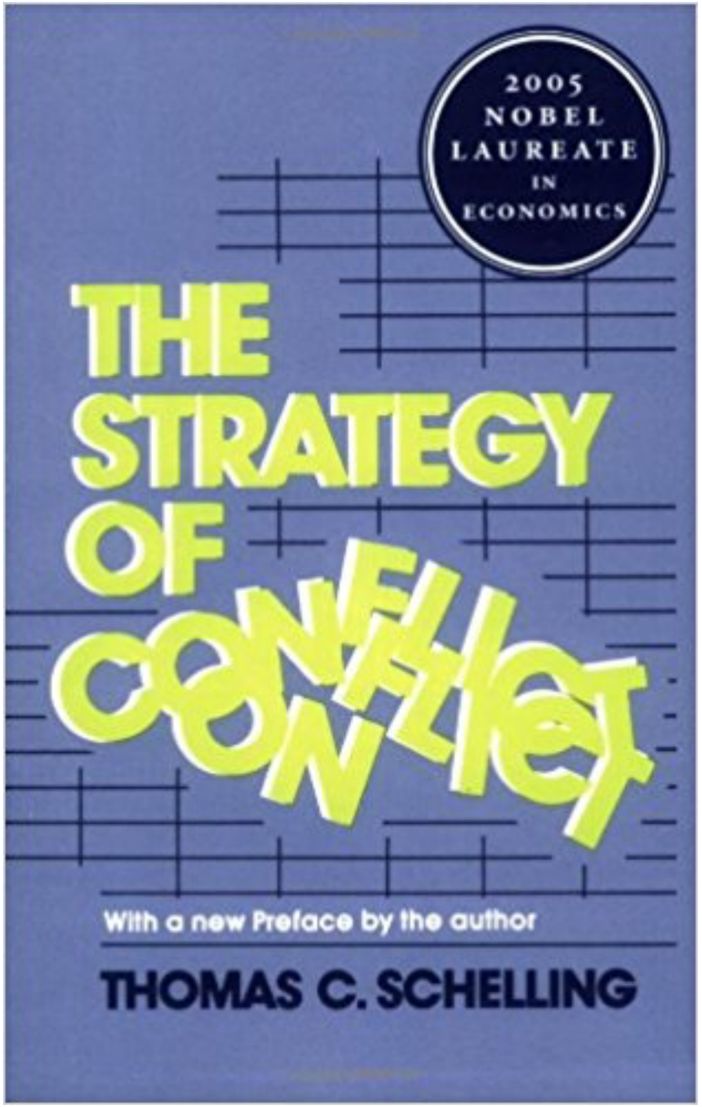
(Vogel et al., 2014)

What is *said* and what is *meant*

"glasses"



Coordination games

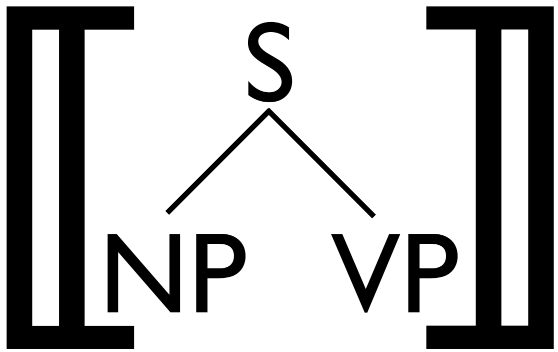


Formalizing theories of semantics & pragmatics

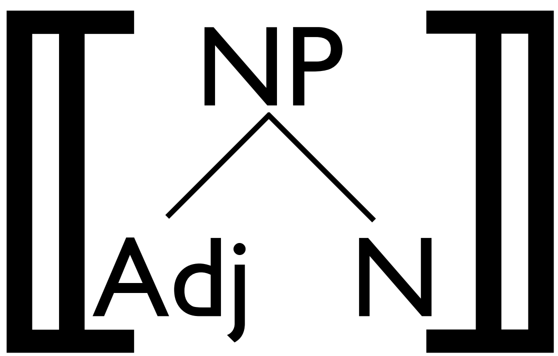
- How does human language achieve its unbounded and highly context-dependent expressive capacity?
- **Semantics**: the "literal" meanings of words and the rules of composition by which words are combined
- **Pragmatics**: how a speaker's communicative intent is inferred from literal meaning in context
  A. *I could really use a cup of coffee.*
  B. *There's a good place called Area Four nearby.*
- Probabilistic models over rich logical structures finally allow us to formalize joint semantic/pragmatic models
- Allows us to connect insights about linguistic meaning from across cognitive science—linguistics, AI, cognitive psychology, social cognition, philosophy

Semantics: principle of **compositionality**

The meaning of a complex expression is determined by the rules by which the expression is formed as applied to the meaning of the expression's subparts.



$= [\![VP]\!]([\![NP]\!])$



$= [\![Adj]\!]([\![N]\!])$

...

Pragmatics: Grice, 1975

Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. One might label this the COOPERATIVE PRINCIPLE.

Grice's maxims (in his own words)

- **Quality**: Try to make your contribution one that is true, i.e.:
  - Do not say what you believe to be false.
  - Do not say that for which you lack adequate evidence.
- **Quantity**:
  - Make your contribution as informative as is required (for the current purposes of the exchange).
  - Do not make your contribution more informative than is required.
- **Relation**: Be relevant
- **Manner**: Be perspicuous, i.e.:
  - Avoid obscurity of expression
  - Avoid ambiguity
  - Be brief
  - Be orderly

(Grice, 1975)

Generating implicatures

- Assuming that the maxims hold often allows listeners to infer meaning intentions on the part of the speaker that go beyond the literal meaning of the speaker's utterance
- These additional meaning intentions are **implicatures.**

Examples of the maxims in action

- Example:
  A. *I could really use a cup of coffee.*
  B. *There's a good place called Area Four nearby.*
- Assuming the maxims of **Quality** (be truthful) and **Relation** (be relevant) holds allows B to understand A's declarative statement as a request for information, and allows A to understand B's response as providing that information

Examples of the maxims in action

- Example: A and B are late in their senior year of high school and discussing college applications by text.
  A. *How did your applications go?*
  B. *I got into some of my top-choice schools*
- In addition to the Maxims of Quality and Relation, assuming the Maxim of **Quantity** holds allows A to infer that there were some of B's top-choice schools that B did *not* get into

Examples of the maxims in action

- Example: A performed a duet. C was in the audience and relates the experience to B, who was not.

  B. *How was the performance?*

  C. *A got all the notes in the right order.*

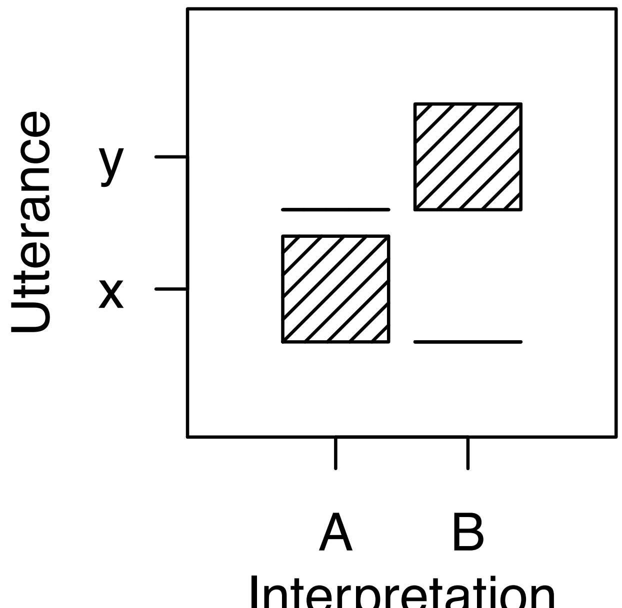
- The maxim of **Manner** licenses the inference that A's performance may not have been that great.

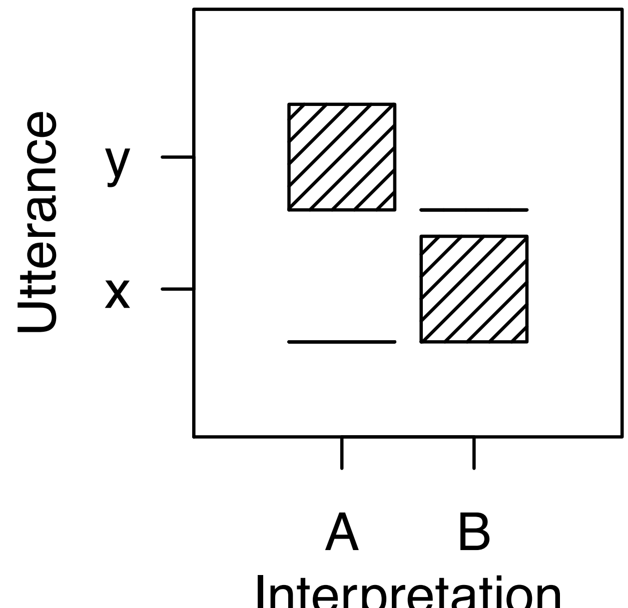
A simple communication game

- The ***speaker*** knows which of two states {**A,B**} holds of the world
- She can transmit one of two messages {*x*,*y*} to the ***listener*** to signal which world state holds
- Speaker and listener have as common ground:
  - A ***prior distribution*** on world state P(**A**), P(**B**)
  - Knowledge that messages *x* and *y* are equal in cost
  - That the game is purely cooperative

A simple communication game

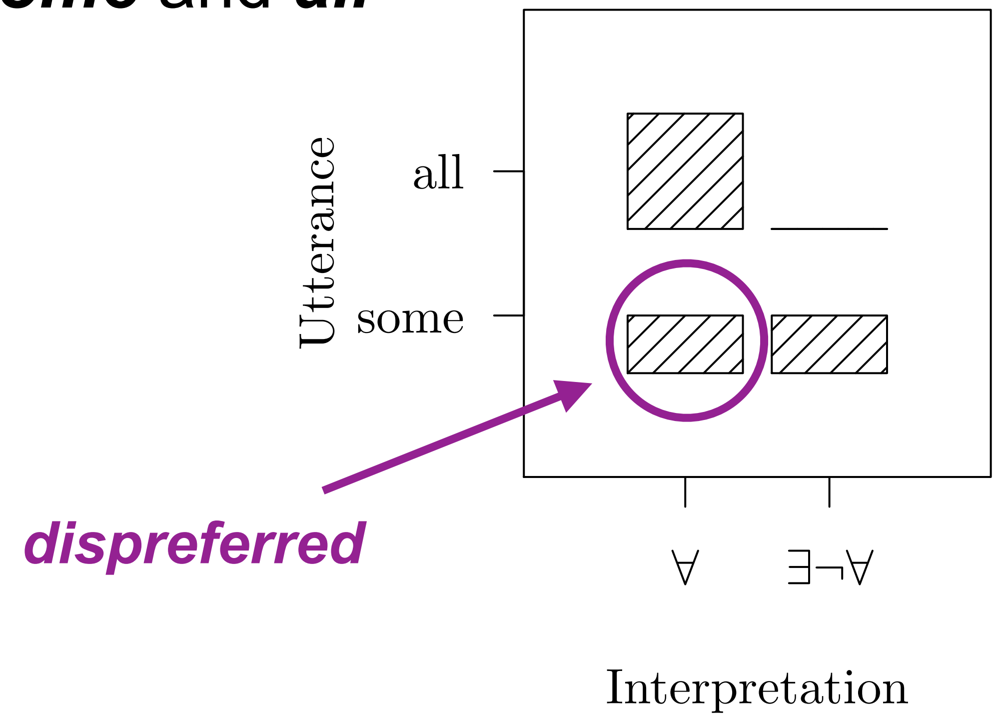
- A ***Pareto optimal strategy*** is one that delivers the highest possible reward to all players
  - A speaker who knows **A**
  - A speaker who knows **B**
  - A listener who hears ***x***
  - A listener who hears ***y***
- There are two Pareto optimal strategies in this game:





Efficiency limits of literal meaning

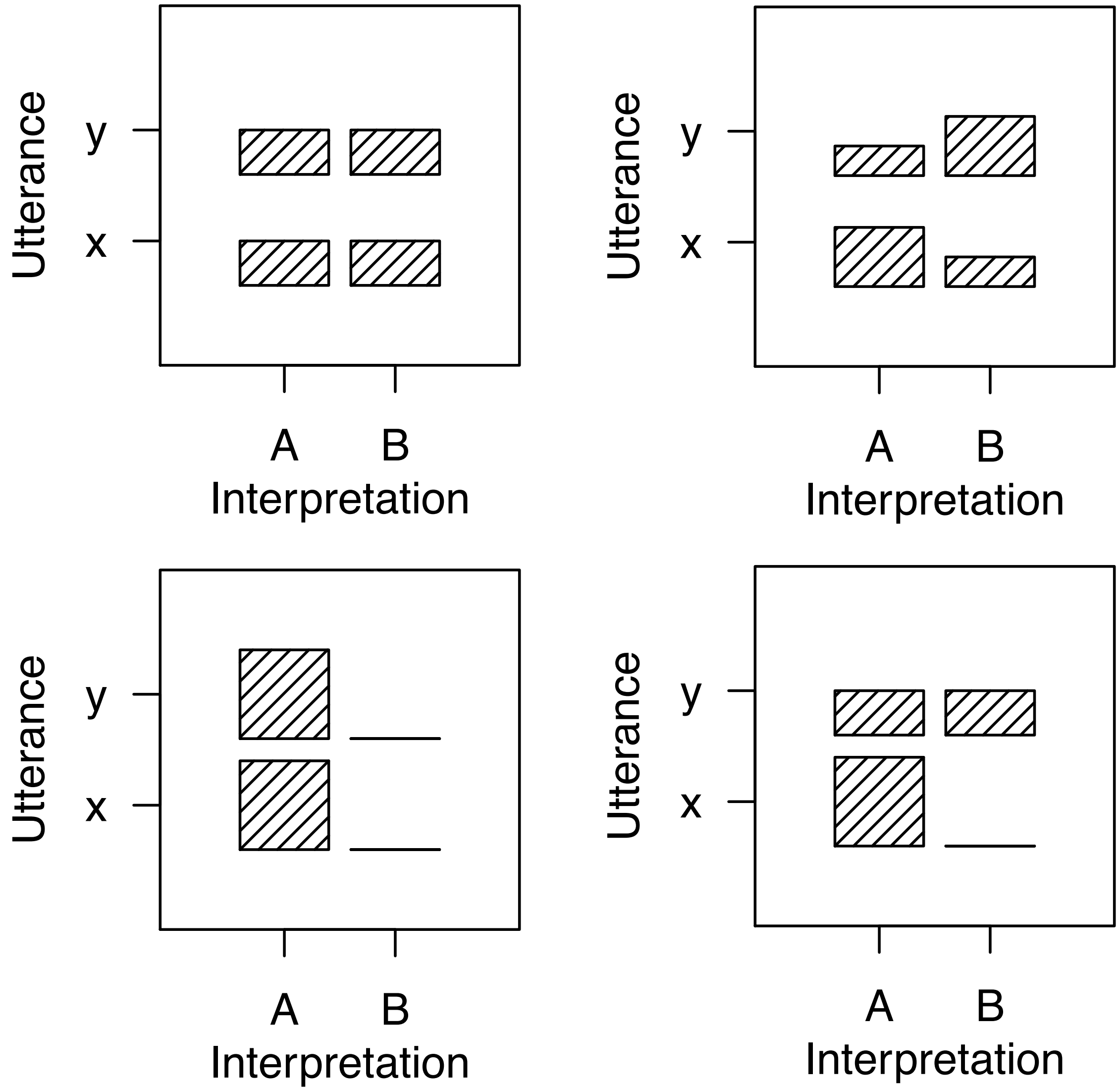
- But literal meanings don't hand us Pareto optimality
- A simple example: ***some*** and ***all***



- "Remarkable" fact: the interpretation of ***some*** that is responsible for suboptimality is dispreferred!

A simple communication game

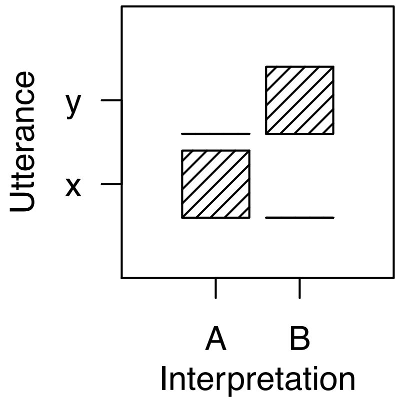
- There are also many *non*-Pareto-optimal strategies:

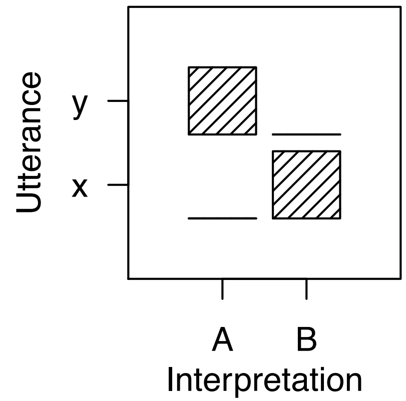


...

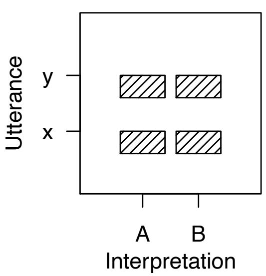
A simple communication game

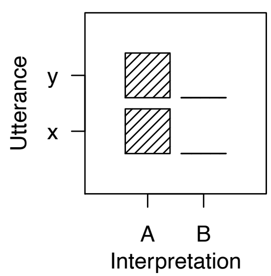
- *Efficient* communication would involve getting as close as possible to Pareto-optimal strategies…

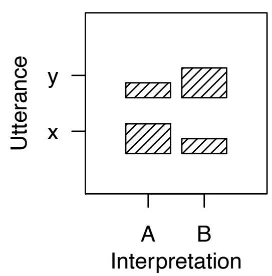


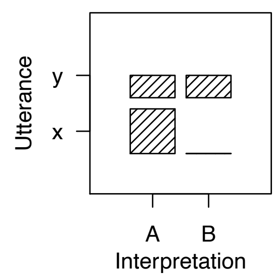


- …and away from the suboptimal strategies









…

- …but without conventions, there's no way to do this reliably!

Scalar implicature

- Consider the conventions offered us by ***some*** and ***all***
- Two meanings: $\forall$, $E\neg\forall$
- Two signals:
  - ***all*** is compatible only with meaning $\forall$
  - ***some*** is compatible with both meaning $\forall$ and meaning $E\neg\forall$
- For simplicity, assume prior $P(E\neg\forall)=P(E\neg\forall)=1/2$

Bayesian theories of pragmatics

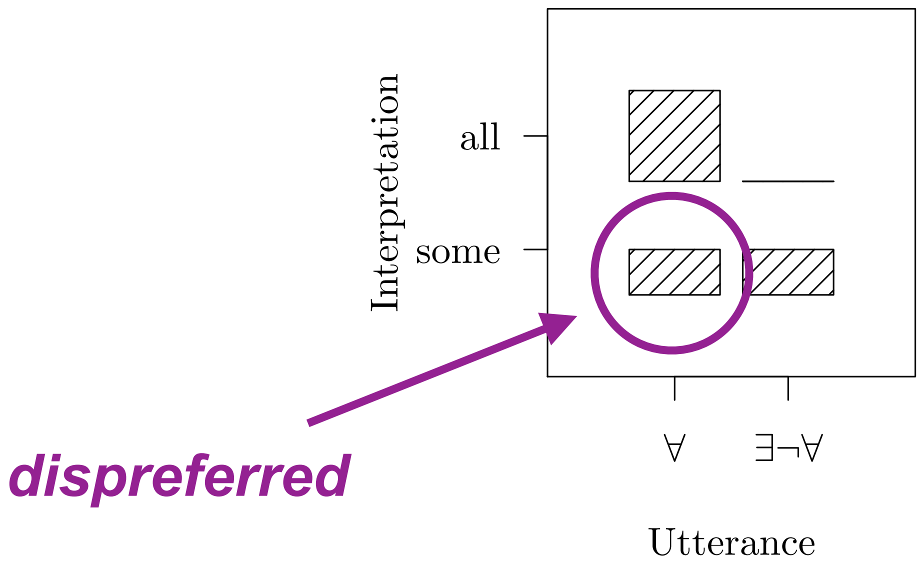
Assumptions:

- Speaker and listener beliefs represented as probability distributions over world states
- Joint communicative goal:
  - align the listener's beliefs with those of the speaker
  - but maintain brevity while doing so!
- Grammar and the literal meanings of words are common knowledge between speaker and listener
- Speaker and listener can recursively reason (probabilistically) about each other

(Frank & Goodman, 2012; Bergen et al., 2016; Goodman & Frank, 2016; Franke, 2009; Jäger & Ebert, 2009; Jäger, 2011)

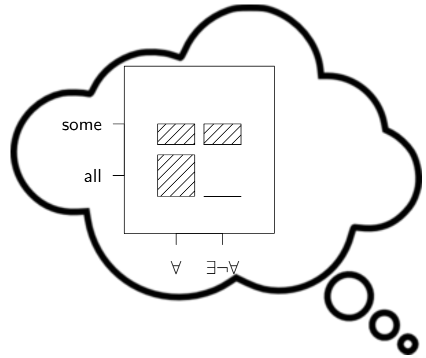
Scalar implicature

- Simple model of *literal interpretation*:
  - Listener rules out meanings incompatible with message
  - Among meanings compatible with message, prefers those with higher prior probability
- Literal interpretation matrix for *some/all*:



- This is non-Pareto—!
- —and it fails to capture human preferences

The Rational Speech-Act (RSA) model



m	P(*u*|*m*)
∃¬∀	?
∀	?

Modeling hypothesis: speakers are ***softmax decision-theoretic agents*** *that prefer* ***high-utility utterances:***

$$\text{Utility}^{(1)}(u;m) = \log P^{(0)}_{\text{Listener}}(m|u) - c(u)$$

How ***surprised*** *would* ***u*** *leave the listener about my intended meaning* ***m****?*

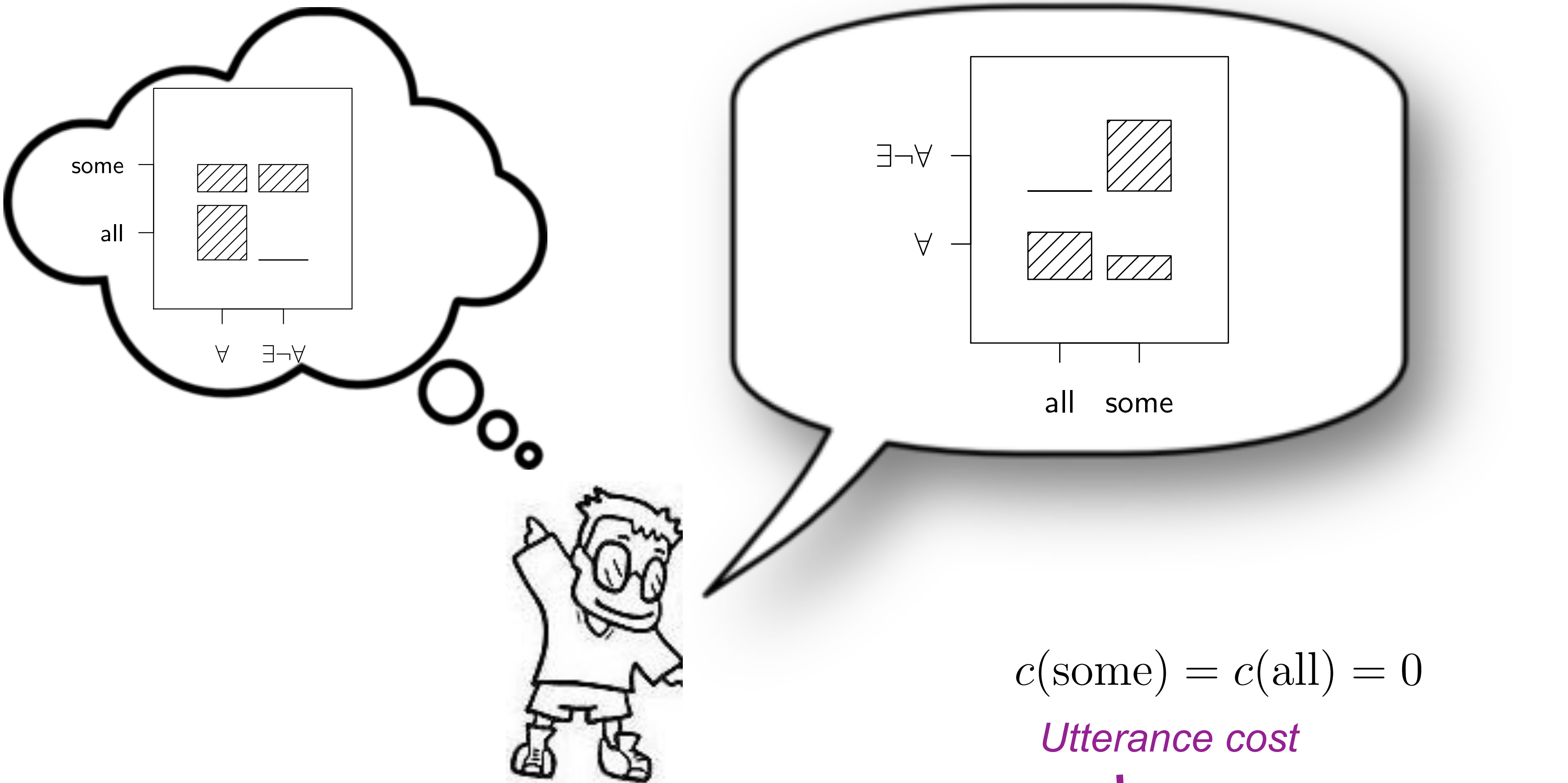
How intrinsically ***costly*** *(time-consuming, effortful, taboo, …) is the utterance* ***u****?*

$$P^{(1)}_{\text{Speaker}}(u\,|\,m) \propto \left(\exp\left[\text{Utility}^{(1)}(u;m)\right]\right)^{\lambda}$$

Softmax optimality parameter

(Frank & Goodman, 2012; Goodman & Frank, 2016)

The Rational Speech-Act (RSA) model



$$c(\text{some}) = c(\text{all}) = 0$$

Utterance cost

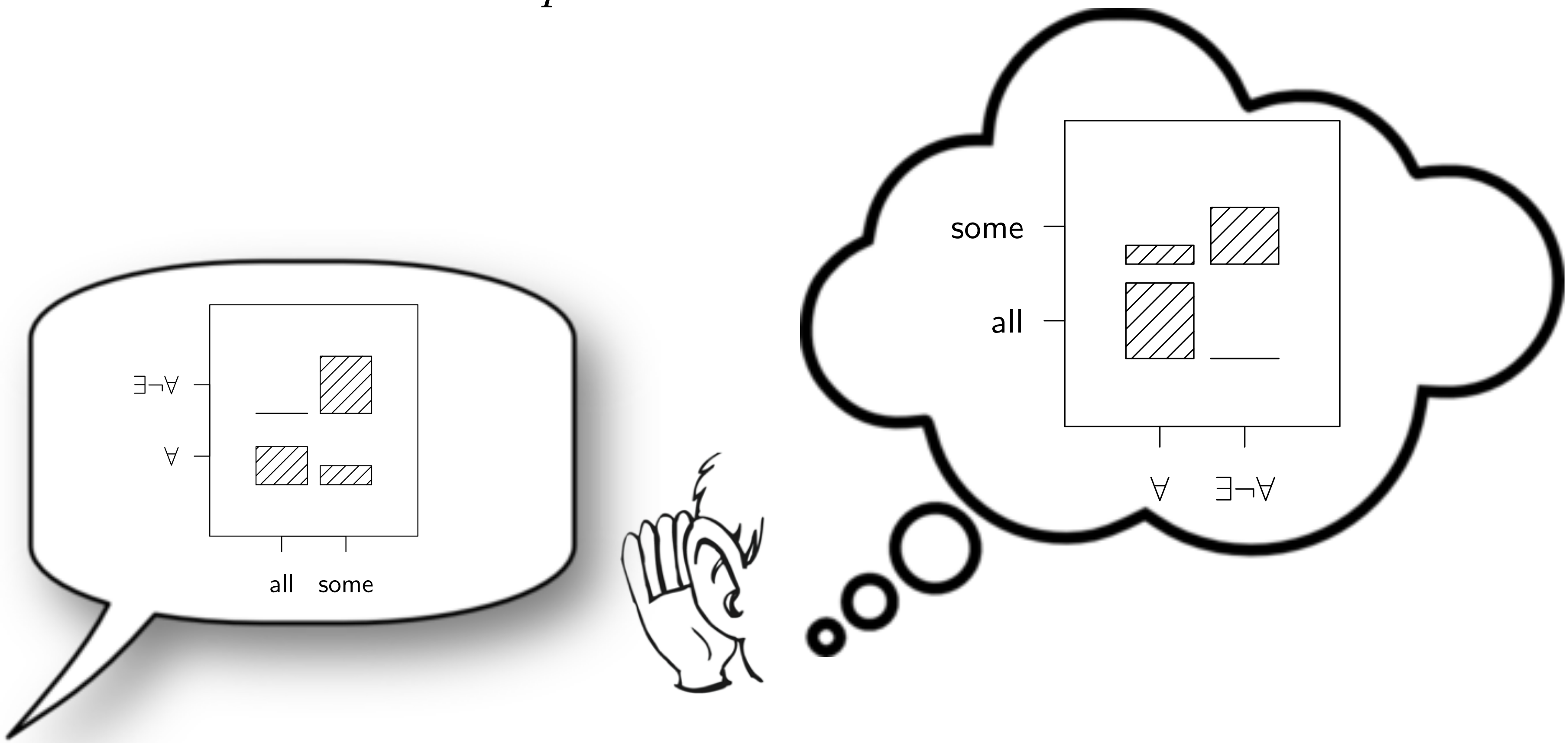
$$P^{(1)}_{Speaker}(u|m) \propto \left[P^{(0)}_{Listener}(m|u)e^{-c(u)}\right]^{\lambda} \qquad \lambda = 1$$

"Greedy optimality" parameter

(Frank & Goodman, 2012;
Goodman & Stühlmuller, 2013)

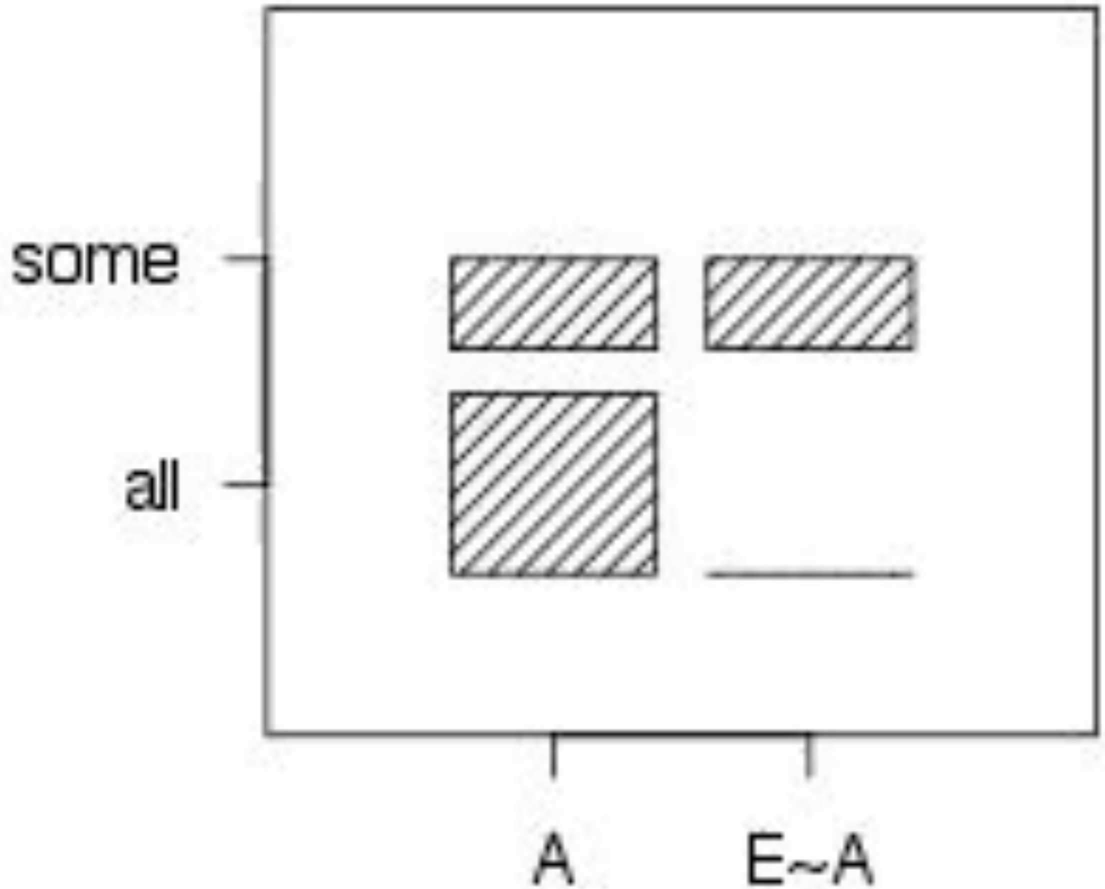
Scalar implicature in RSA: listening

$$P^{(1)}_{Listener}(m|u) \propto P^{(1)}_{Speaker}(u|m)P(m)$$



Speaker—listener recursion in RSA

- The process of recursion strengthens the implicature



Conceptual framing

- Speaker and listener got (close) to a Pareto-optimal strategy by combining two ingredients:
  - Language knowledge (lexicon/grammar) as the *raw materials* for initial solutions to the communication game
  - General principles of socio-cognitive reasoning to craft these raw materials into more efficient solutions
- These two ingredients together allow discourse participants to do *so much more* than either one alone

Levinson's (2000) typology of implicature

Q-implicature (Horn's Q)	M-implicature (Horn's "division of pragmatic labor")	I-implicature (Horn's R, sort of)
What isn't said isn't meant	*Align utterance simplicity with situation stereotypicality*	*Interpret utterances as the prototypical case*
Pat has three children ↓ Pat has **exactly three** children	I started the car ↓ ...by **just** turning the key	The cup is on the table ↓ It's **in contact with** the table

Q/I Tradeoff

I injured a child ↓ I didn't injure **my** child	I got the car to start ↓ ...I needed to do **more than just** turn the key	I injured a finger ↓ I injured **my own** finger

Can we explain this typology from basic principles in a probabilistic pragmatic framework, respecting linguistic form, semantic composition, and world knowledge?

Q/I tradeoff in rational speech-act theory

I injured a child→it WASN'T my child

$$c(\text{my}) = c(\text{a}) = 0$$

$$c(\text{someone else's}) = 1$$

$$P(\text{MINE}) = \frac{1}{2}$$

I broke a finger→it WAS my finger

Assumed alternatives:

*I injured **my** child*

*I injured **a** child*

*I injured **someone else's** child*

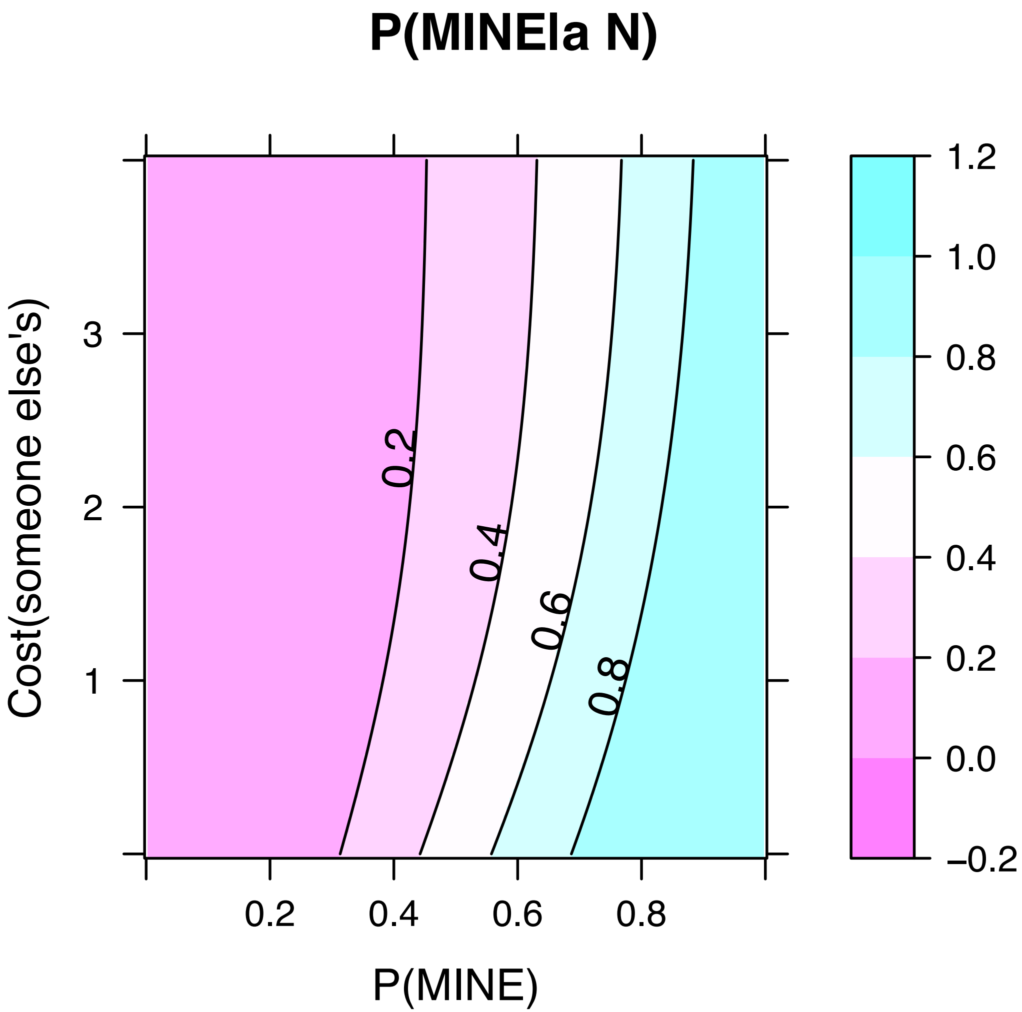
$$c(\text{my}) = c(\text{a}) = 0$$

$$c(\text{someone else's}) = 1$$

$$P(\text{MINE}) = \frac{5}{6}$$

Q/I tradeoff in rational speech-act theory

- Prior probability and simplicity trade off against one another
- But they aren't symmetric!



A rich testbed for exploring Q/I tradeoff

The man injured a finger *The man injured a child*

His? **Someone else's?**

- Five predictions from the rational speech-act model:
  1. Judgments should track prior event probabilities

     The man broke a nose *The python broke a nose*
  2. Judgments should be *other* skewed relative to prior
  3. Relational nouns should favor *own* judgments

     The man injured a child *The father injured a child*
  4. "Only-one-of" nouns should favor *other* judgments

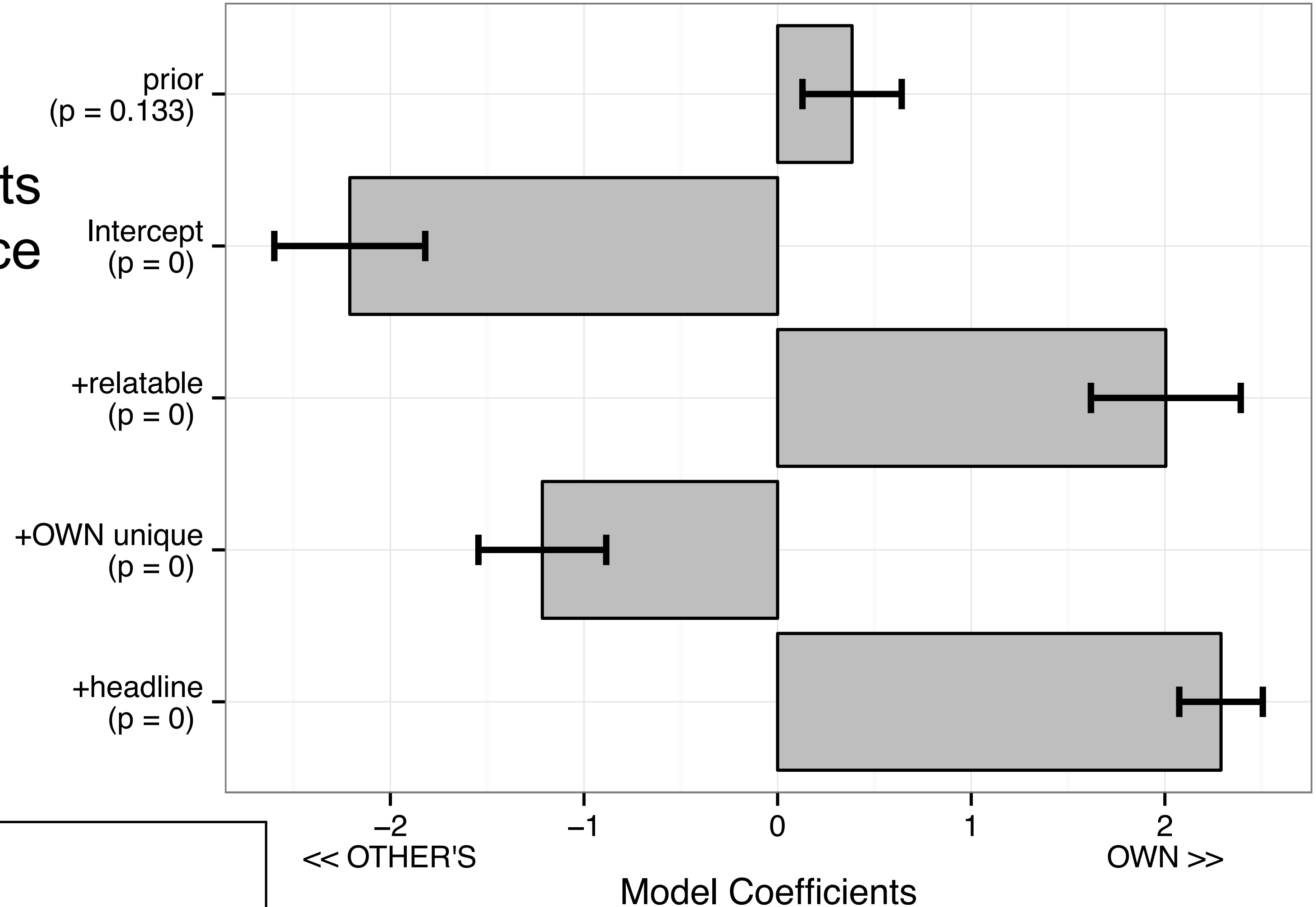
     The man broke a finger *The man broke a nose*
  5. Allowing null determiners should favor *own* judgments

     The man injured a child *Man injured child*

(Poppels & Levy, 2015)

A rich testbed for exploring Q/I tradeoff

- 1348 judgments of 108 sentence prompts
- Multivariate mixed-effects logistic regression analysis



1. Effect of prior
2. Overall *other* skew
3. Relational nouns favor *own*
4. "Only-one-of" nouns favor *other*
5. Null determiners favor *own*

(Poppels & Levy, 2015)

Adjectives: a range of semantic types

- **Intersective**: *living*, *blue*
- **Scalar/Gradable**:
  - **Relative**: *short*, *expensive*
  - **Absolute**: *dangerous*, *full*

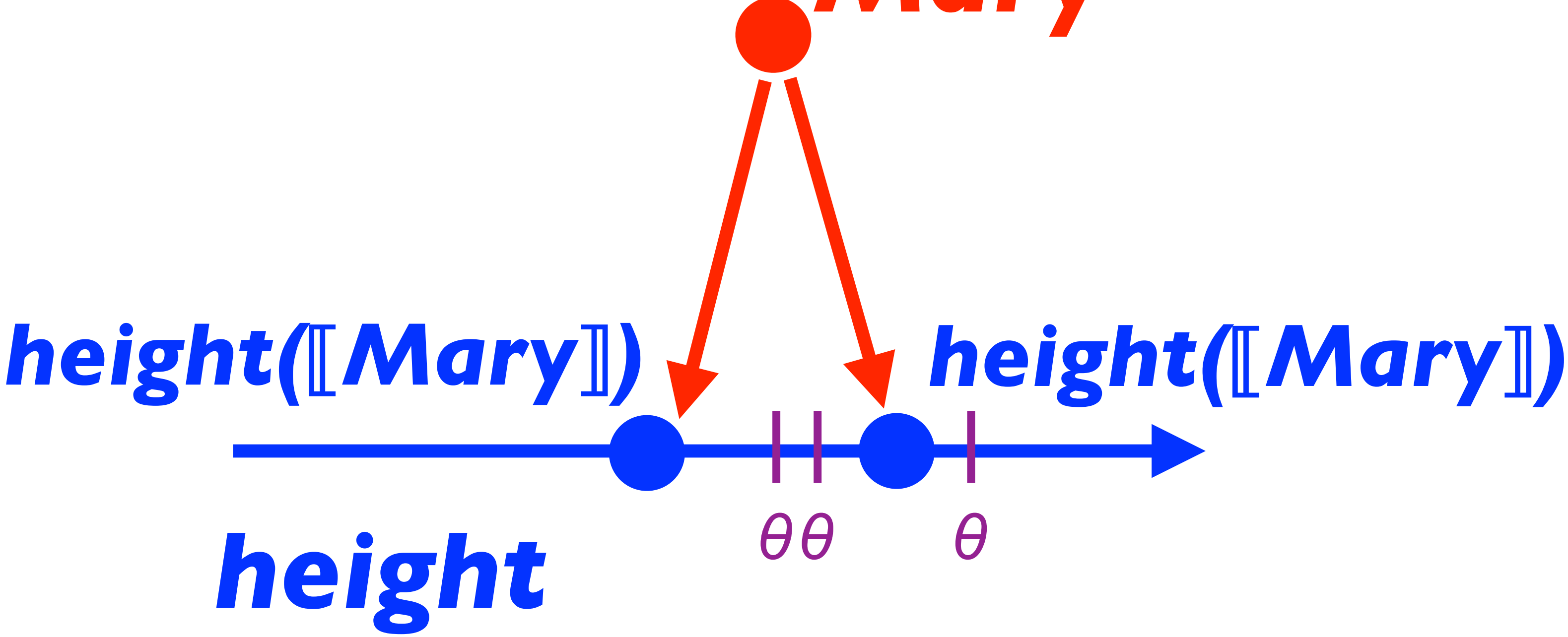
Today

- **Non-intersective**: *possible*, *alleged*
- **Anti-intersective**: *former*, *counterfeit*

Degree semantics for scalar adjectives

- The meaning of a scalar adjective like *tall* does two things
  1. Projects a referent onto some **value** on a *scale*
  2. Predicates that that **value** is greater than some **threshold** θ

Mary is tall TRUE FALSE



(Kennedy, 2007)

Observations regarding degree semantics

- Differences in scale structure can predict validity of compositions

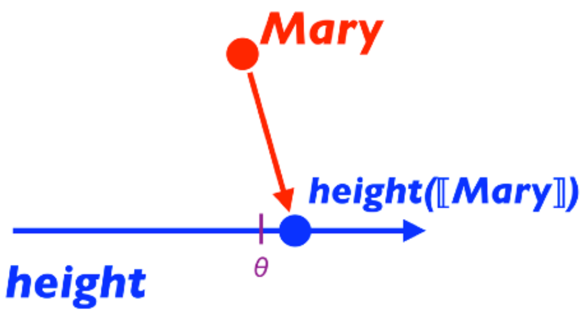
fullness

✓ *The glass is perfectly full.*
✓ *The glass is perfectly empty.*

danger

✓ *The neighborhood is perfectly safe.*
✳ *The neighborhood is perfectly dangerous.*

What the degree semantics doesn't say



- This is a very elegant model
  - The abstractness of the model allows for context-sensitivity
- But it doesn't say *how* this context-sensitivity is achieved!
  - How does *tall elephant* turn out to mean something different from *tall mouse*?
  - How can the same *individual* be evaluated as either tall or not tall in different contexts?

Stephen Curry is tall.

Stephen Curry is a tall basketball player.

(Stephen Curry is 6'2"; this is the 12th percentile of NBA player heights)

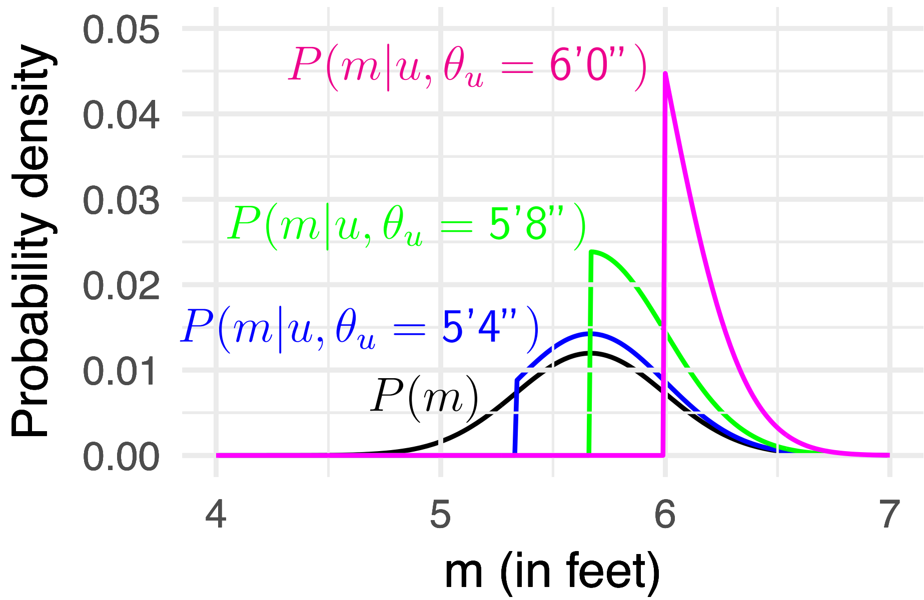
Towards a pragmatic model for scalar adjectives

- Desiderata
  - Inference on a continuum of possible scalar values
  - A threshold representation

The Lassiter & Goodman model

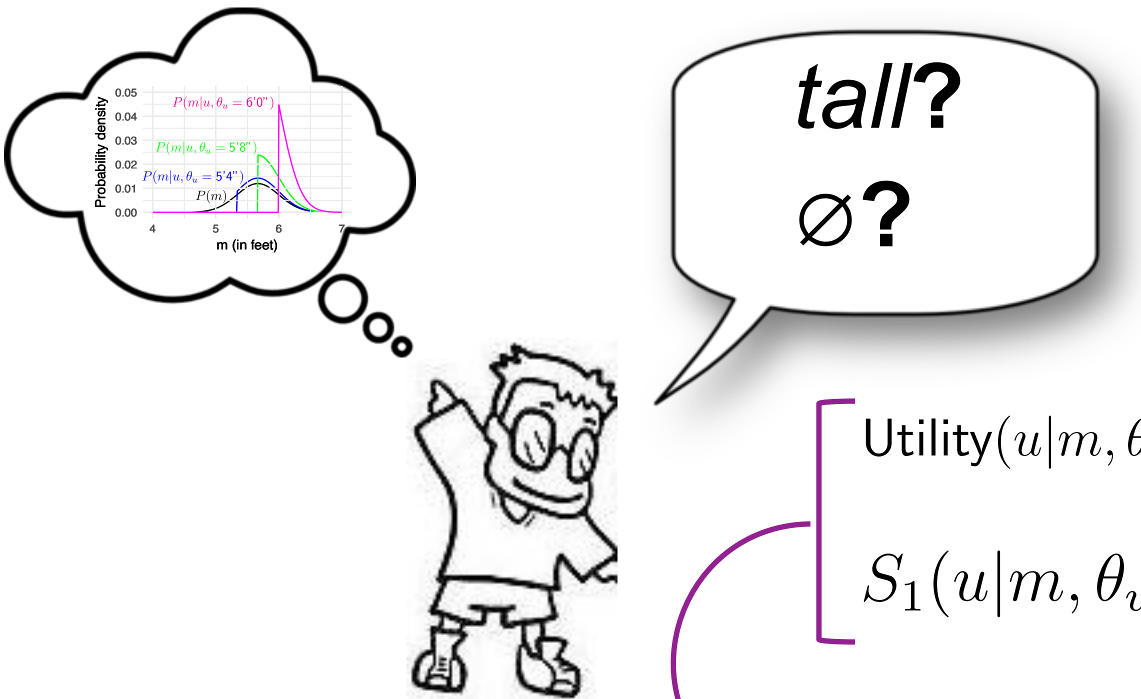
- The literal-listener model of interpretation:

$$L_0(m|u,\theta) \propto \begin{cases} P(m) & m \geq \theta \\ 0 & \text{otherwise} \end{cases}$$



A speaker model

- Assume a set of *alternative utterances* available to speaker
  - For "Pat ate some of the cookies", alternatives were *some/all*
  - For "I injured a finger", alternatives were *a/my/someone else's*
- Here, we assume alternatives (to start) *tall* and **silence** (∅)



Lassiter & Goodman's cost assumption:

$$\text{cost}(\text{tall}) = \text{cost}(\emptyset) + 2$$

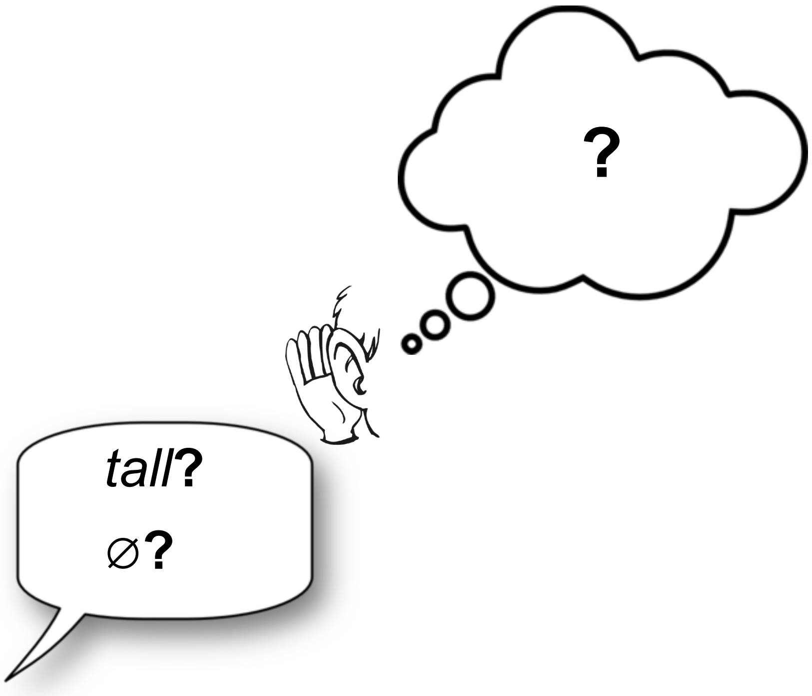
$$\text{Utility}(u|m,\theta_u) = \log L_0(m|u,\theta_u) - \text{cost}(u)$$

$$S_1(u|m,\theta_u) \propto e^{\text{Utility}(u|m,\theta_u)}$$

$$S_1(u|m,\theta_u) \propto \frac{L_0(m|u,\theta_u)}{e^{\text{cost}(u)}}$$

A pragmatic listener

Pragmatic listener is a standard Bayesian comprehender:



$$L_1(m, \theta_u | u) \propto S_1(u | m, \theta_u) P(m, \theta_u)$$

What do we do with this joint distribution?

Proposal: they are conditionally independent…

$$L_1(m, \theta_u | u) \propto S_1(u | m, \theta_u) P(m) P(\theta_u)$$

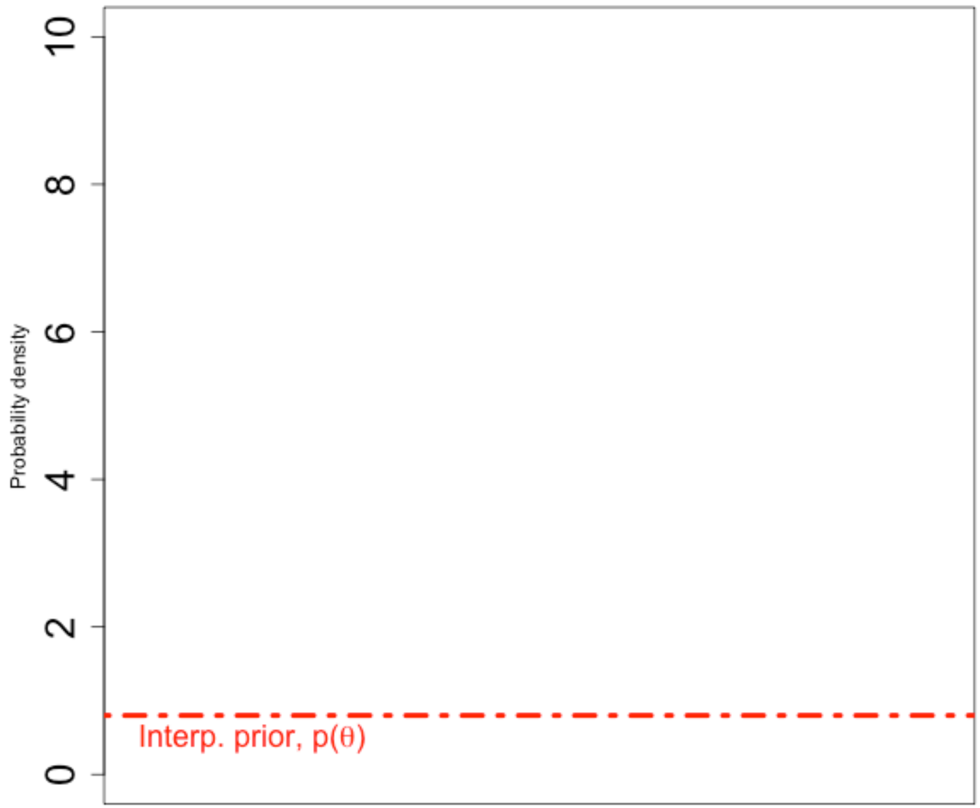
…and θ_u has a ***uniform*** *prior:*

$$L_1(m, \theta_u | u) \propto S_1(u | m, \theta_u) P(m)$$

This is a proposal of non-trivial theoretical depth and interest; let's discuss!

Visualizing pragmatic listener inferences

$$L_1(m, \theta_u|u) \propto S_1(u|m, \theta_u)P(m)\underline{P(\theta_u)}$$

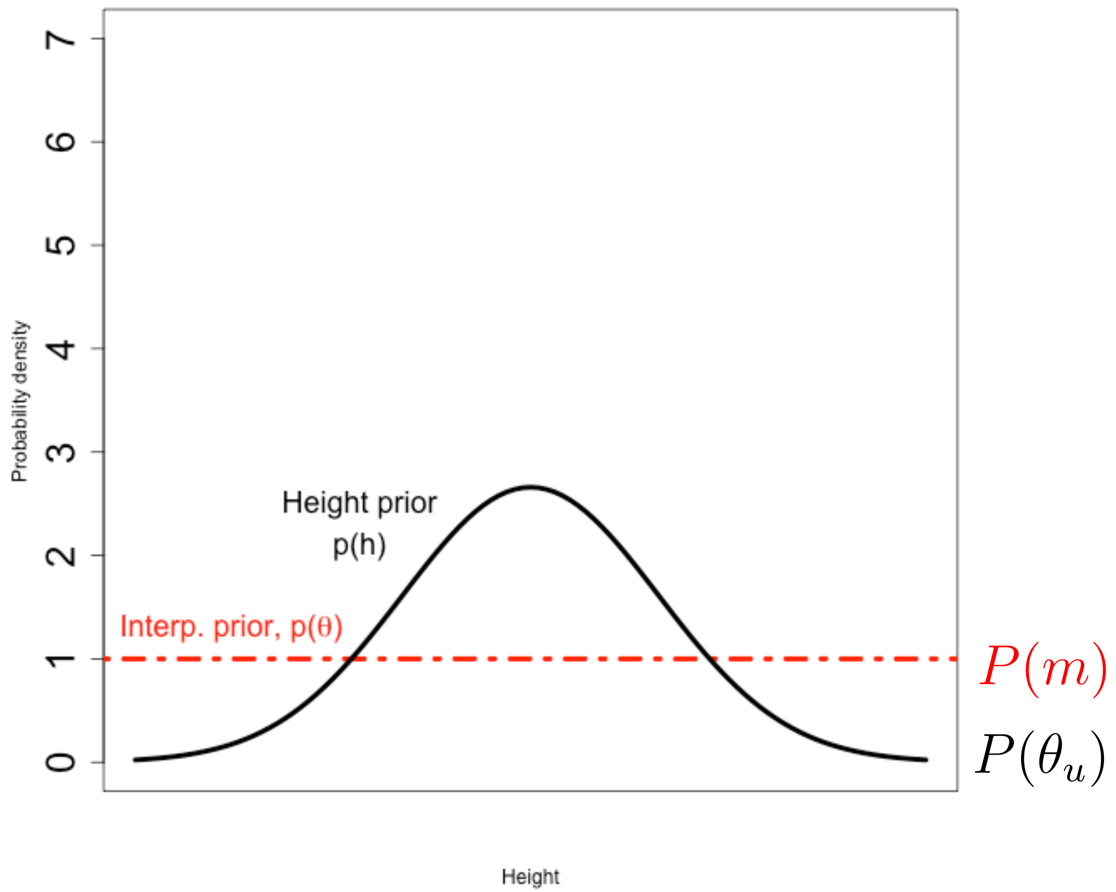


$P(\theta_u)$

(Figures due to Dan Lassiter)

Visualizing pragmatic listener inferences

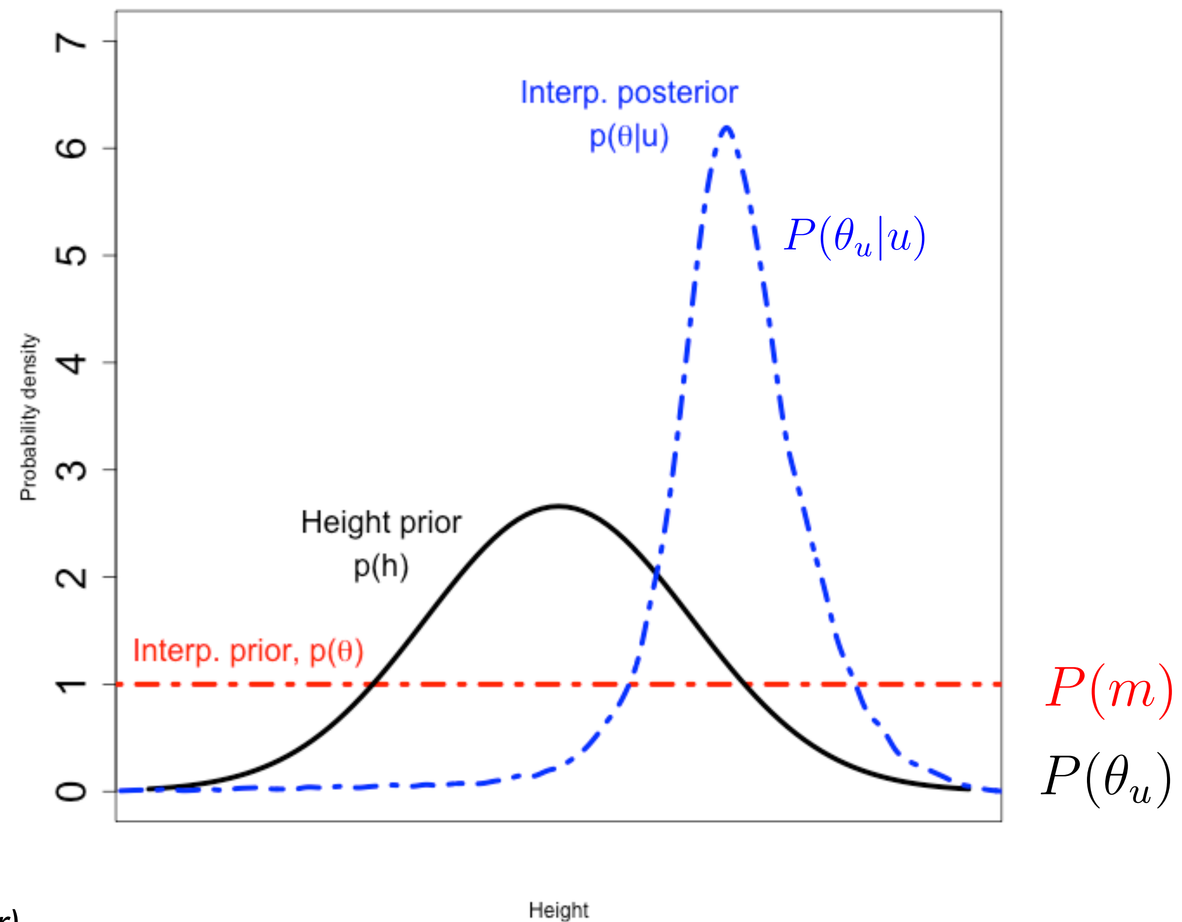
$$L_1(m, \theta_u | u) \propto S_1(u | m, \theta_u) \underline{P(m)} P(\theta_u)$$



(Figures due to Dan Lassiter)

Visualizing pragmatic listener inferences

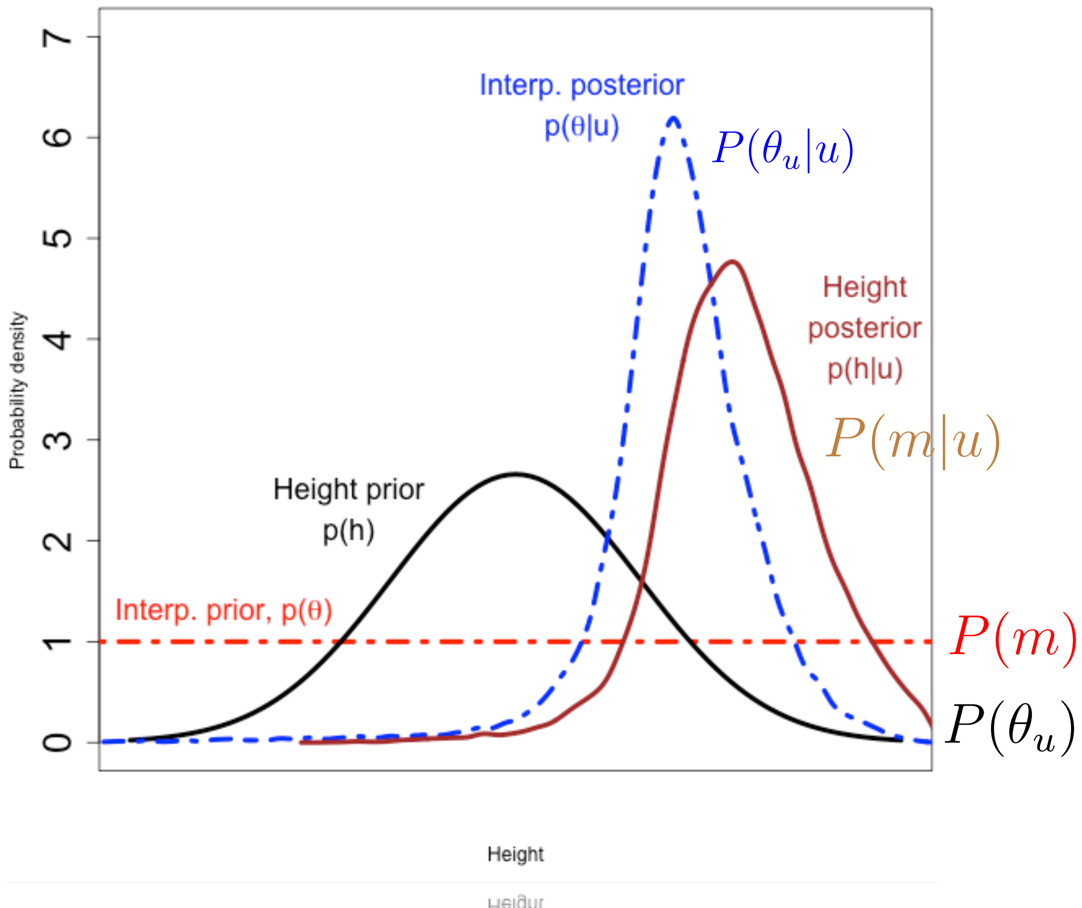
$$L_1(m, \underline{\theta_u} | \underline{u}) \propto S_1(u | m, \theta_u) P(m) P(\theta_u)$$



(Figures due to Dan Lassiter)

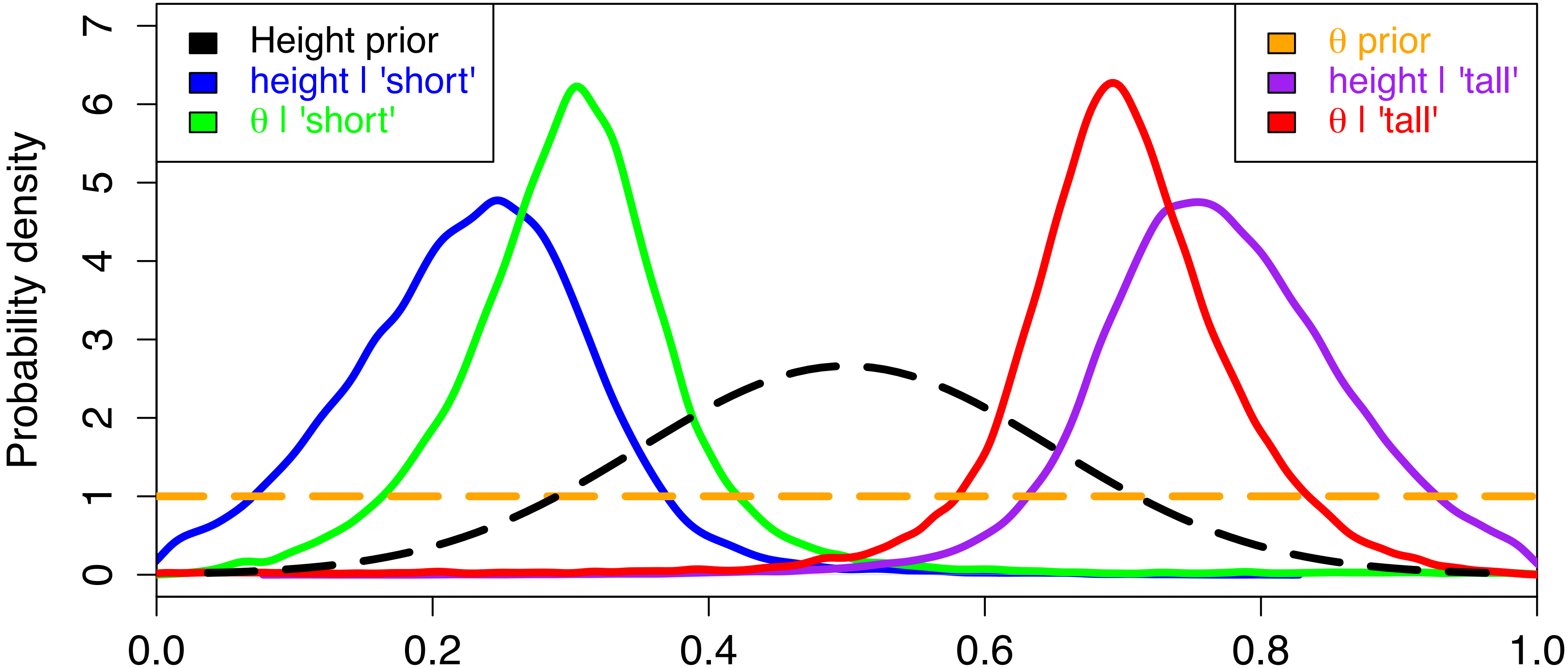
Visualizing pragmatic listener inferences

$$L_1(m, \underline{\theta_u} | \underline{u}) \propto S_1(u|m, \theta_u) P(m) P(\theta_u)$$



(Figures due to Dan Lassiter)

Antonyms



(Due to Dan Lassiter)

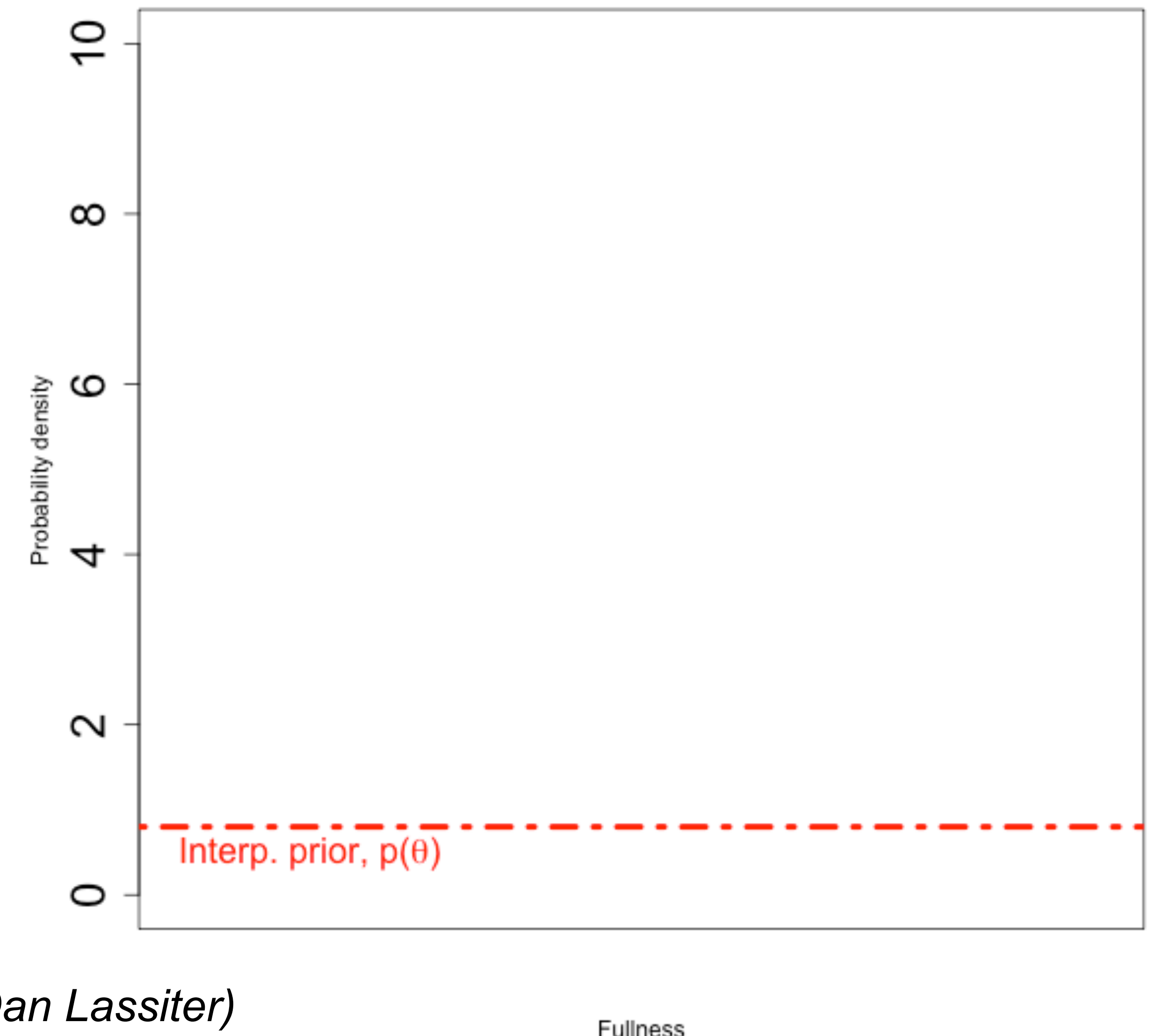
Absolute adjectives

- *full/empty, wet/dry, safe/dangerous, ...*
  - meanings are less (not?) context-dependent
  - meanings are sharp(er)
  - reference classes apparently not relevant to interpretation

(Due to Dan Lassiter)

The pragmatic model on *full*

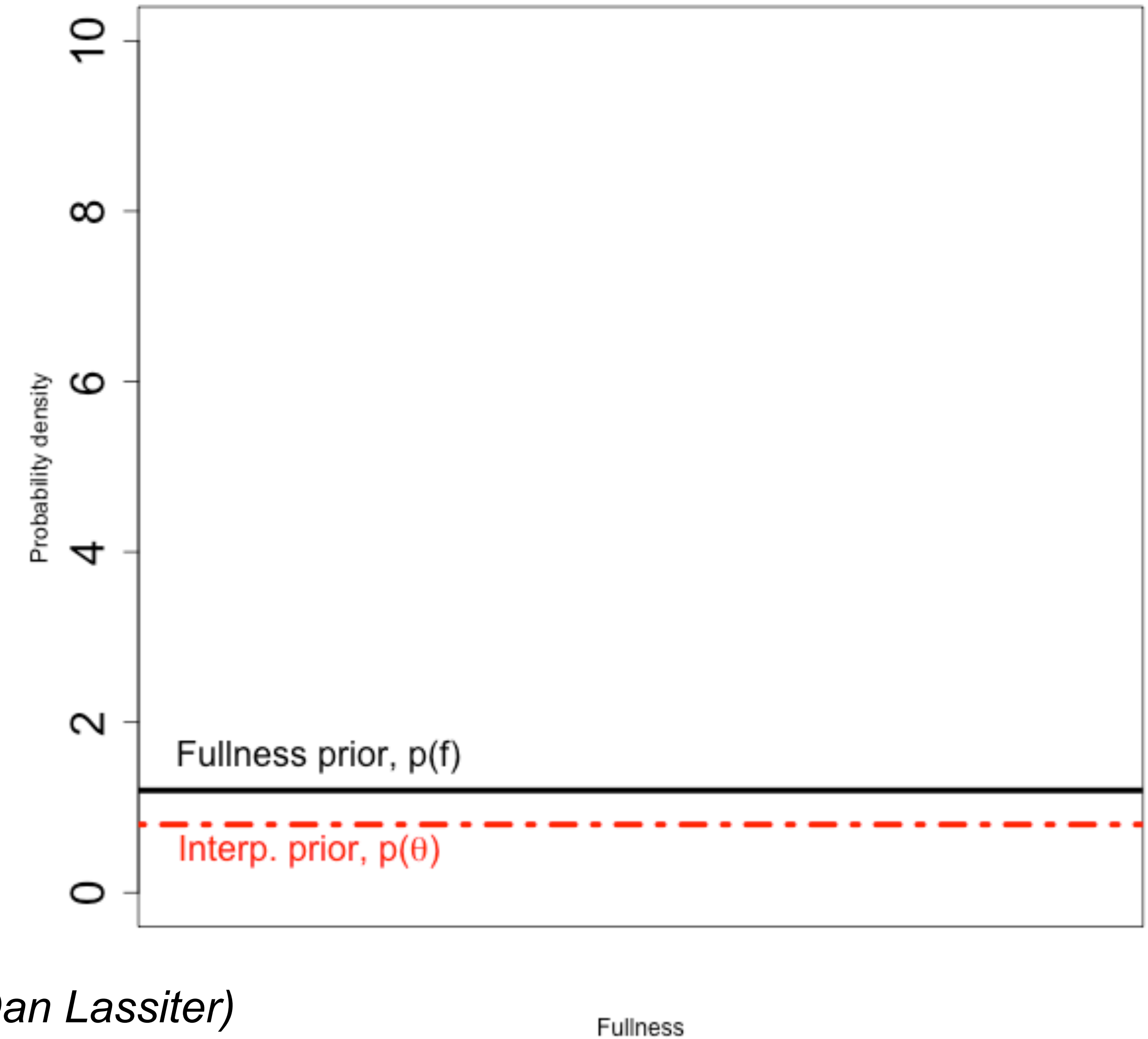
- Crucially, fullness is a ***bounded*** scale!



(Graph due to Dan Lassiter)

The pragmatic model on *full*

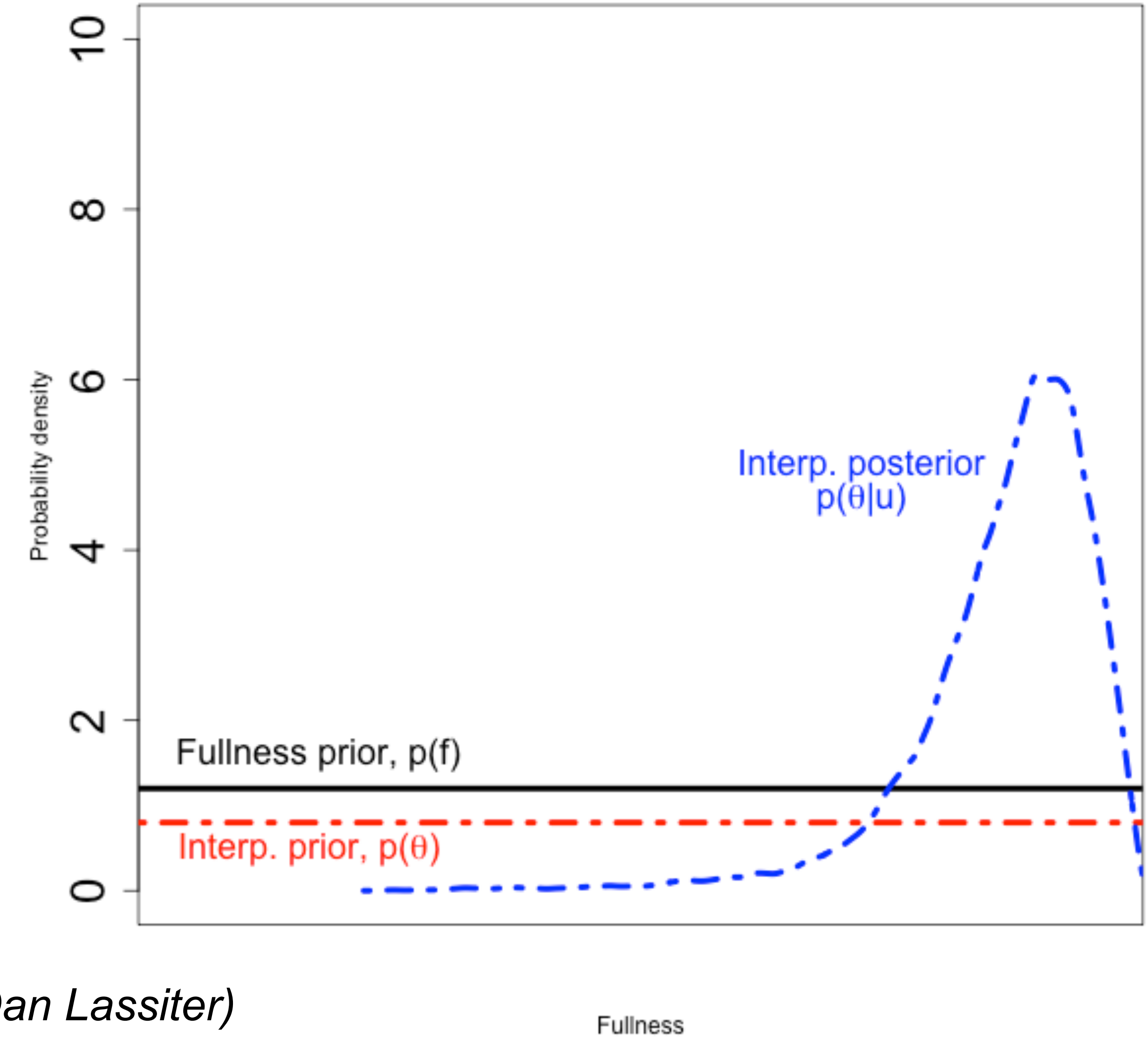
- Crucially, fullness is a ***bounded*** scale!



(Graph due to Dan Lassiter)

The pragmatic model on *full*

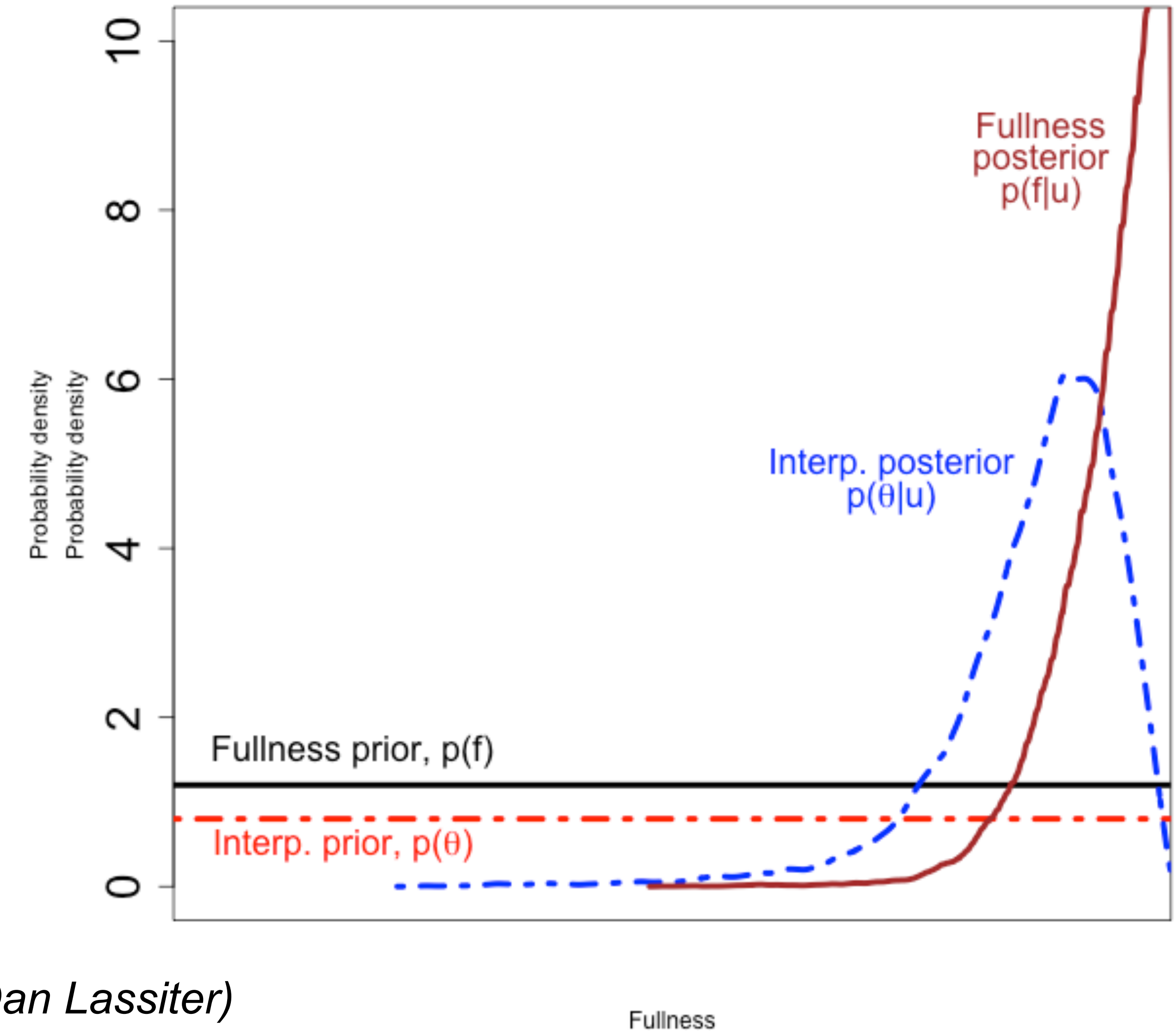
- Crucially, fullness is a ***bounded*** scale!



(Graph due to Dan Lassiter)

The pragmatic model on *full*

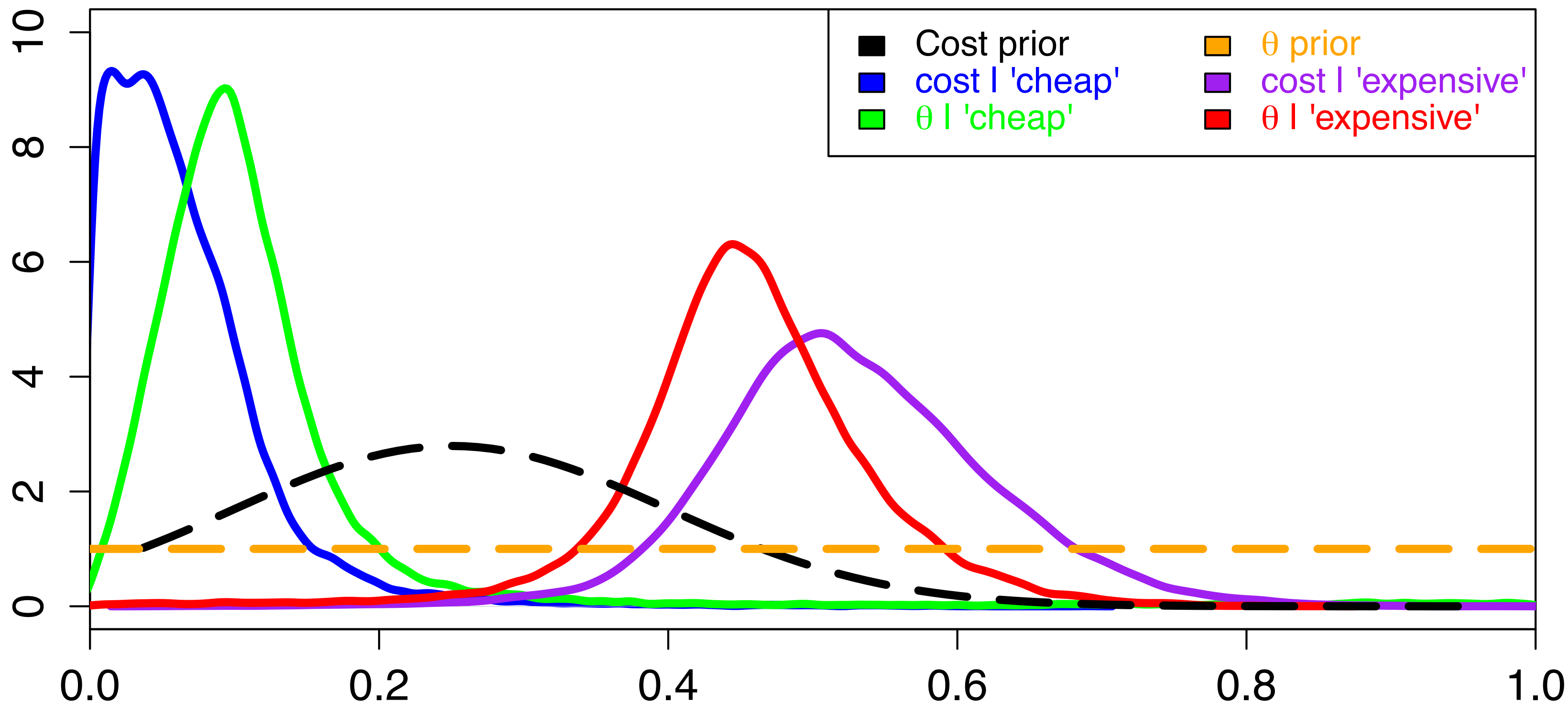
- Crucially, fullness is a ***bounded*** scale!



(Graph due to Dan Lassiter)

Bounds on scales

- On the Lassiter & Goodman model, asymmetries in the interpretations of adjectives arise naturally as a consequence of the prior



(Graph due to Dan Lassiter)

Summary

- Scalar adjectives are a simple example, but pose an additional challenge for pragmatics models
- Some part of the *literal meaning of an utterance* must get contextually determined
- This is one of the simplest examples of interleaving of semantic representation and probabilistic pragmatic inference
- Pieces of the puzzle:
  - Logical semantic representations
  - Latent-variable treatment of pieces of these representations
  - Prior probabilities on likely speaker meanings
  - Joint, utility-driven posterior inference on latent semantic variables and speaker meaning