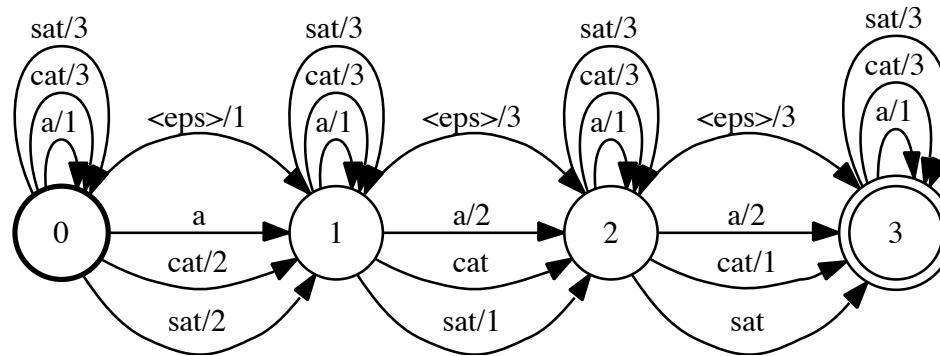


# Noisy-channel sentence comprehension theory



Roger Levy

9.19: Computational Psycholinguistics

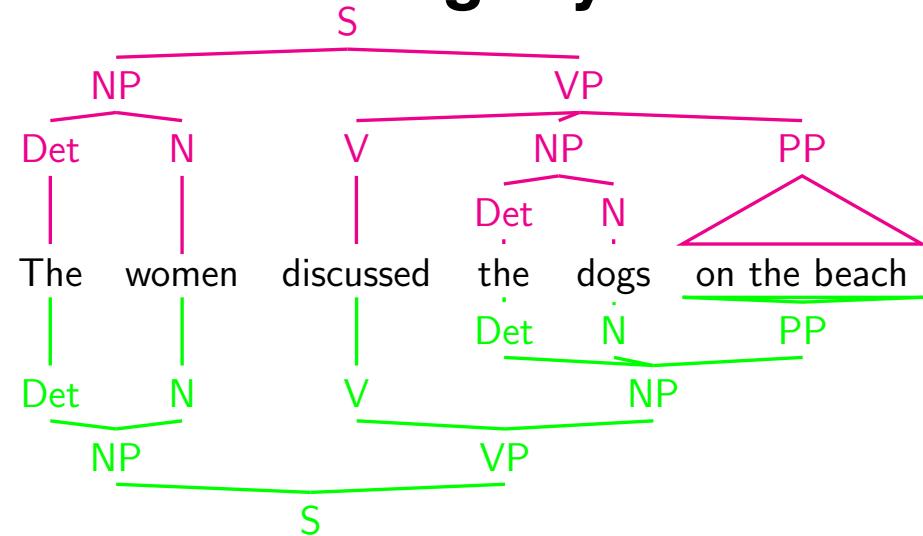
17 November 2021

# Today's agenda

- Review principles of rational analysis and its application to theory of language comprehension
- Examine a phenomenon challenging for surprisal theory
- Propose a noisy-channel processing theory, using information theory and probabilistic grammars
- Develop a hypothesis within the theory for the challenging phenomenon
- Empirically test a key prediction of the theory

# Challenges for efficient linguistic communication

## Ambiguity



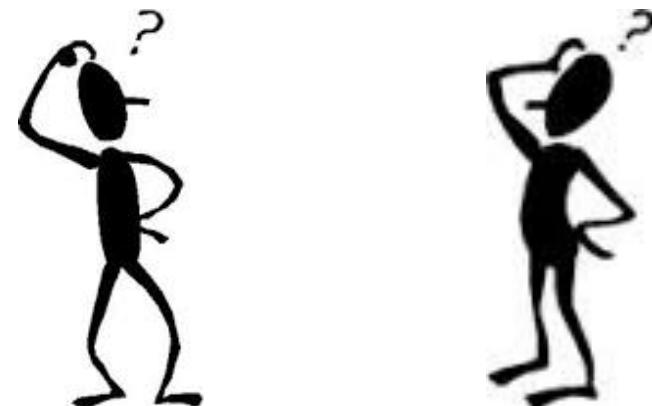
## Environmental noise



## Memory Limitations



## Incomplete knowledge of one's interlocutors



# Rational analysis

---

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations
  4. Derive predicted optimal behavior given 1—3

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations
  4. Derive predicted optimal behavior given 1—3
  5. Compare predictions with empirical data

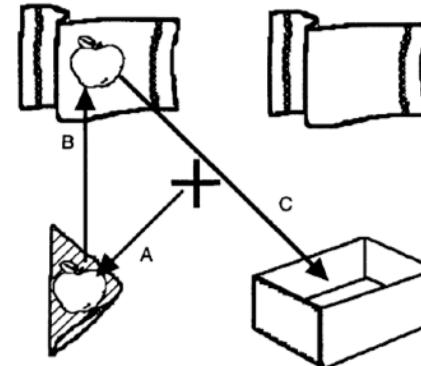
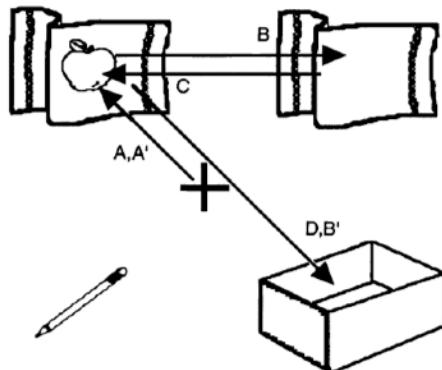
# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations
  4. Derive predicted optimal behavior given 1—3
  5. Compare predictions with empirical data
  6. If necessary, iterate 1—5

# Efficient comprehension as rational, goal-driven

- Online sentence comprehension is hard
- But lots of information sources can be usefully brought to bear to help with the task
- Therefore, it would be *rational* for people to use *all information sources available*, whenever possible
- This is what *incrementality* is
- We have lots of evidence that people do this often
- How do we reconcile these information sources?



*“Put the apple on the towel in the box.” (Tanenhaus et al., 1995, Science)*

# Comprehenders as reverse engineers

---

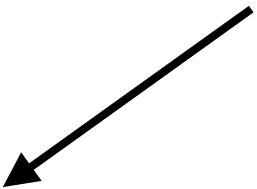
Discourse goals [eat tastier food]

# Comprehenders as reverse engineers

---

Discourse goals [eat tastier food]

Planned  
communicative  
acts



[ask dinner partner  
for spices]

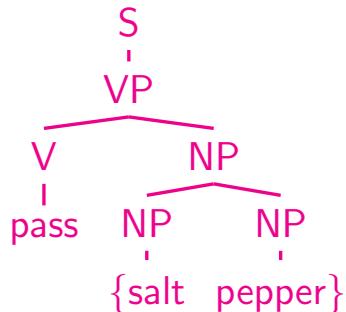
# Comprehenders as reverse engineers

Discourse goals [eat tastier food]

Planned  
communicative  
acts

[ask dinner partner  
for spices]

Lexicalization  
& constituency



# Comprehenders as reverse engineers

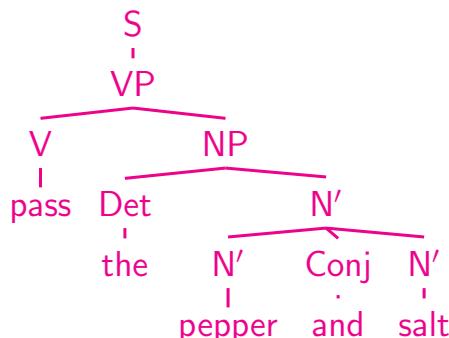
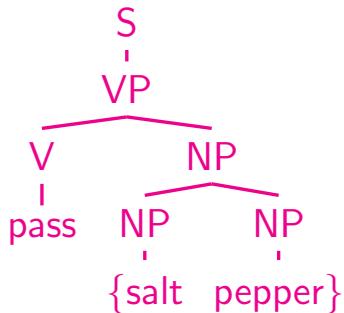
Discourse goals [eat tastier food]

Planned  
communicative  
acts

[ask dinner partner  
for spices]

Lexicalization  
& constituency

Linearization  
decisions

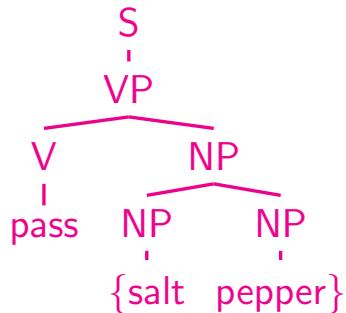


# Comprehenders as reverse engineers

# Planned communicative acts

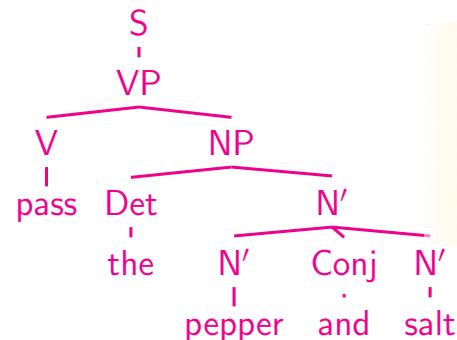
[ask dinner partner  
for spices]

# Lexicalization & constituency

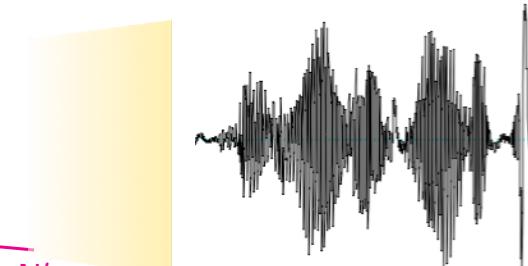


# Discourse goals [eat tastier food]

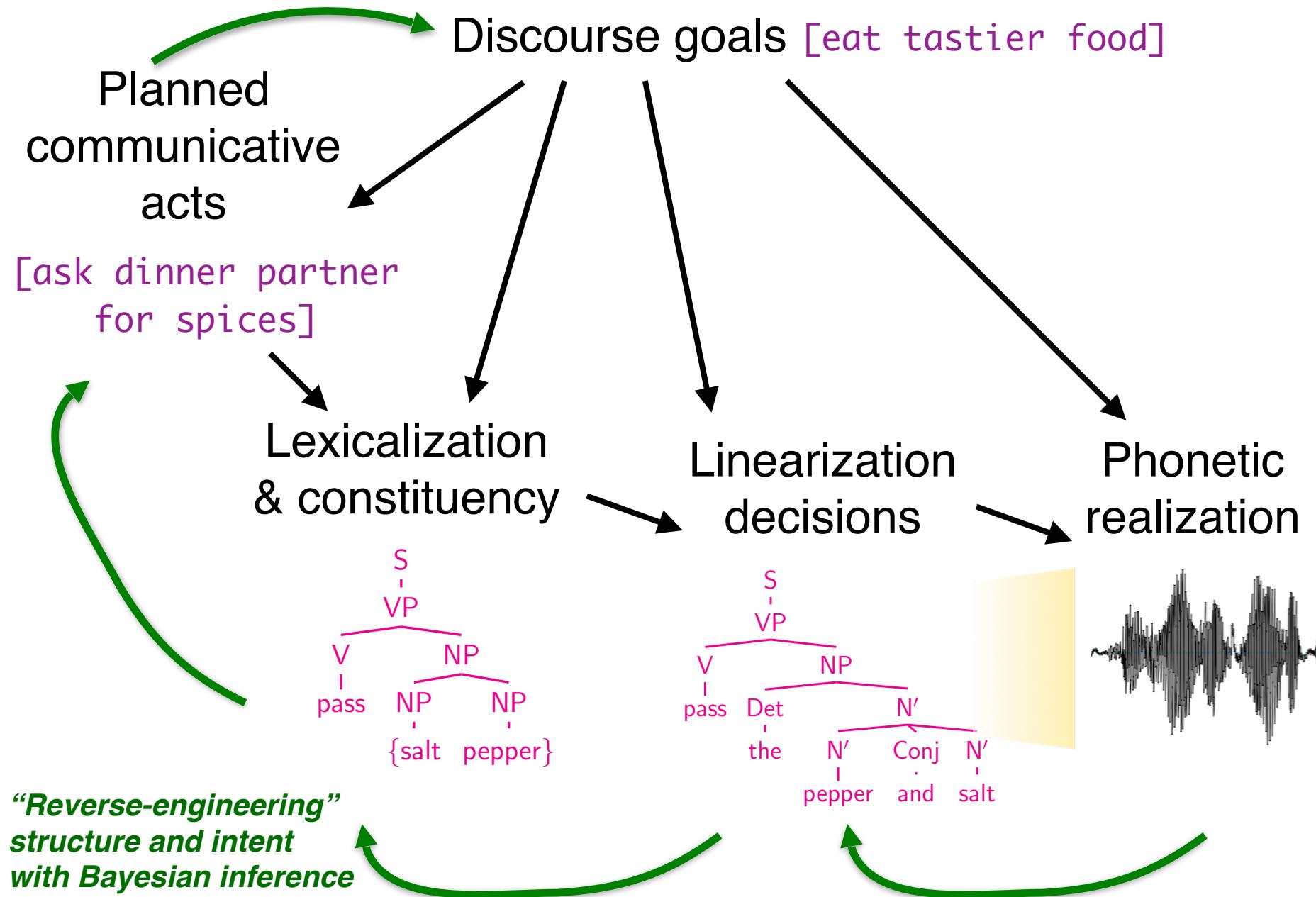
# Linearization decisions



## Phonetic realization



# Comprehenders as reverse engineers



# Surprisal summary: psycholinguistic evidence

# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

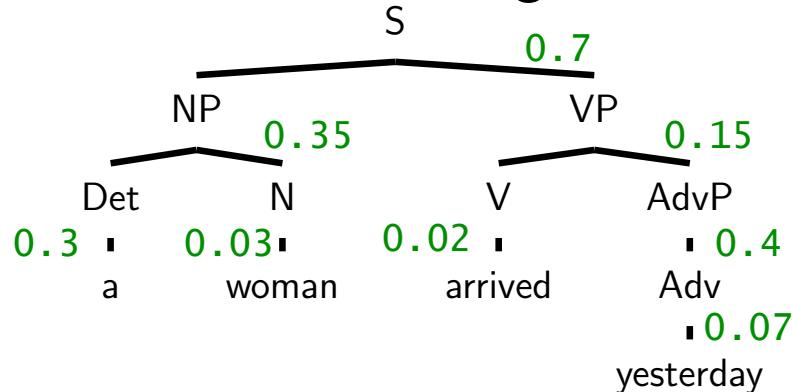
# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

- Probabilistic grammar



$$\begin{aligned} P(T) &= 0.7 * 0.35 * 0.15 * 0.3 * 0.03 * 0.02 * 0.4 * 0.07 \\ &= 1.85 \cdot 10^{-7} \end{aligned}$$

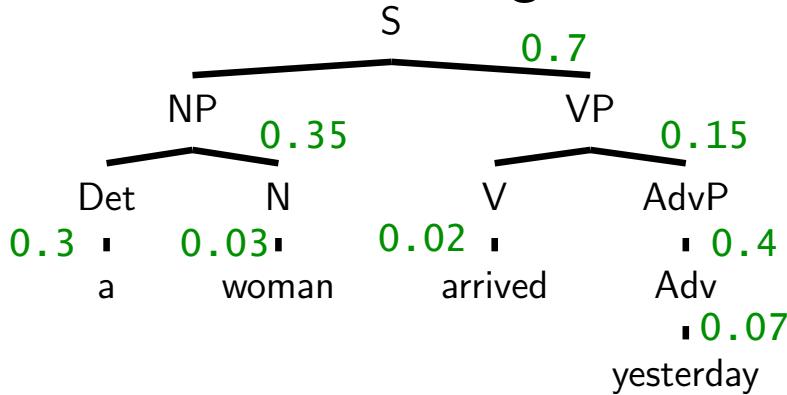
# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

- Probabilistic grammar



$$\begin{aligned} P(T) &= 0.7 * 0.35 * 0.15 * 0.3 * 0.03 * 0.02 * 0.4 * 0.07 \\ &= 1.85 \cdot 10^{-7} \end{aligned}$$

- Surprisal

$$\begin{aligned} \text{Surprisal}(w_i) &\equiv \log \frac{1}{P(w_i|\text{CONTEXT})} \\ &\approx \log \frac{1}{P(w_i|w_1 \dots w_{i-1})} \end{aligned}$$

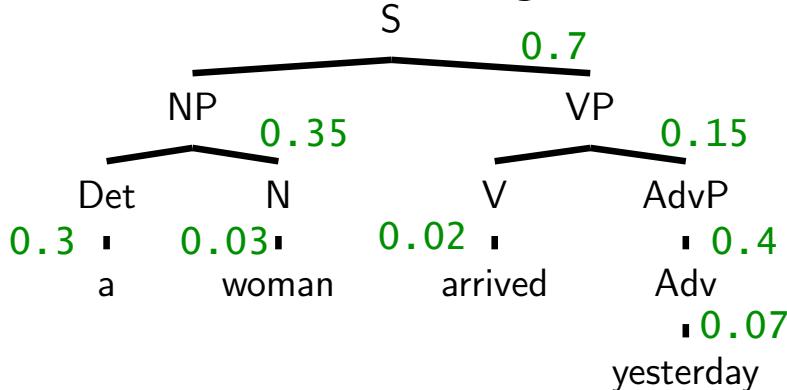
# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

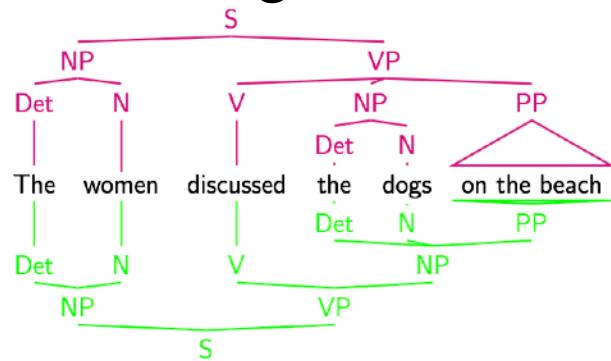
- Probabilistic grammar



- Surprisal

$$\begin{aligned}\text{Surprisal}(w_i) &\equiv \log \frac{1}{P(w_i|\text{CONTEXT})} \\ &\approx \log \frac{1}{P(w_i|w_1 \dots w_{i-1})}\end{aligned}$$

- Global disambiguation



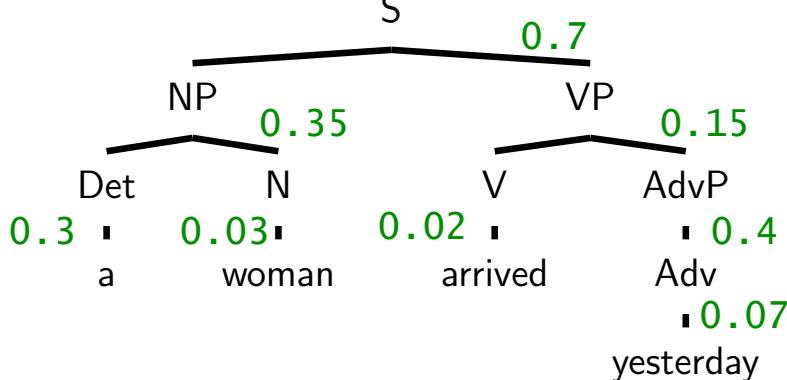
# Surprisal summary: psycholinguistic evidence

# Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

- Probabilistic grammar



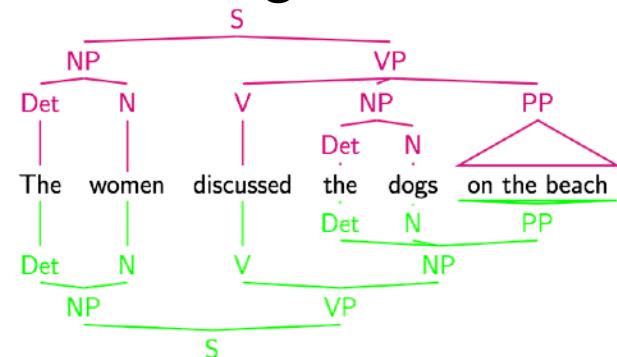
$$P(T) = 0.7 \cdot 0.35 \cdot 0.15 \cdot 0.3 \cdot 0.03 \cdot 0.02 \cdot 0.4 \cdot 0.07 \\ = 1.85 \cdot 10^{-7}$$

- Surprisal

$$\text{Surprisal}(w_i) \equiv \log \frac{1}{P(w_i|\text{CONTEXT})}$$

$$[\approx \log \frac{1}{P(w_i|w_1 \dots w_{i-1})}]$$

- ## • Global disambiguation



- Garden-pathing

*When the dog scratched the vet removed the muzzle.*

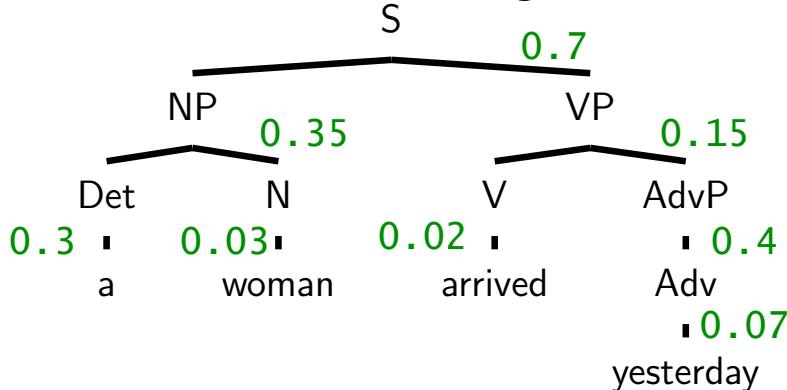
# Surprisal summary: psycholinguistic evidence

Problems addressed by a theory consisting of:

- Bayesian inference

$$P(\text{Str}|\text{Input}) \propto P(\text{Input}|\text{Str})P(\text{Str})$$

- Probabilistic grammar

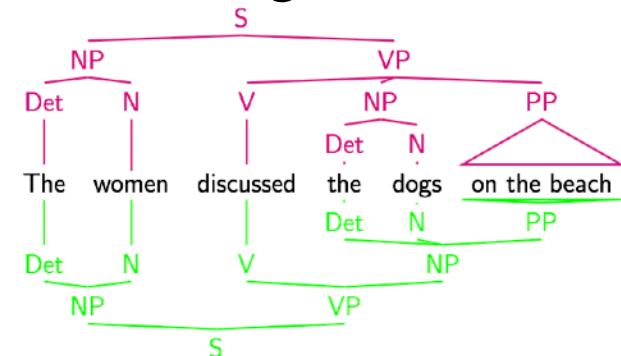


$$P(T) = 0.7 * 0.35 * 0.15 * 0.3 * 0.03 * 0.02 * 0.4 * 0.07 \\ = 1.85 \cdot 10^{-7}$$

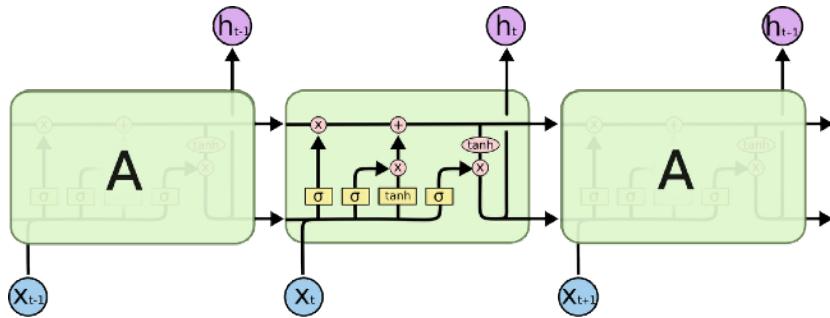
- Surprisal

$$\text{Surprisal}(w_i) \equiv \log \frac{1}{P(w_i|\text{CONTEXT})} \\ \left[ \approx \log \frac{1}{P(w_i|w_1 \dots w_{i-1})} \right]$$

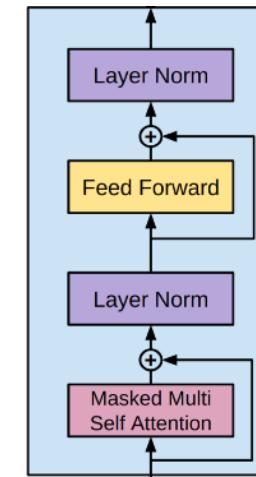
- Global disambiguation



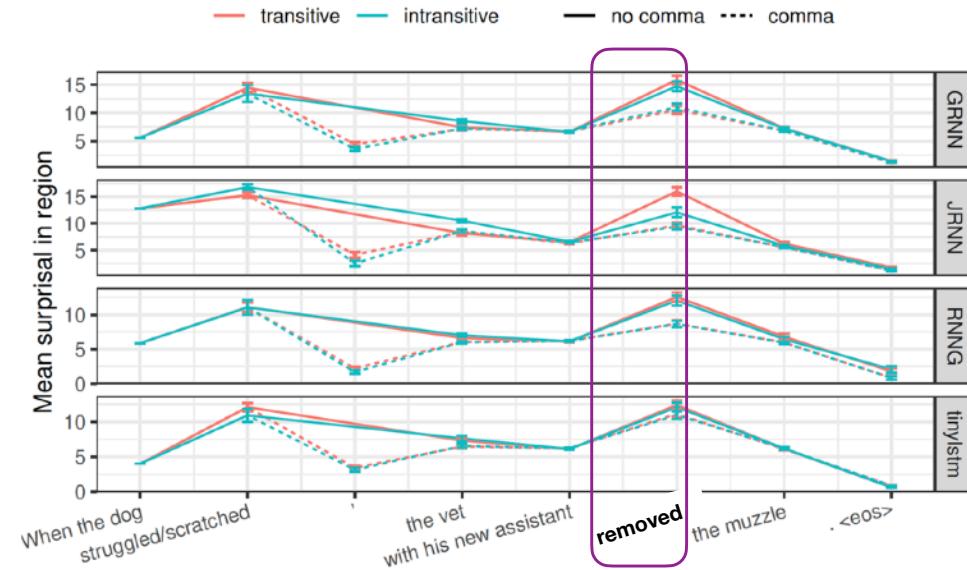
# Syntax-like surprisal from deep-learning models



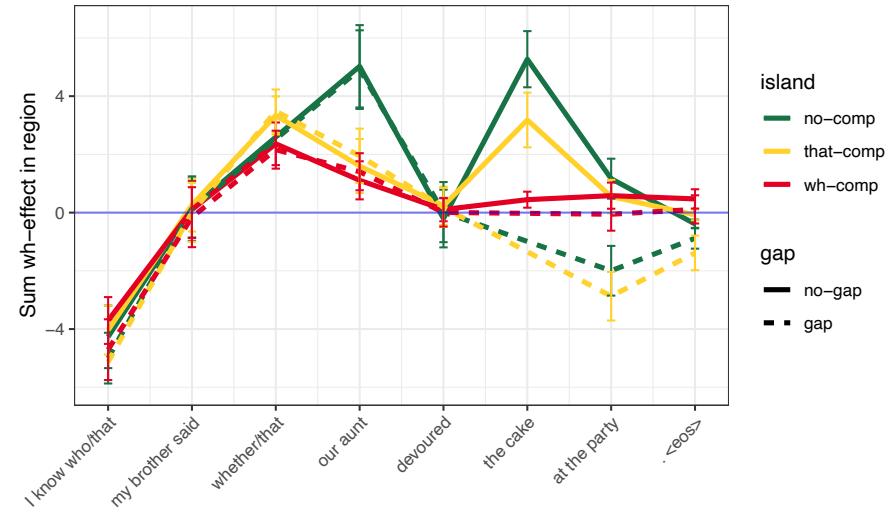
(Elman, 1990; Hochreiter & Schmidhuber, 1997)



(Vaswani et al., 2017; Radford et al., 2018, 2019)



(Futrell et al. 2019, NAACL)



(Wilcox et al., 2018, BlackBox NLP)

# An incremental inference puzzle for surprisal

---

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:  
(a) *The coach smiled at the player tossed the frisbee.*

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

  - (b) *The coach smiled at the player **thrown** the frisbee.*

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

  - (b) *The coach smiled at the player **thrown** the frisbee.*
  - (c) *The coach smiled at the player **who was thrown** the frisbee.*

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

  - (b) *The coach smiled at the player **thrown** the frisbee.*
  - (c) *The coach smiled at the player **who was thrown** the frisbee.*
  - (d) *The coach smiled at the player **who was tossed** the frisbee.*

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

  - (b) *The coach smiled at the player **thrown** the frisbee.*
  - (c) *The coach smiled at the player **who was thrown** the frisbee.*
  - (d) *The coach smiled at the player **who was tossed** the frisbee.*
- Readers boggle at “tossed” in (a), but not in (b-d)

# An incremental inference puzzle for surprisal

---

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

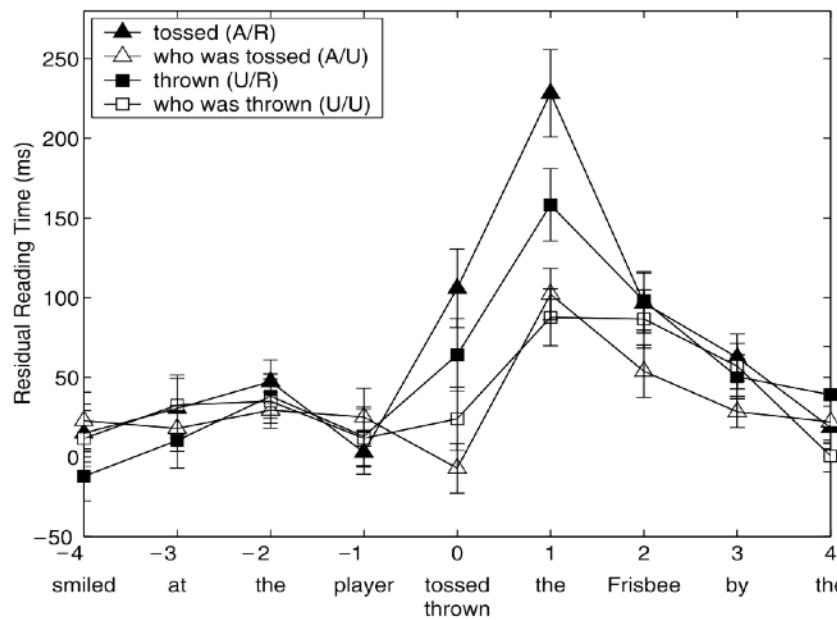
  - (b) *The coach smiled at the player **thrown** the frisbee.*
  - (c) *The coach smiled at the player **who was thrown** the frisbee.*
  - (d) *The coach smiled at the player **who was tossed** the frisbee.*
- Readers boggle at “tossed” in (a), but not in (b-d)

# An incremental inference puzzle for surprisal

- Try to understand this sentence:
  - The coach smiled at the player tossed the frisbee.*

...and contrast this with:

- The coach smiled at the player thrown the frisbee.*
  - The coach smiled at the player who was thrown the frisbee.*
  - The coach smiled at the player who was tossed the frisbee.*
- Readers boggle at “tossed” in (a), but not in (b-d)



# An incremental inference puzzle for surprisal

- Try to understand this sentence:

(a) *The coach smiled at the player tossed the frisbee.*

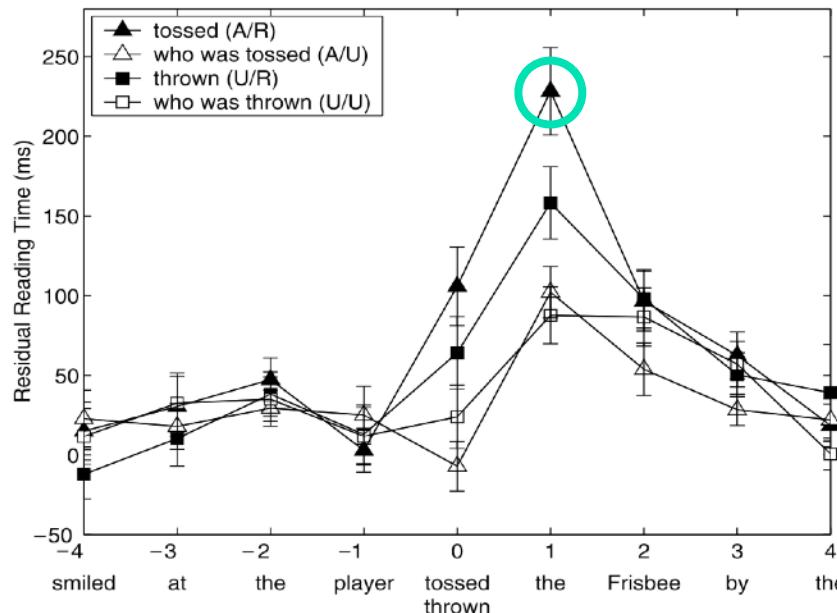
...and contrast this with:

(b) *The coach smiled at the player thrown the frisbee.*

(c) *The coach smiled at the player who was thrown the frisbee.*

(d) *The coach smiled at the player who was tossed the frisbee.*

- Readers boggle at “tossed” in (a), but not in (b-d)

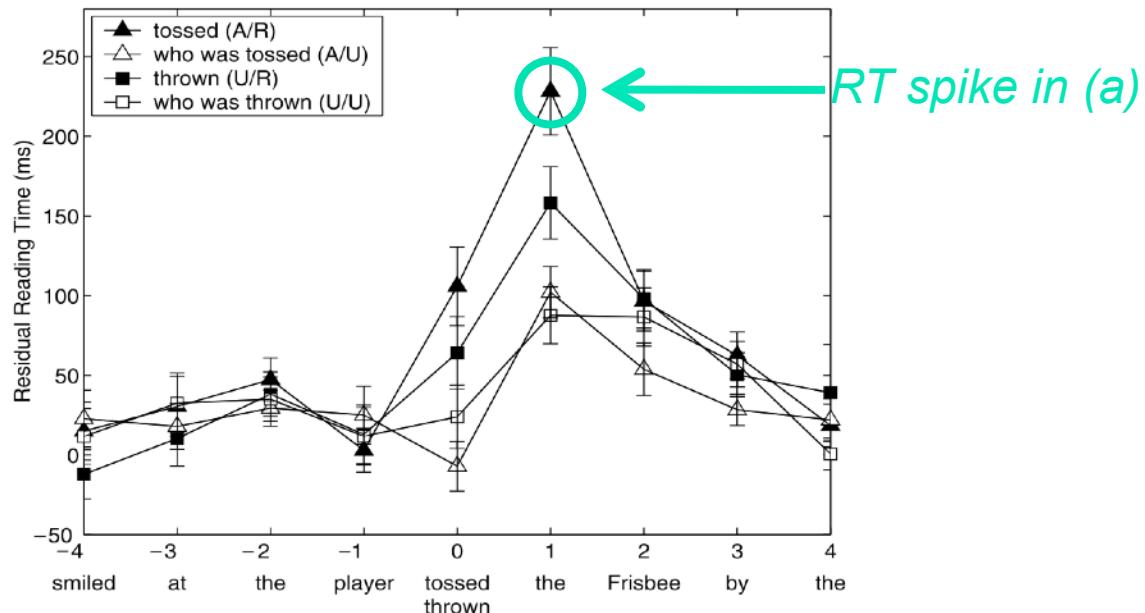


# An incremental inference puzzle for surprisal

- Try to understand this sentence:
  - (a) *The coach smiled at the player tossed the frisbee.*

...and contrast this with:

  - (b) *The coach smiled at the player thrown the frisbee.*
  - (c) *The coach smiled at the player who was thrown the frisbee.*
  - (d) *The coach smiled at the player who was tossed the frisbee.*
- Readers boggle at “tossed” in (a), but not in (b-d)



# Why is *tossed/thrown* interesting?

# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - *The woman brought the sandwich...tripped*

# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - The woman brought the sandwich...tripped*  
*verb?*  
*participle?*

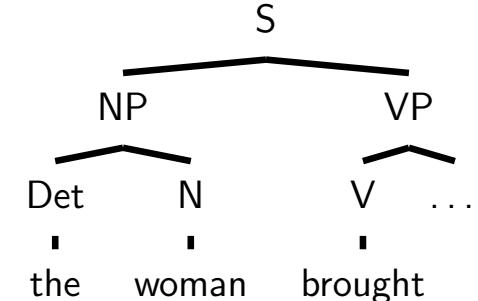
# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

- The woman brought the sandwich...tripped*

verb?  
participle?



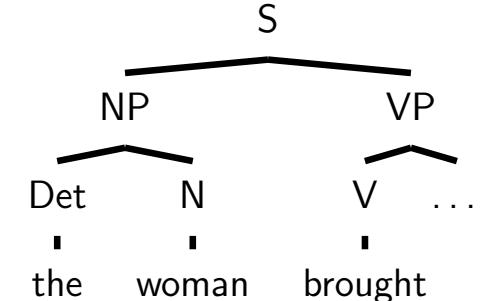
# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:

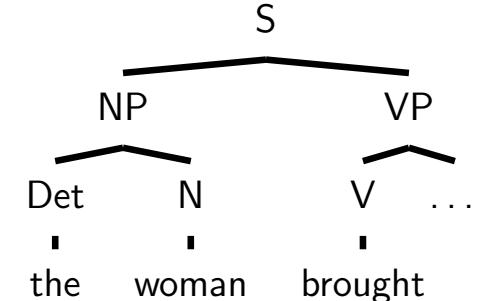
# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:
  - The coach smiled at the player tossed...*

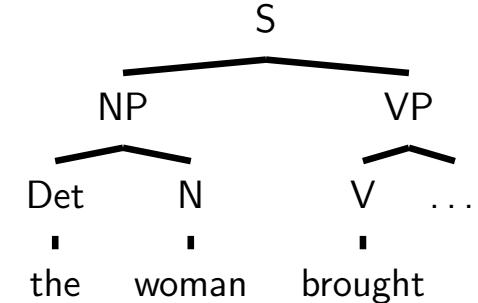
# Why is *tossed/thrown* interesting?

---

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:

- The coach smiled at the player tossed...*

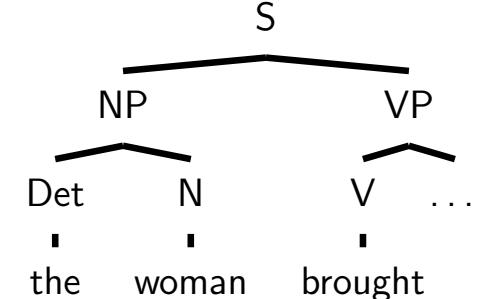
verb?  
participle?

# Why is *tossed/thrown* interesting?

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

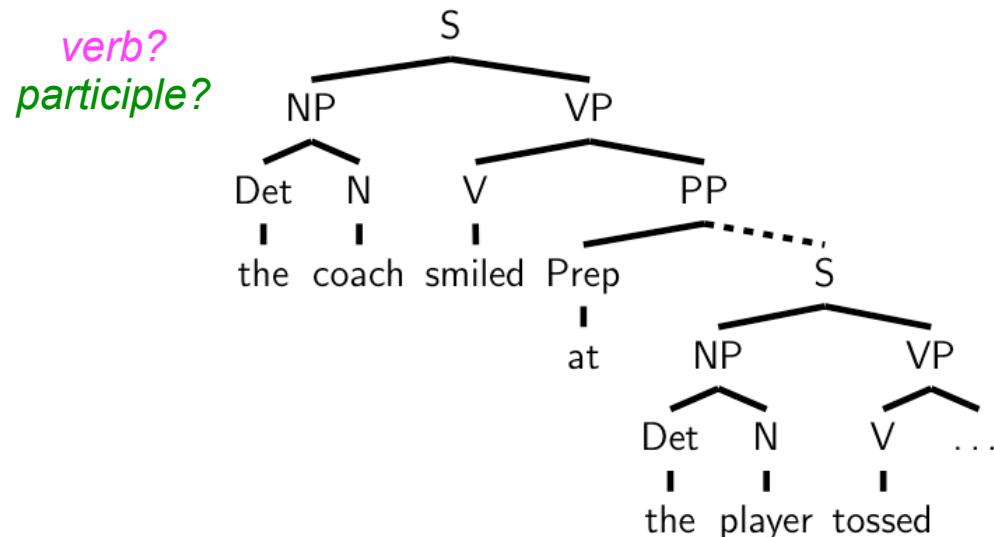
- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:

- The coach smiled at the player tossed...*

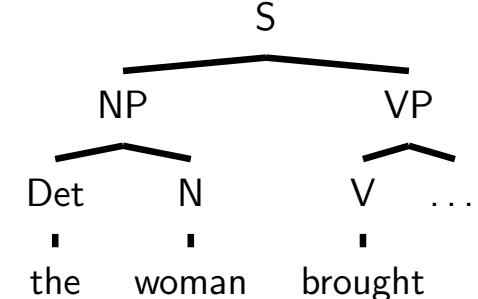


# Why is *tossed/thrown* interesting?

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

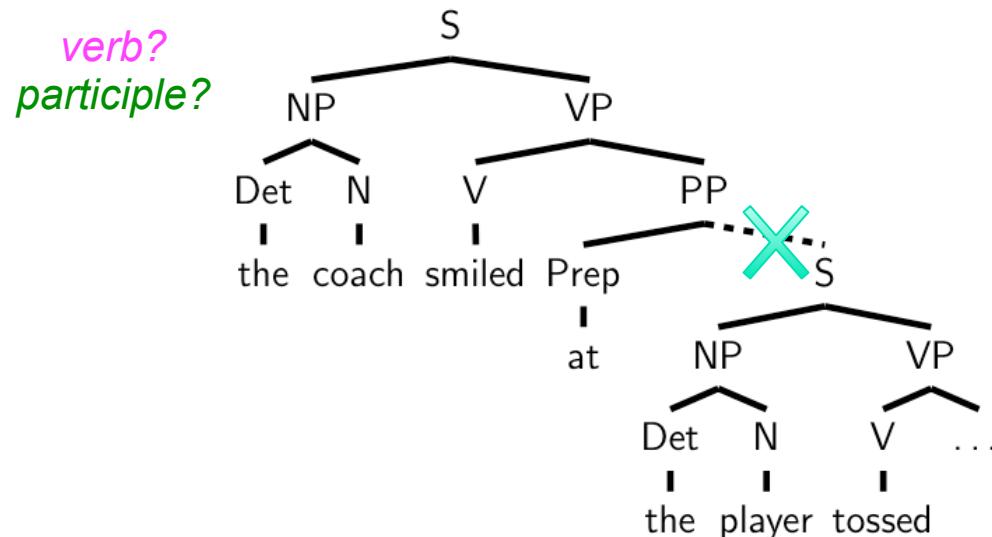
- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:

- The coach smiled at the player tossed...*

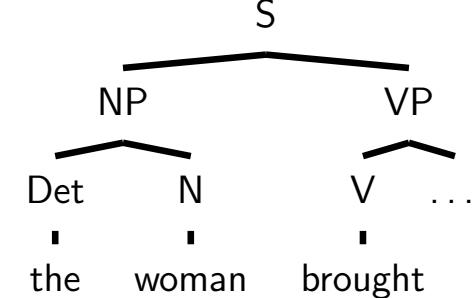


# Why is *tossed/thrown* interesting?

- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation

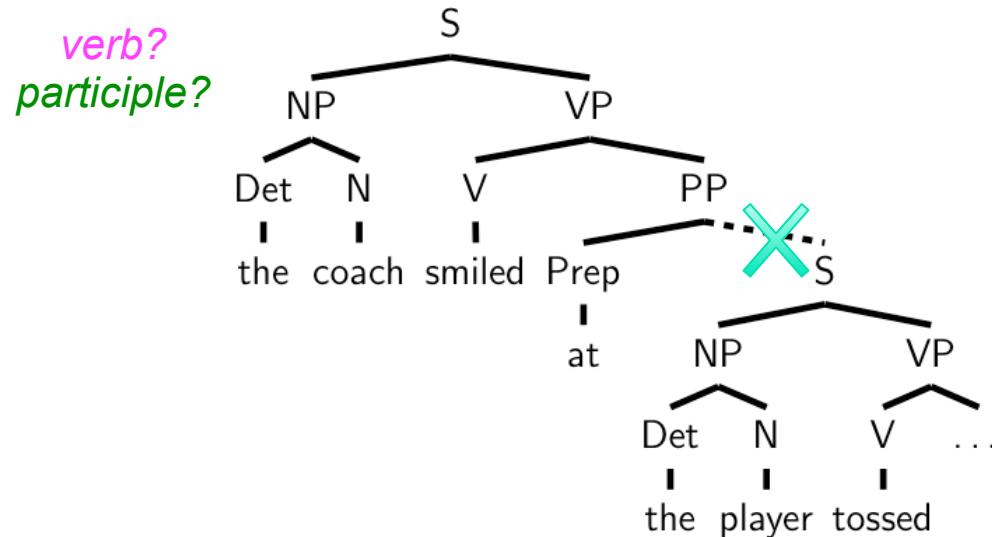
- The woman brought the sandwich...tripped*

verb?  
participle?



- But now context “should” rule out the garden path:

- The coach smiled at the player tossed...*



- A challenge for rational models: **failure to condition on relevant context**

# Uncertain input in language comprehension

---

# Uncertain input in language comprehension

---

- Previous state of the art models for ambiguity resolution ≈ probabilistic incremental parsing

# Uncertain input in language comprehension

---

- Previous state of the art models for ambiguity resolution ≈ probabilistic incremental parsing
- Simplifying assumption:
  - Input is *clean* and *perfectly-formed*
  - No uncertainty about input is admitted

# Uncertain input in language comprehension

---

- Previous state of the art models for ambiguity resolution ≈ probabilistic incremental parsing
- Simplifying assumption:
  - Input is *clean* and *perfectly-formed*
  - No uncertainty about input is admitted
- Intuitively seems patently wrong...
  - We sometimes *misread* things
  - We can also *proofread*

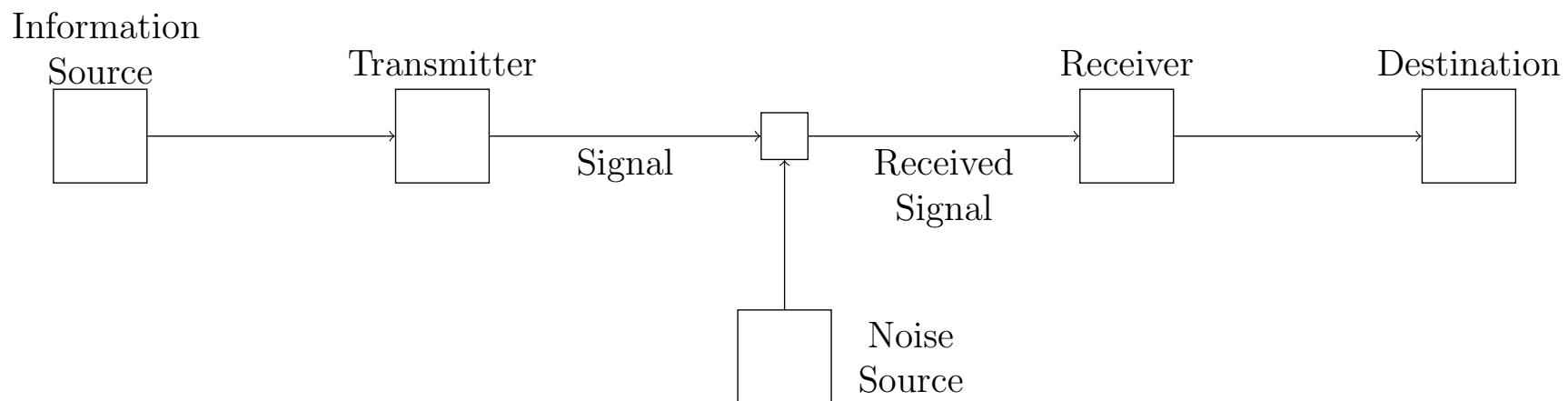
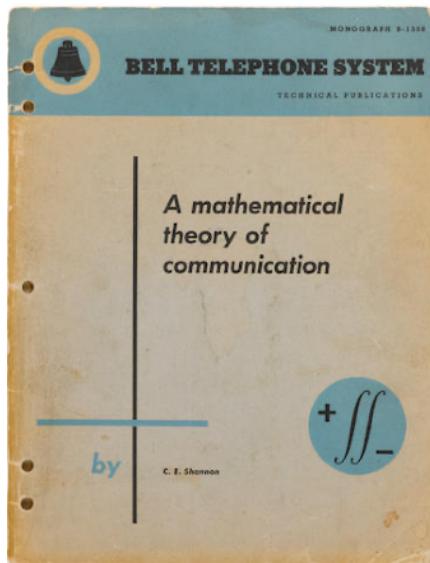
# Uncertain input in language comprehension

---

- Previous state of the art models for ambiguity resolution ≈ probabilistic incremental parsing
- Simplifying assumption:
  - Input is *clean* and *perfectly-formed*
  - No uncertainty about input is admitted
- Intuitively seems patently wrong...
  - We sometimes *misread* things
  - We can also *proofread*
- Leads to two questions:
  1. What might a model of sentence comprehension under uncertain input look like?
  2. What interesting consequences might such a model have?

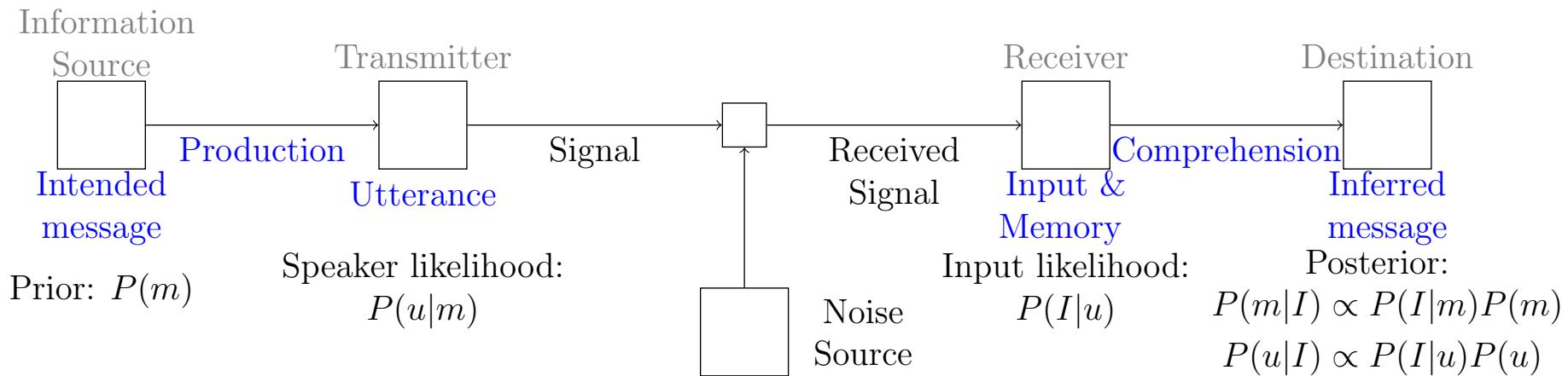
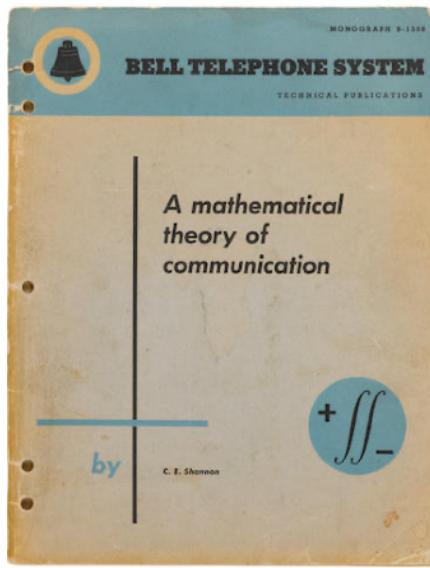
# Noisy-channel theory of language processing

(Shannon, 1948)



# Noisy-channel theory of language processing

(Shannon, 1948)



# Noisy-channel sentence processing

---

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

# Noisy-channel sentence processing

---

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

# Noisy-channel sentence processing

---

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

- If we don't observe a sentence but only a noisy input  $I$ :

$$P_G(T|I) \propto \sum_{\mathbf{w}} P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

# Noisy-channel sentence processing

---

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

- If we don't observe a sentence but only a noisy input  $I$ :

$$P_G(T|I) \propto \sum_{\mathbf{w}} P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

- If we know true sentence  $\mathbf{w}^*$  but not input  $I$ :

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I P_C(\mathbf{w}|I, \mathbf{w}^*)P_T(I|\mathbf{w}^*) dI$$

# Noisy-channel sentence processing

---

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

- If we don't observe a sentence but only a noisy input  $I$ :

$$P_G(T|I) \propto \sum_{\mathbf{w}} P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

- If we know true sentence  $\mathbf{w}^*$  but not input  $I$ :

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I P_C(\mathbf{w}|I, \mathbf{w}^*)P_T(I|\mathbf{w}^*) dI$$

*comprehender's  
model*

# Noisy-channel sentence processing

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

- If we don't observe a sentence but only a noisy input  $I$ :

$$P_G(T|I) \propto \sum_{\mathbf{w}} P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

- If we know true sentence  $\mathbf{w}^*$  but not input  $I$ :

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I P_C(\mathbf{w}|I, \mathbf{w}^*) P_T(I|\mathbf{w}^*) dI$$

*comprehender's model* true model

# Noisy-channel sentence processing

- Standard probabilistic sentence processing:

$$P_G(T|\mathbf{w}) \propto P(\mathbf{w}|T)P(T) = P(T, \mathbf{w})$$

- If we don't observe a sentence but only a noisy input  $I$ :

$$P_G(T|I) \propto \sum_{\mathbf{w}} P(I|T, \mathbf{w}) P(\mathbf{w}|T) P(T)$$

$$P_G(\mathbf{w}|I) \propto \sum_T P(I|T, \mathbf{w})P(\mathbf{w}|T)P(T)$$

- If we know true sentence  $w^*$  but not input  $l$ :  *true model*

$$P(\mathbf{w}|\mathbf{w}^*) = \int_I P_C(\mathbf{w}|I, \mathbf{w}^*) P_T(I|\mathbf{w}^*) dI$$

# *comprehender's model*

$$= P_C(\mathbf{w}) \int_I \frac{P_C(I|\mathbf{w})P_T(I|\mathbf{w}^*)}{P_C(I)} dI$$

$$\propto Q(\mathbf{w}, \mathbf{w}^*)$$

Levy (2008, EMNLP)

# Representing noisy input

---

# Representing noisy input

---

- How can we represent the type of noisy input generated by a word sequence?

# Representing noisy input

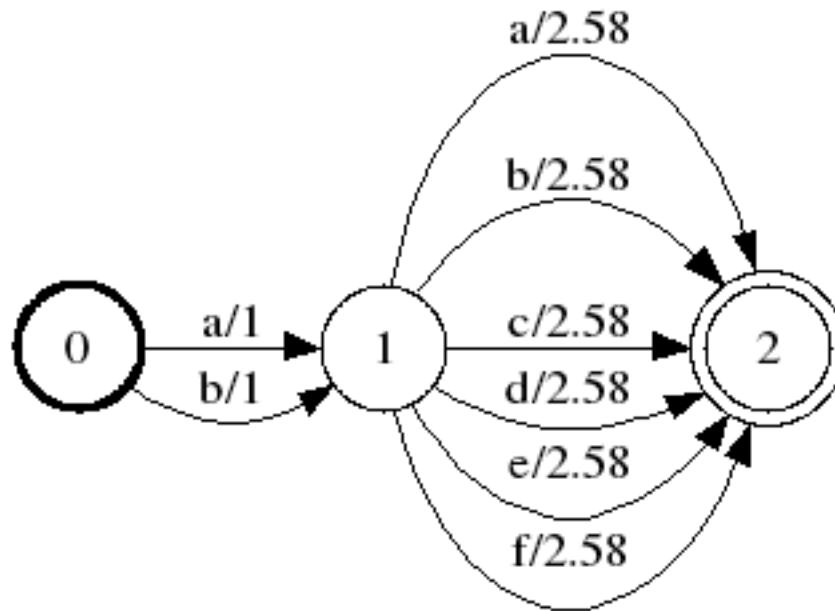
---

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

# Representing noisy input

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

$\text{vocab} = a, b, c, d, e, f$

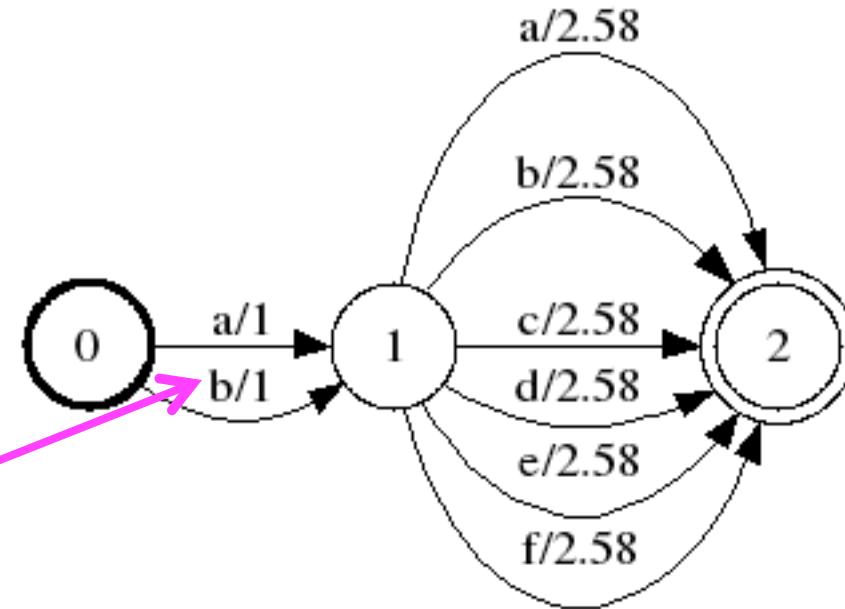


# Representing noisy input

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

$\text{vocab} = a, b, c, d, e, f$

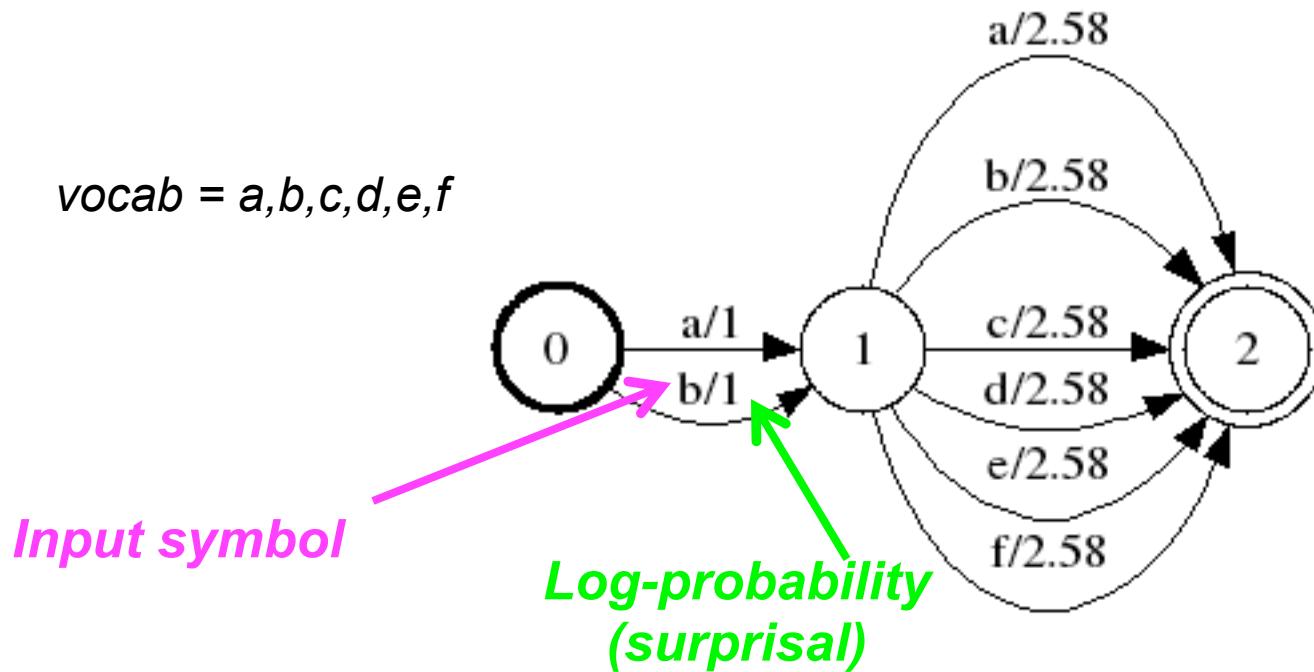
*Input symbol*



# Representing noisy input

- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

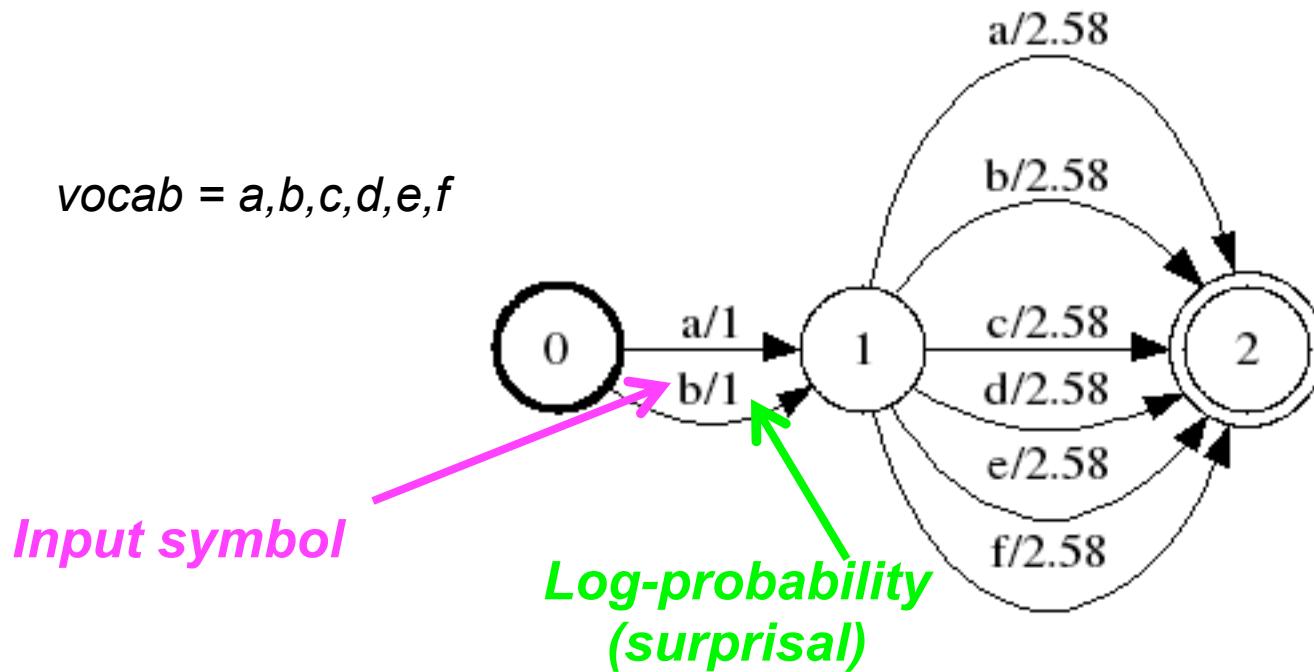
$\text{vocab} = a, b, c, d, e, f$



# Representing noisy input

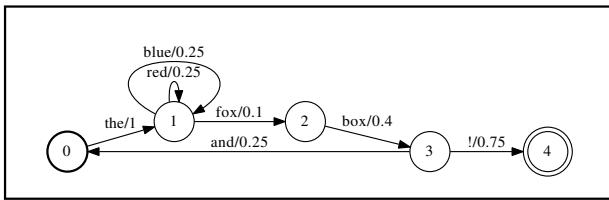
- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) are a good model

$\text{vocab} = a, b, c, d, e, f$



- “Word 1 is a or b, and I have no info about Word 2”

# Weighted finite-state automata



A WEIGHTED FINITE-STATE AUTOMATON (WFSA) consists of a tuple  $(Q, V, S, R)$  such that:

- ▶  $Q$  is a finite set of STATES  $q_0 q_1 \dots q_N$ , with  $q_0$  the designated START STATE;
- ▶  $\Sigma$  is a finite set of terminal symbols;
- ▶  $F \subseteq Q$  is the set of FINAL STATES;
- ▶  $\Delta$  is a finite set of TRANSITIONS each of the form  $q \xrightarrow{i} q'$ , meaning that “if you are in state  $q$  and see symbol  $i$  you can consume it and move to state  $q'$ ”;
- ▶  $\lambda$  is a function mapping transitions to real numbers (weights);
- ▶  $\rho$  is a function mapping final states to real numbers (weights).

# Weighted finite-state automata (2)

- ▶  $Q$  is a finite set of STATES  $q_0 q_1 \dots q_N$ , with  $q_0$  the designated START STATE;
- ▶  $\Sigma$  is a finite set of terminal symbols;
- ▶  $F \subseteq Q$  is the set of FINAL STATES;
- ▶  $\Delta$  is a finite set of TRANSITIONS each of the form  $q \xrightarrow{i} q'$ , meaning that “if you are in state  $q$  and see symbol  $i$  you can consume it and move to state  $q'$ ”;
- ▶  $\lambda$  is a function mapping transitions to real numbers (weights);
- ▶  $\rho$  is a function mapping final states to real numbers (weights).

- ▶  $w_{1\dots N} \in \Sigma^N$  is ACCEPTED or RECOGNIZED by an automaton iff there is a PATH of transitions  $\xrightarrow{1\dots N}$  to a final state  $q^* \in F$  such that

$$q_0 \xrightarrow[1]{w_1} \xrightarrow[2]{w_2} \dots \xrightarrow[N-1]{w_{N-1}} \xrightarrow[N]{w_N} q^*$$

- ▶ The WEIGHT of such a path  $\xrightarrow{1\dots N}$  is the product of the weights of each of the transitions, together with the weight of the final state:

$$P(q_0 \xrightarrow[1]{w_1} \xrightarrow[2]{w_2} \dots \xrightarrow[N-1]{w_{N-1}} \xrightarrow[N]{w_N} q^*) = \rho(q^*) \prod_{i=1}^N \lambda(\xrightarrow{i}) \quad (1)$$

# Probabilistic Linguistic Knowledge

---

# Probabilistic Linguistic Knowledge

---

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)

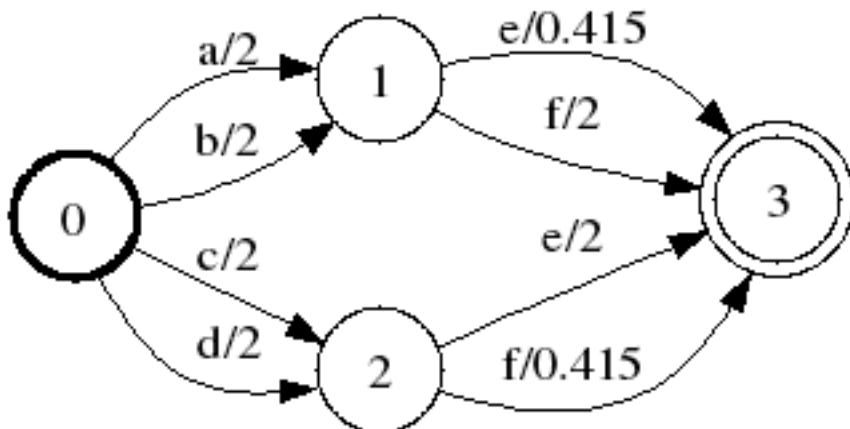
# Probabilistic Linguistic Knowledge

---

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)

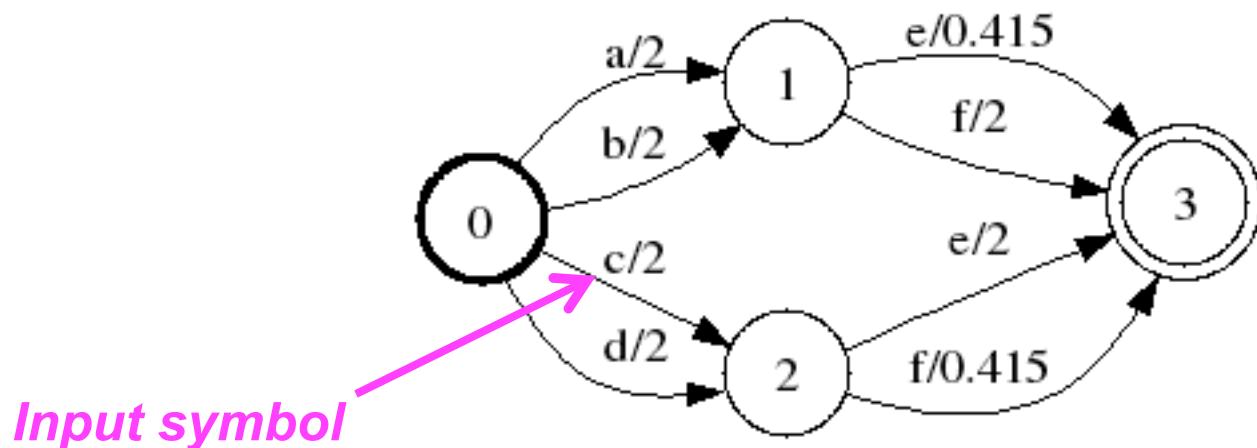
# Probabilistic Linguistic Knowledge

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



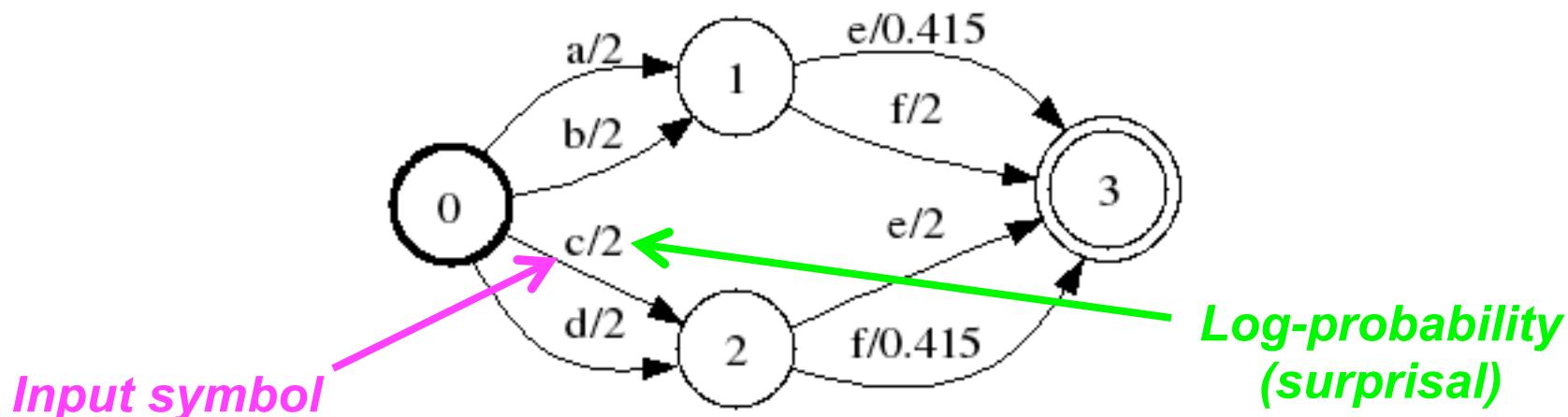
# Probabilistic Linguistic Knowledge

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



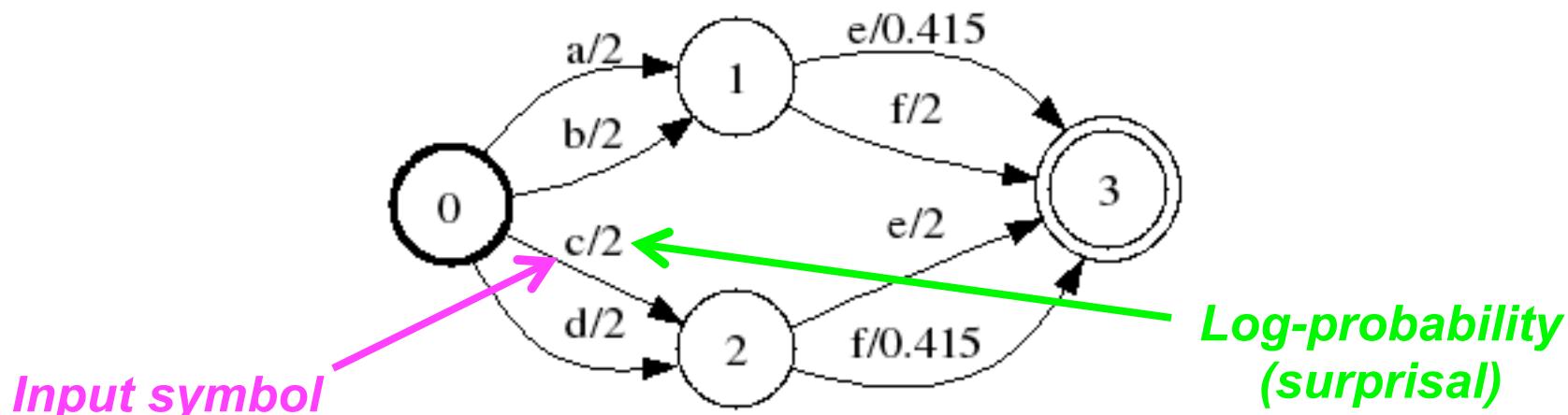
# Probabilistic Linguistic Knowledge

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



# Probabilistic Linguistic Knowledge

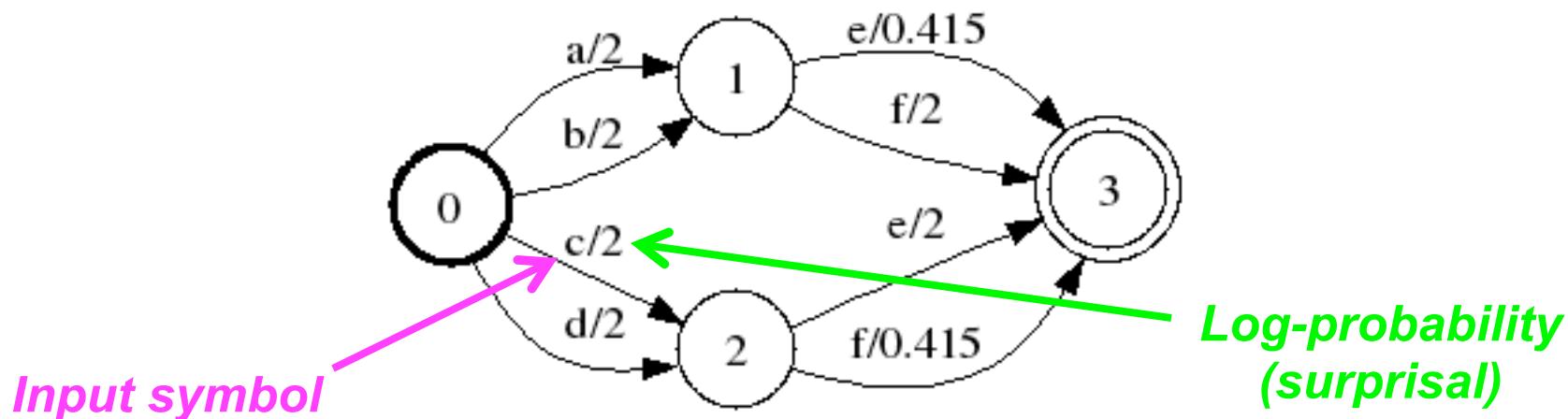
- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



- In position 1,  $\{-a, b, c, d\}$  equally likely; but in position 2:

# Probabilistic Linguistic Knowledge

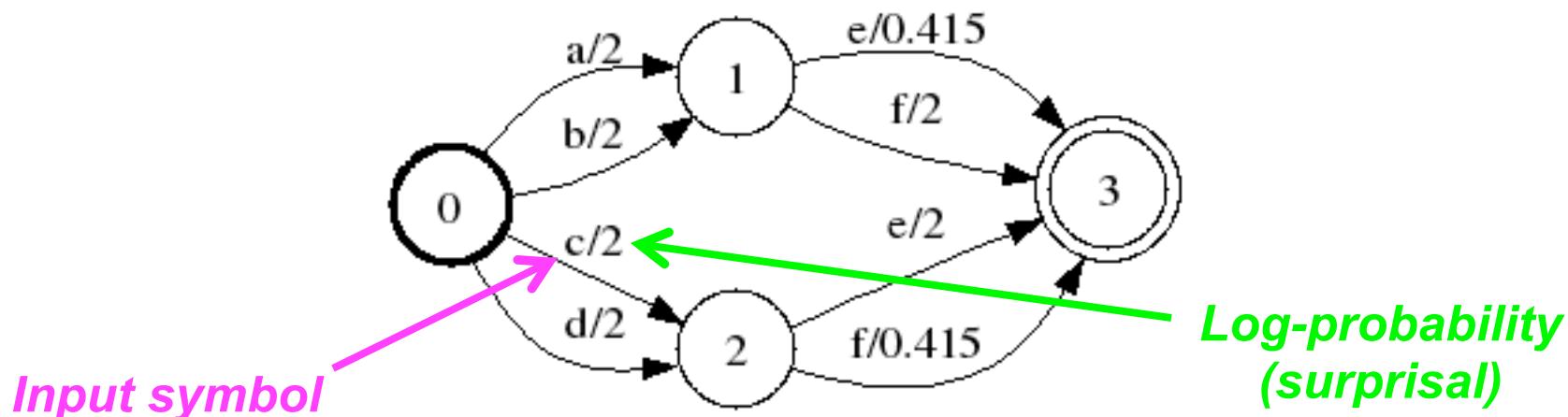
- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



- In position 1,  $\{-a, b, c, d\}$  equally likely; but in position 2:
  - $\{-a, b\}$  are usually followed by e, occasionally by f

# Probabilistic Linguistic Knowledge

- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



- In position 1,  $\{-a, b, c, d\}$  equally likely; but in position 2:
  - $\{-a, b\}$  are usually followed by e, occasionally by f
  - $\{-c, d\}$  are usually followed by f, occasionally by e

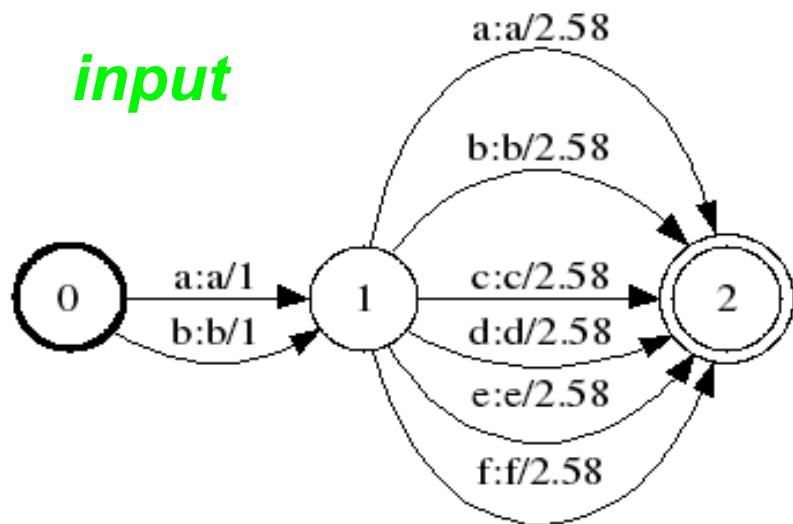
# Combining grammar & uncertain input

---

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*

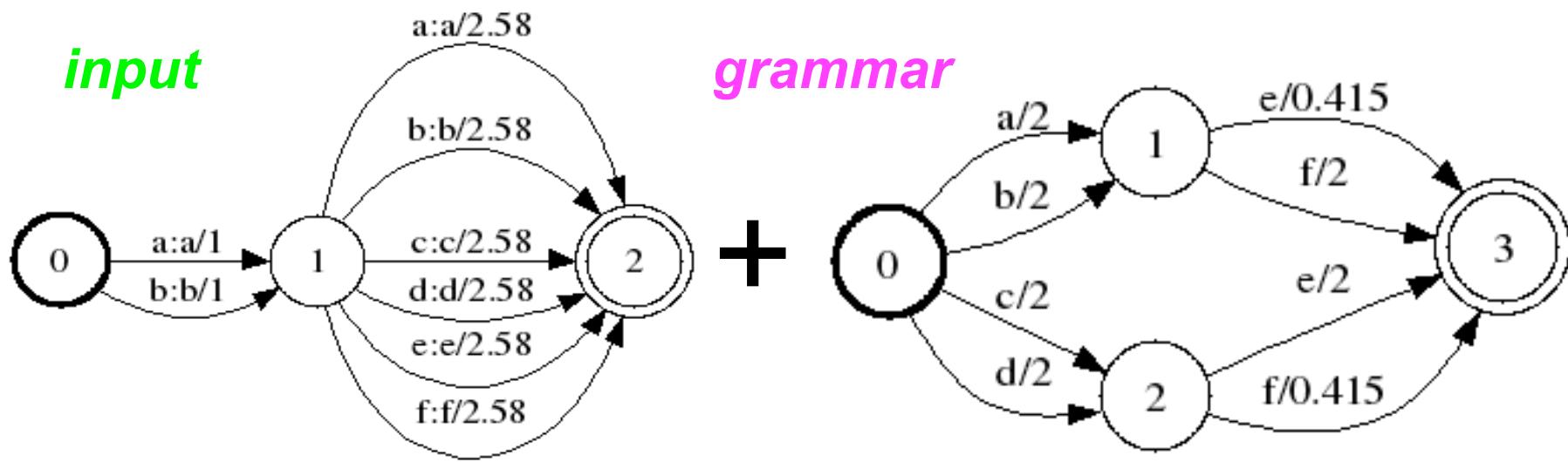
# Combining grammar & uncertain input

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*



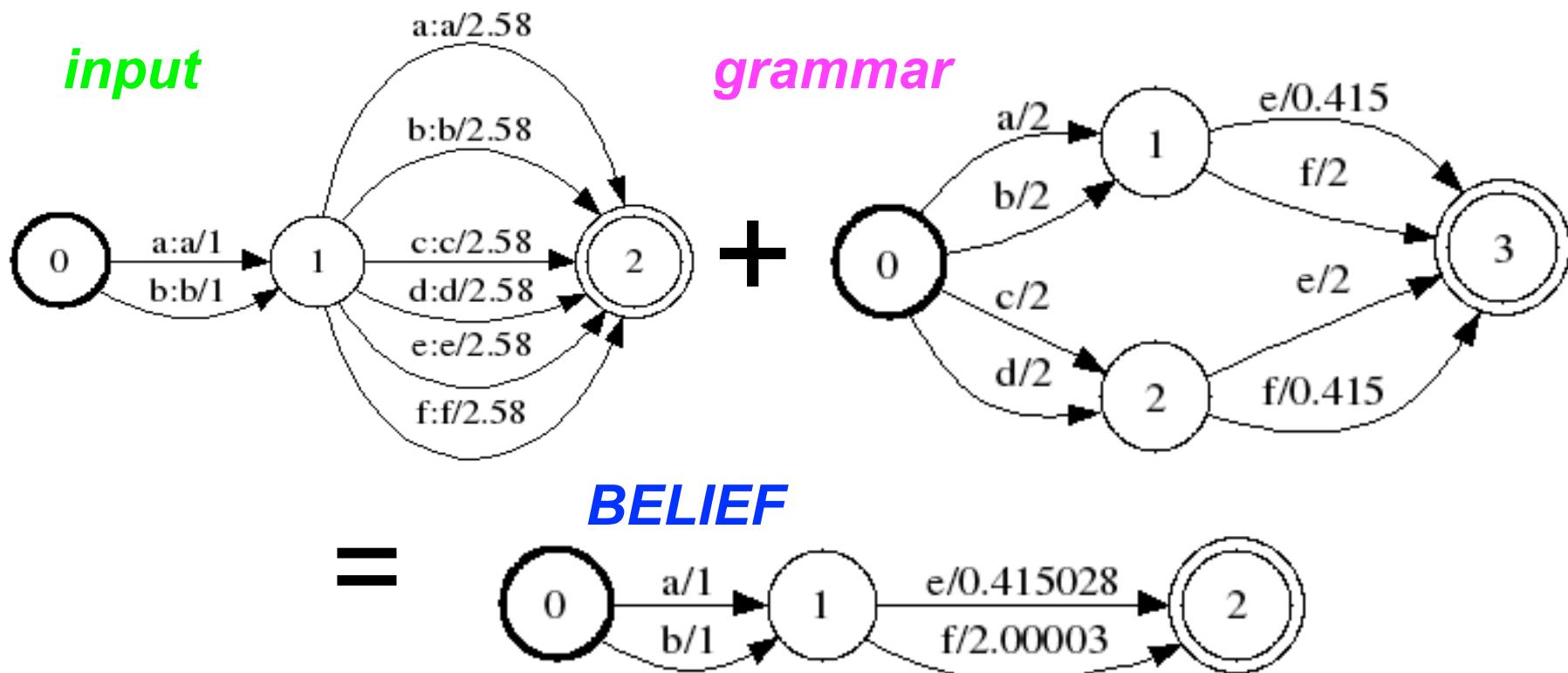
# Combining grammar & uncertain input

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*



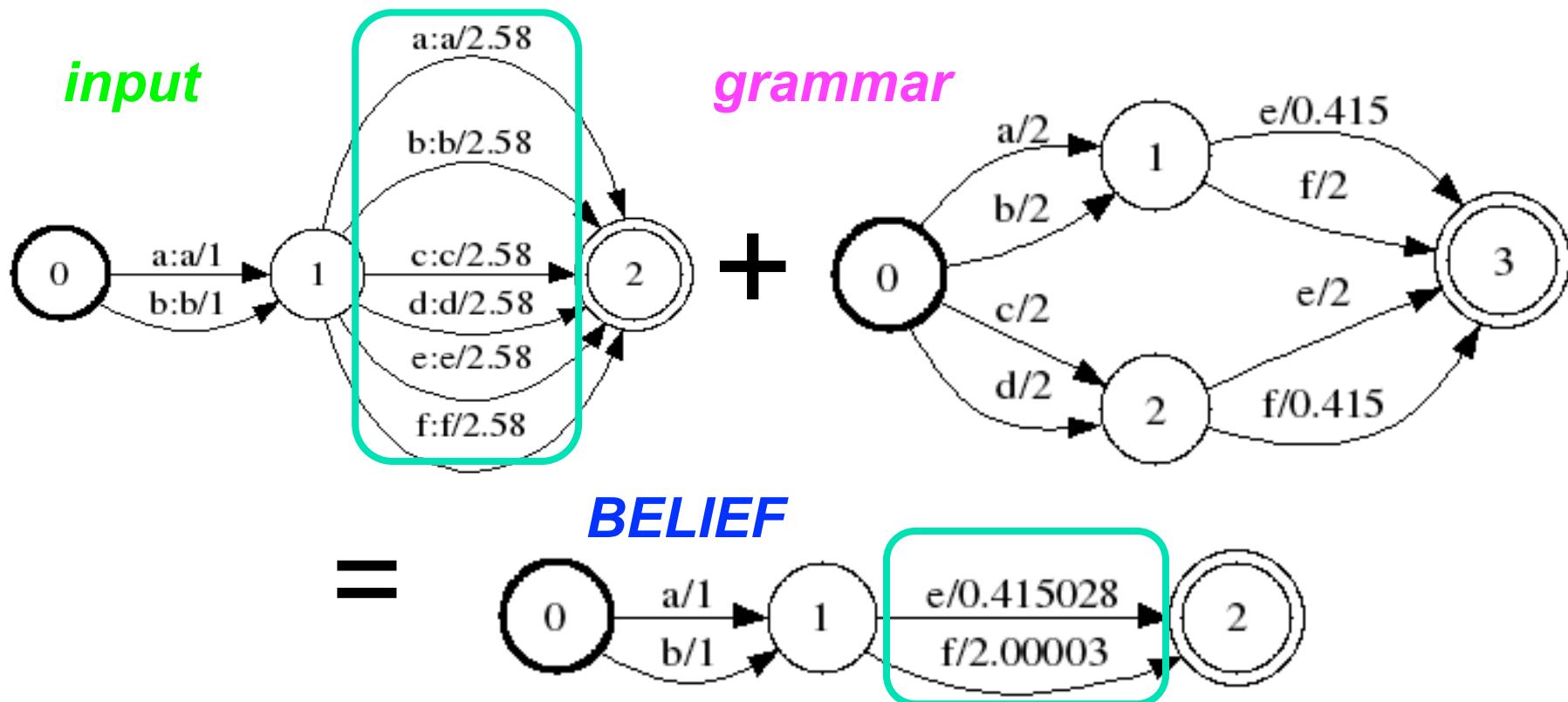
# Combining grammar & uncertain input

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*



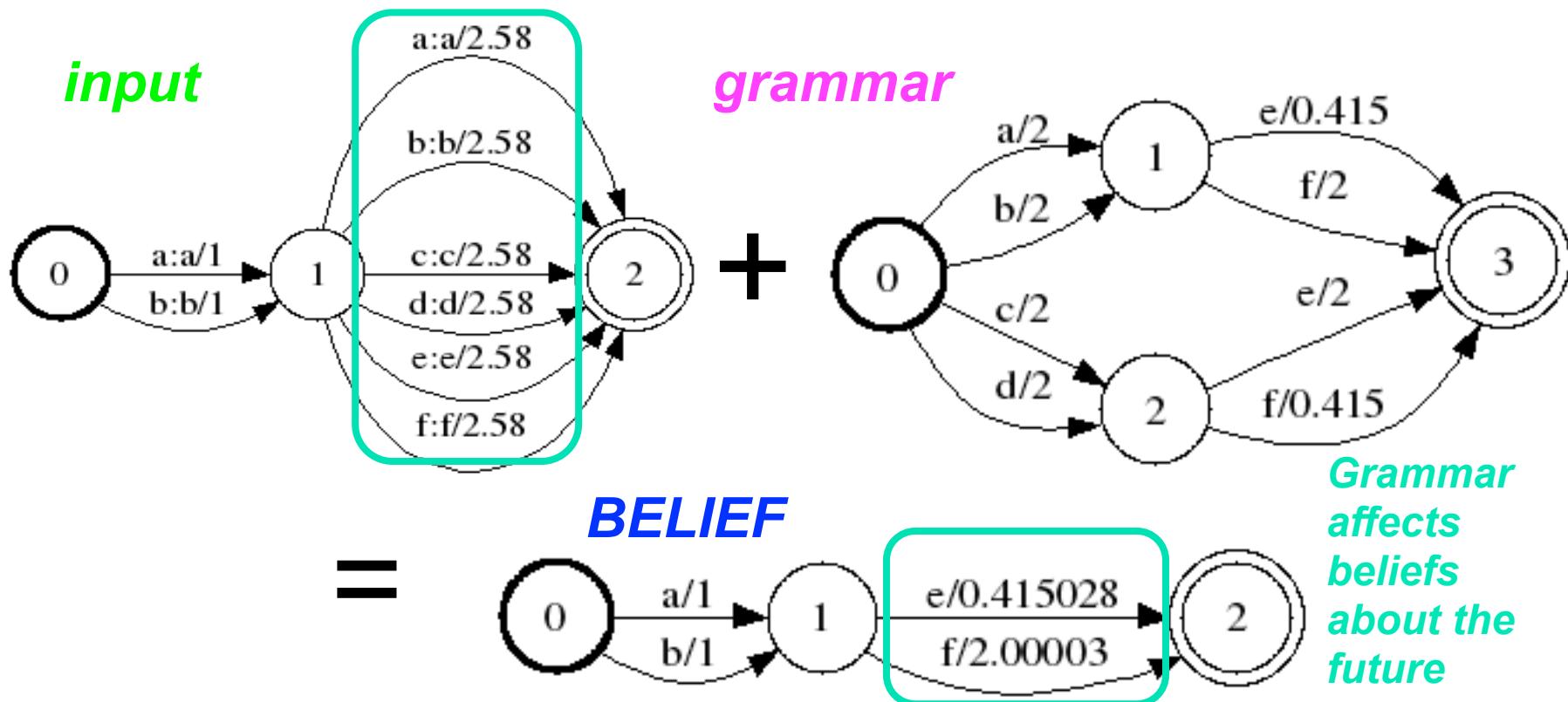
# Combining grammar & uncertain input

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*



# Combining grammar & uncertain input

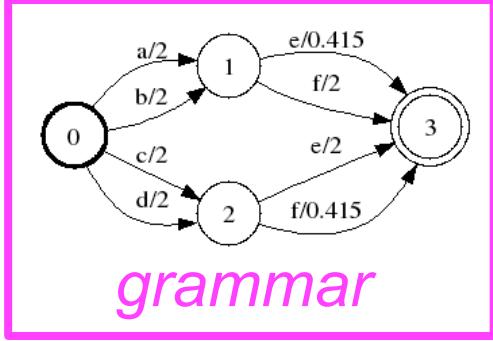
- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *weighted intersection*

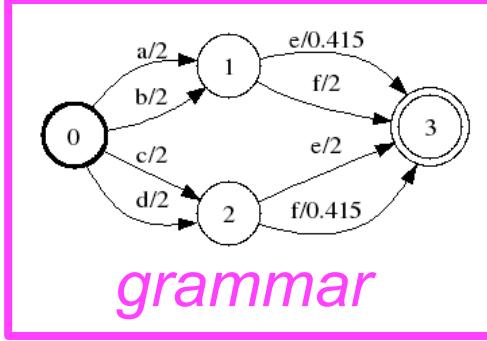


# Revising beliefs about the past

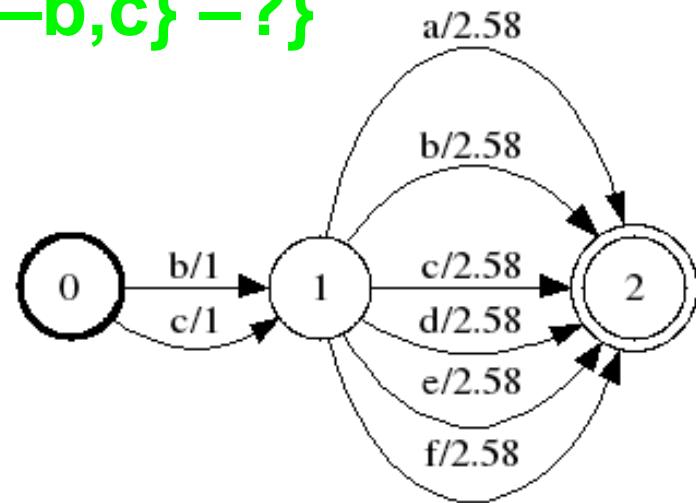
---

- When we're uncertain about the future, grammar + partial input can affect beliefs about what will happen
- With uncertainty of the past, grammar + future input can affect beliefs about *what has already happened*



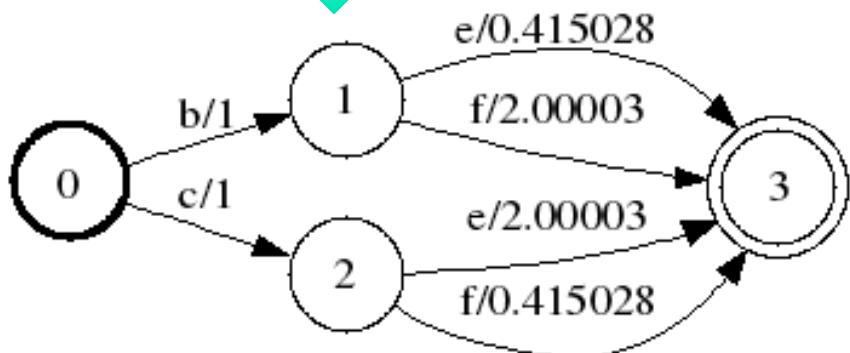
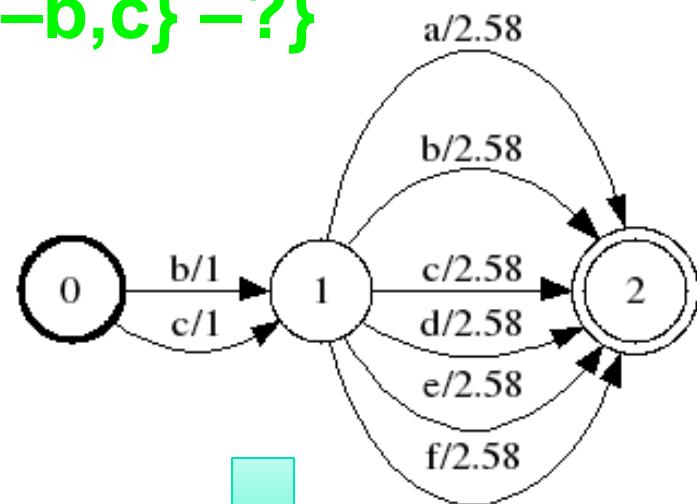
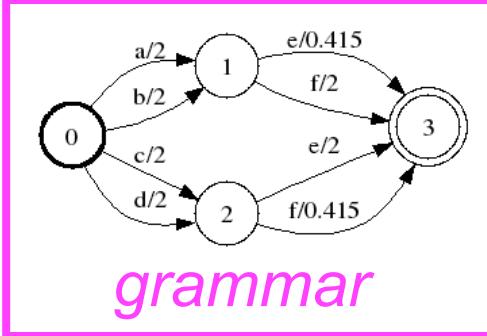


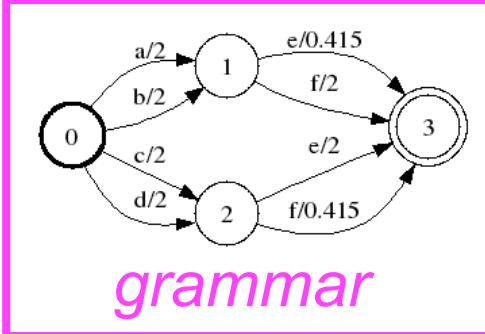
*word 1*  
**-b,c} -?}**



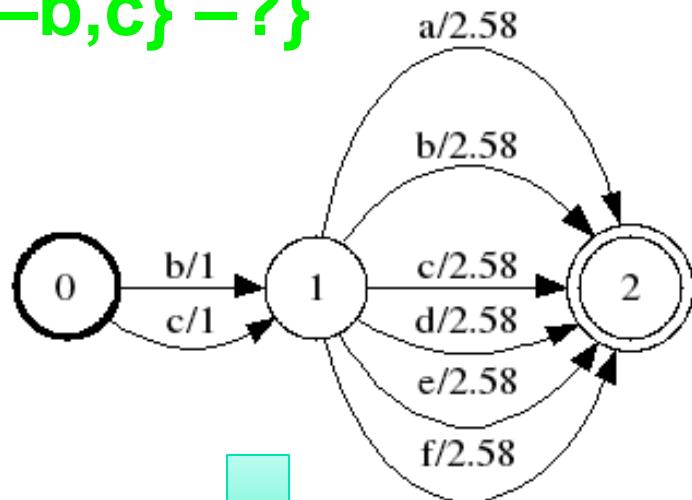
*word 1*

**-b,c} -?}**

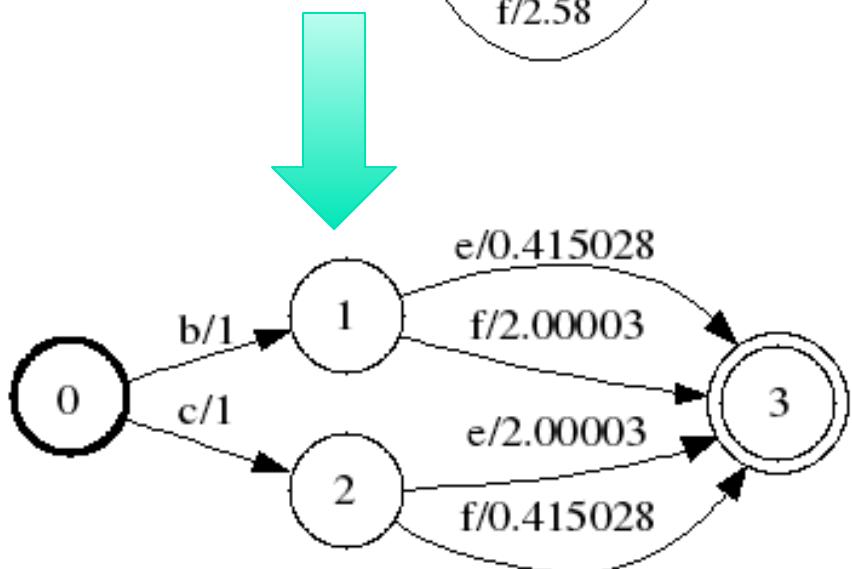
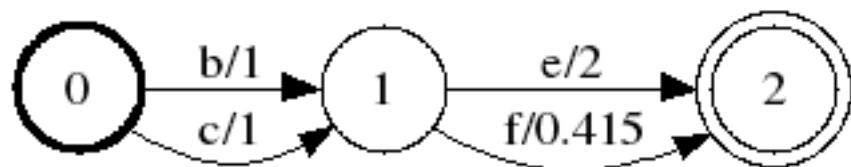


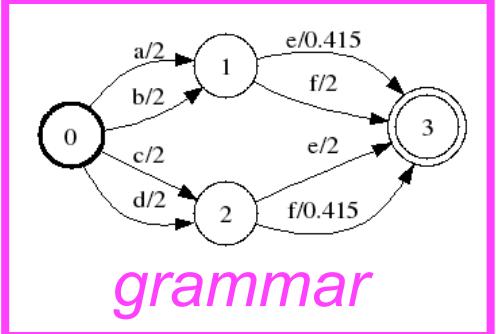


*word 1*  
**-b,c} -?}**

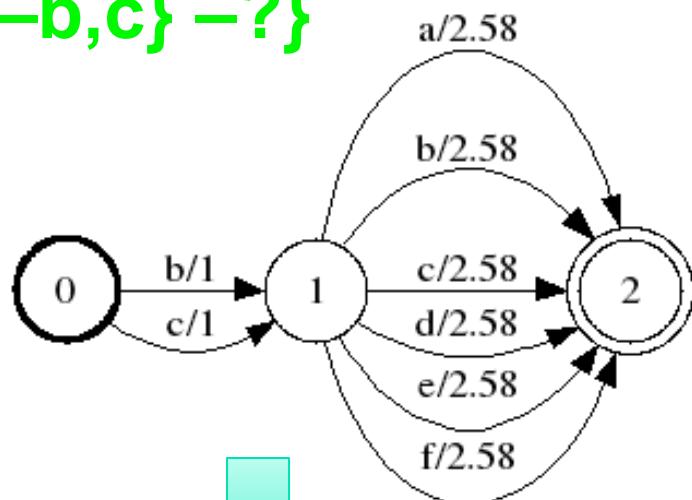


*words 1 + 2*  
**-b,c} -f,e}**

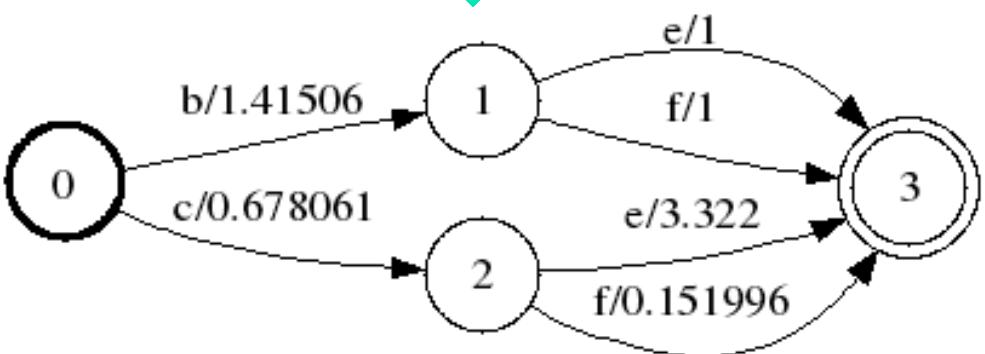
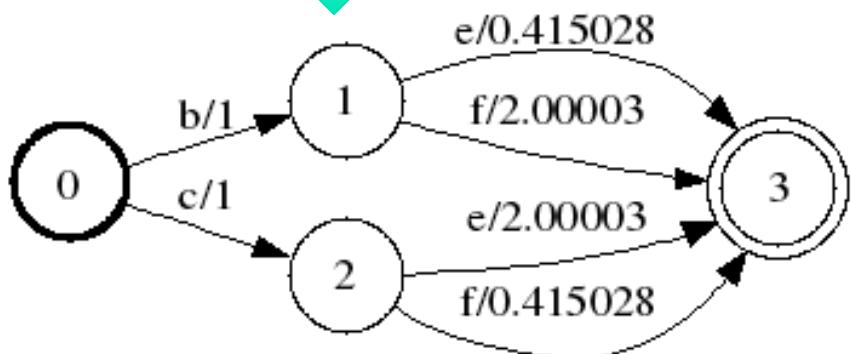
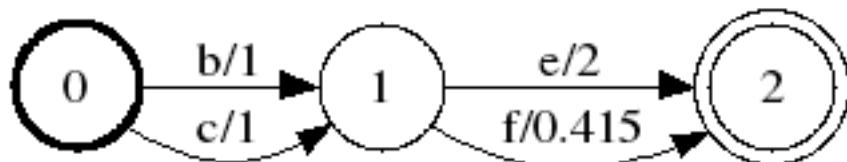


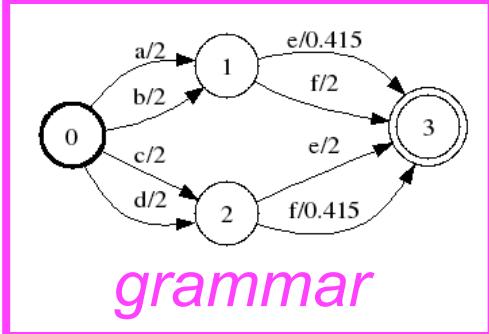


*word 1*  
**-b,c} -?}**



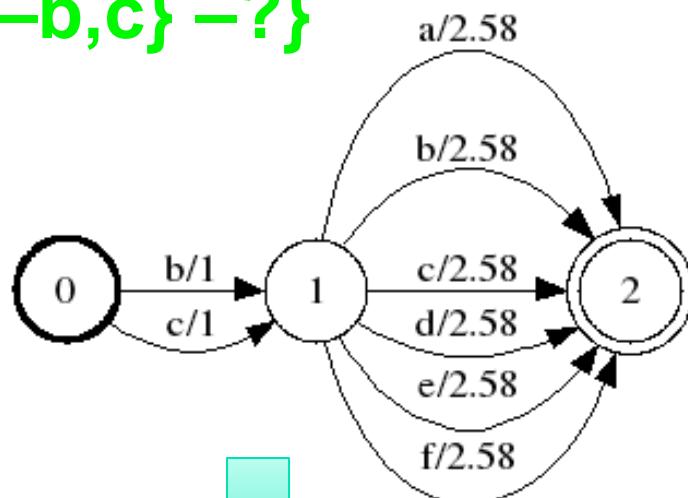
*words 1 + 2*  
**-b,c} -f,e}**





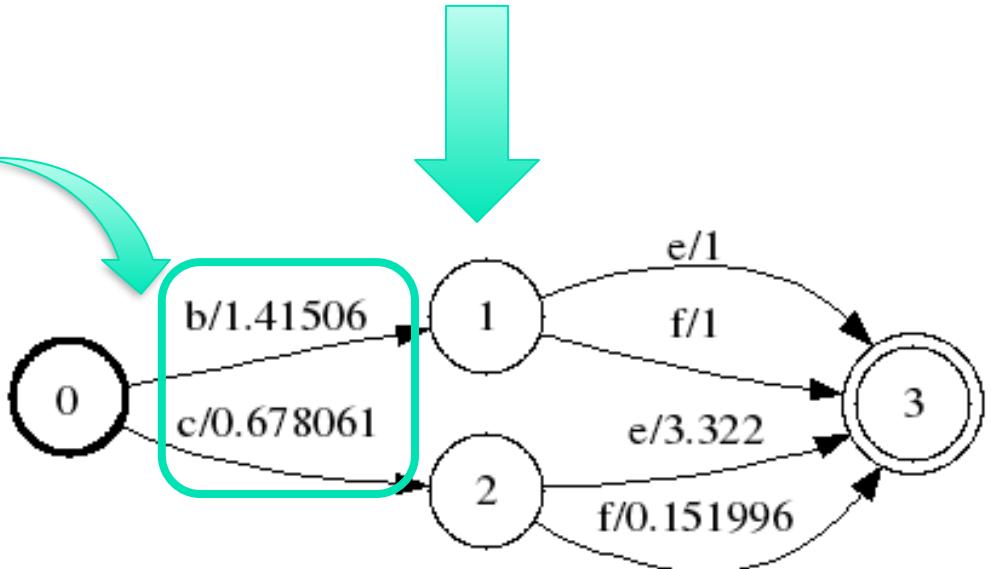
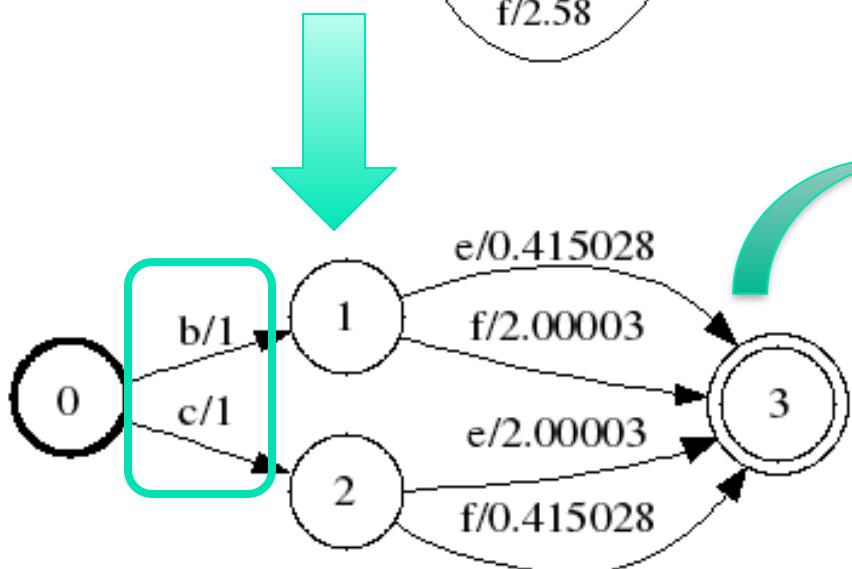
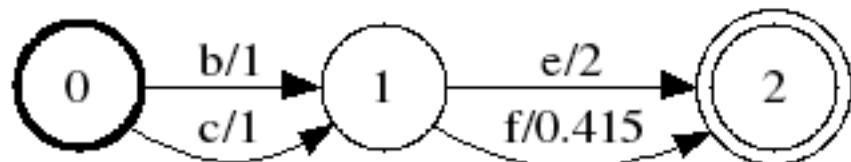
*word 1*

**-b,c} -?}**



*words 1 + 2*

**-b,c} -f,e}**



# The noisy-channel model (**FINAL**)

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto P_C(\mathbf{w})Q(\mathbf{w}, \mathbf{w}^*)$$

# The noisy-channel model (**FINAL**)

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} Q(\mathbf{w}, \mathbf{w}^*)$$

# The noisy-channel model (**FINAL**)

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} \underbrace{Q(\mathbf{w}, \mathbf{w}^*)}_{\text{Expected evidence}}$$

# The noisy-channel model (**FINAL**)

---

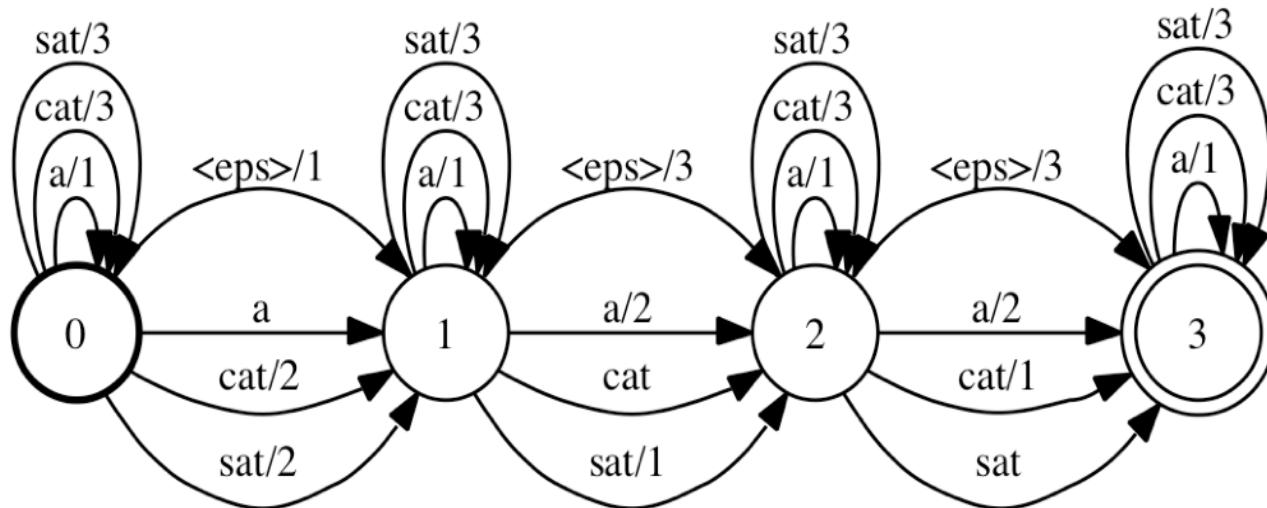
$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} \underbrace{Q(\mathbf{w}, \mathbf{w}^*)}_{\text{Expected evidence}}$$

- For  $Q(\mathbf{w}, \mathbf{w}^*)$ : a WFSA based on Levenshtein distance between words ( $K_{LD}$ ):

# The noisy-channel model (FINAL)

$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} \underbrace{Q(\mathbf{w}, \mathbf{w}^*)}_{\text{Expected evidence}}$$

- For  $Q(\mathbf{w}, \mathbf{w}^*)$ : a WFSA based on Levenshtein distance between words ( $K_{LD}$ ):

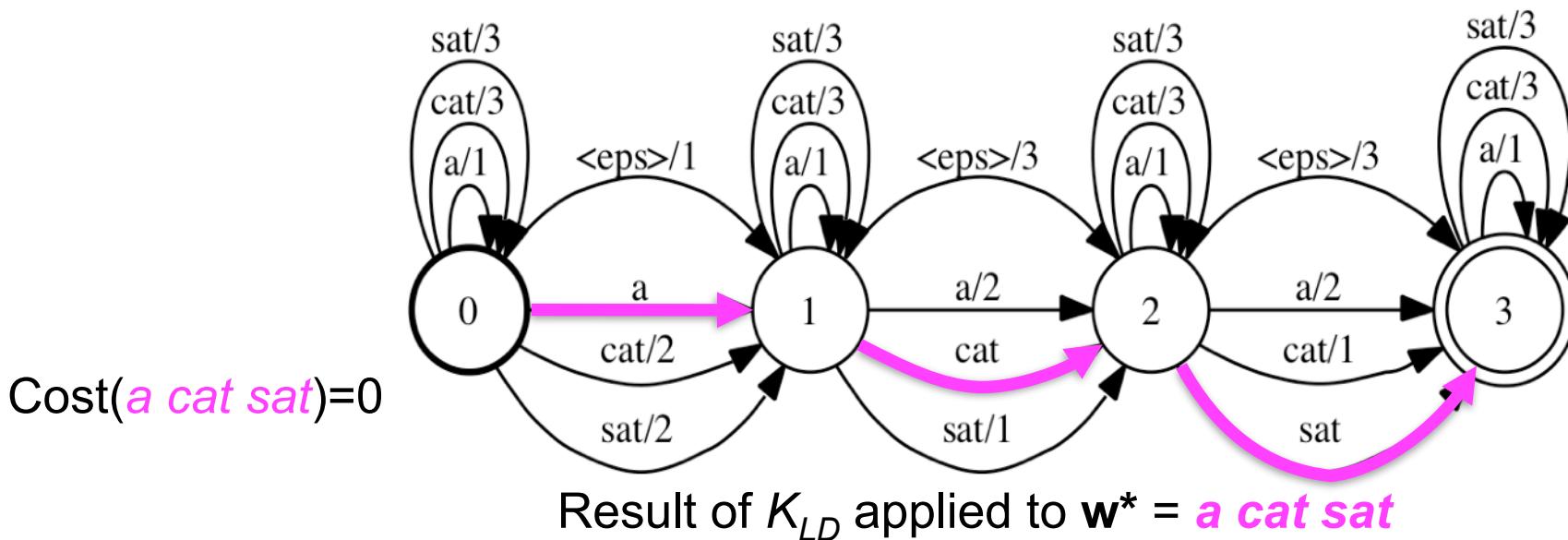


Result of  $K_{LD}$  applied to  $\mathbf{w}^* = \text{a cat sat}$

# The noisy-channel model (FINAL)

$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} \underbrace{Q(\mathbf{w}, \mathbf{w}^*)}_{\text{Expected evidence}}$$

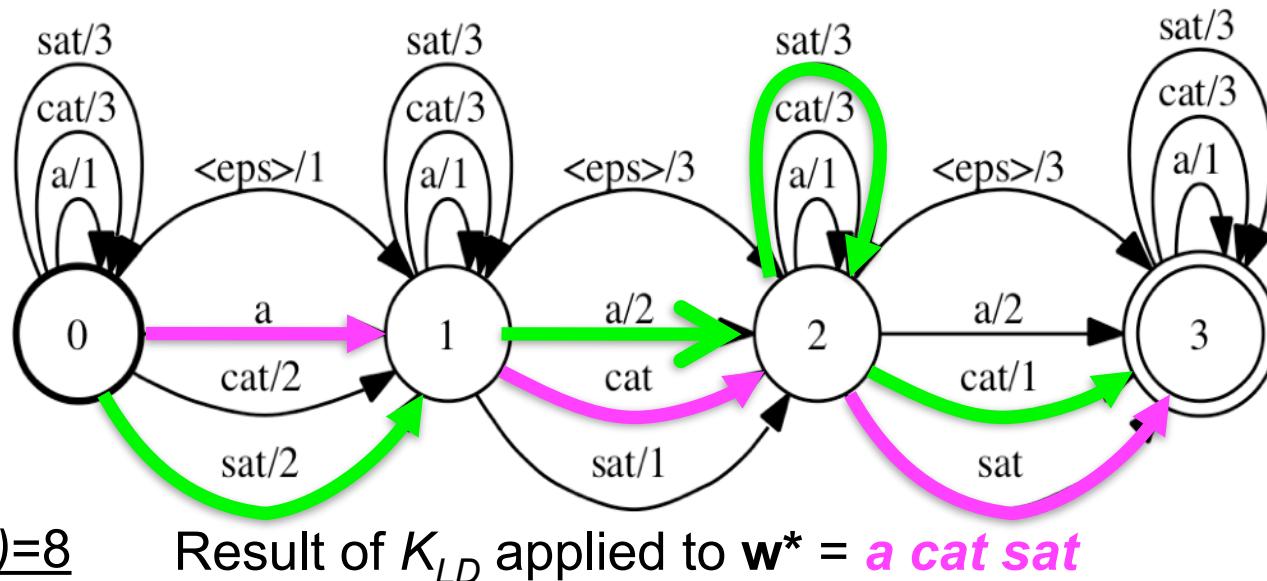
- For  $Q(\mathbf{w}, \mathbf{w}^*)$ : a WFSA based on Levenshtein distance between words ( $K_{LD}$ ):



# The noisy-channel model (FINAL)

$$P(\mathbf{w}|\mathbf{w}^*) \propto \underbrace{P_C(\mathbf{w})}_{\text{Prior}} \underbrace{Q(\mathbf{w}, \mathbf{w}^*)}_{\text{Expected evidence}}$$

- For  $Q(\mathbf{w}, \mathbf{w}^*)$ : a WFSA based on Levenshtein distance between words ( $K_{LD}$ ):



# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations
  4. Derive predicted optimal behavior given 1—3
  5. Compare predictions with empirical data
  6. If necessary, iterate 1—5

# Incremental inference under uncertain input

---

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(and?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(and?)  
(as?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(and?)  
(and?)  
(as?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

- Near-neighbors make the “incorrect” analysis “correct”:

(and?)  
(as?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(and?) (and?)  
(as?) (that?)  
          (who?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(that?)            (and?)            (and?)  
                      (as?)                (that?)  
    (who?)

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

- Near-neighbors make the “incorrect” analysis “correct”:

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(that?) (and?) (and?)  
(who?) (as?) (that?)  
                          (who?)

Any of these changes makes **tossed** a main verb!!!

*The coach smiled at the player **tossed** the frisbee*

# Incremental inference under uncertain input

---

- Near-neighbors make the “incorrect” analysis “correct”:

(that?) (and?) (and?)  
(who?) (as?) (that?)  
Any of these changes  
makes **tossed** a main  
verb!!!

*The coach smiled at the player **tossed** the frisbee*

- Hypothesis: the boggle at “tossed” involves *what the comprehender wonders whether she might have seen*

# Rational analysis

---

- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively
  1. Specify precisely the goals of the cognitive system
  2. Formalize model of the environment adapted to
  3. Make minimal assumptions re: computational limitations
  4. Derive predicted optimal behavior given 1—3
  5. Compare predictions with empirical data
  6. If necessary, iterate 1—5

# The core of the intuition

---

*the coach smiled...*

# The core of the intuition

---

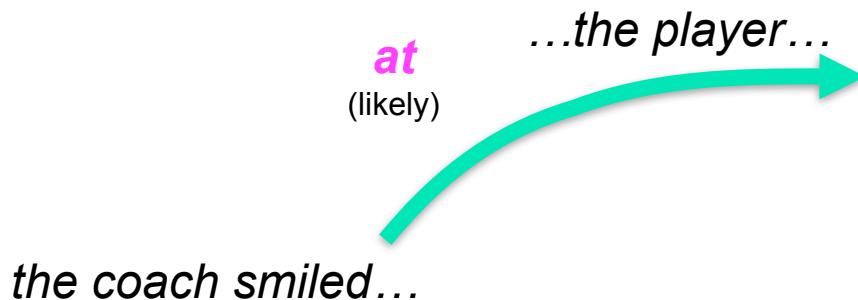
- Grammar & input come together to determine two possible “paths” through the partial sentence:

*the coach smiled...*

# The core of the intuition

---

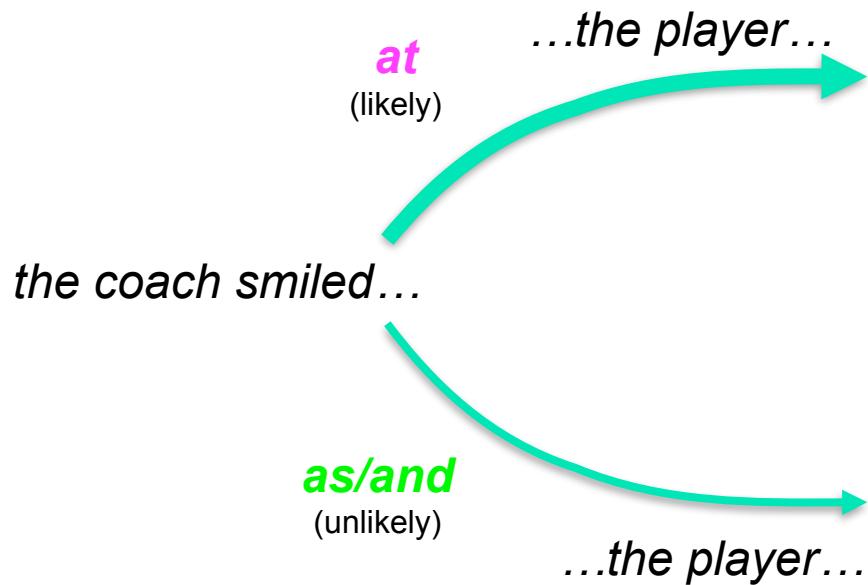
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



# The core of the intuition

---

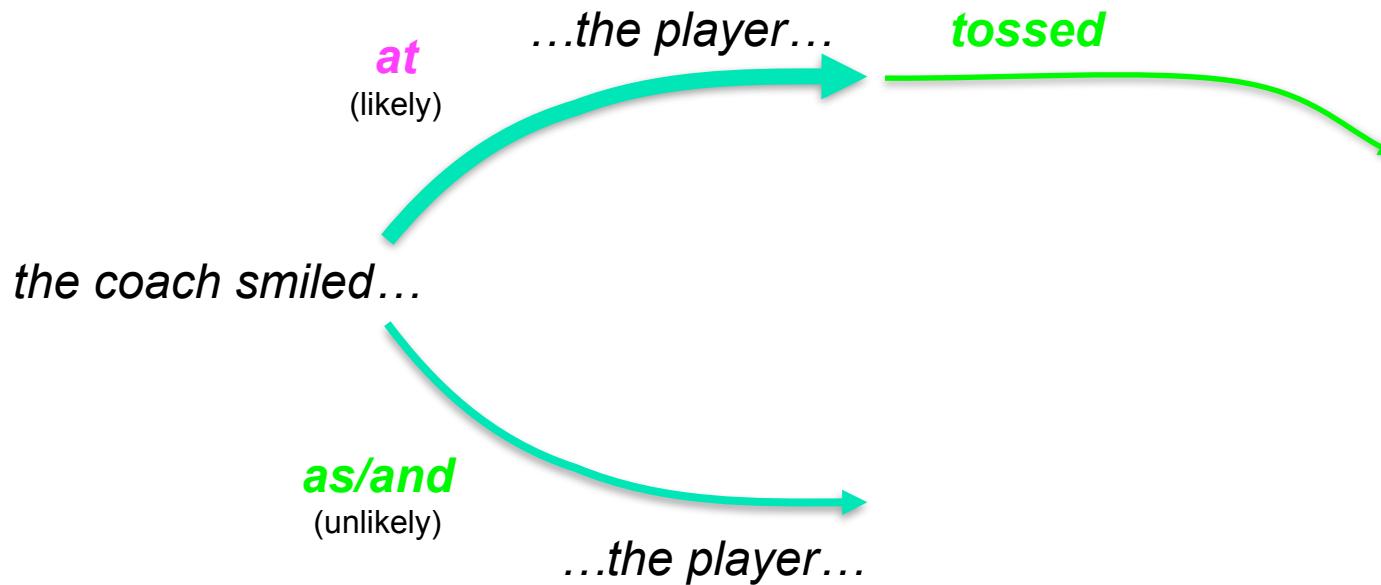
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



# The core of the intuition

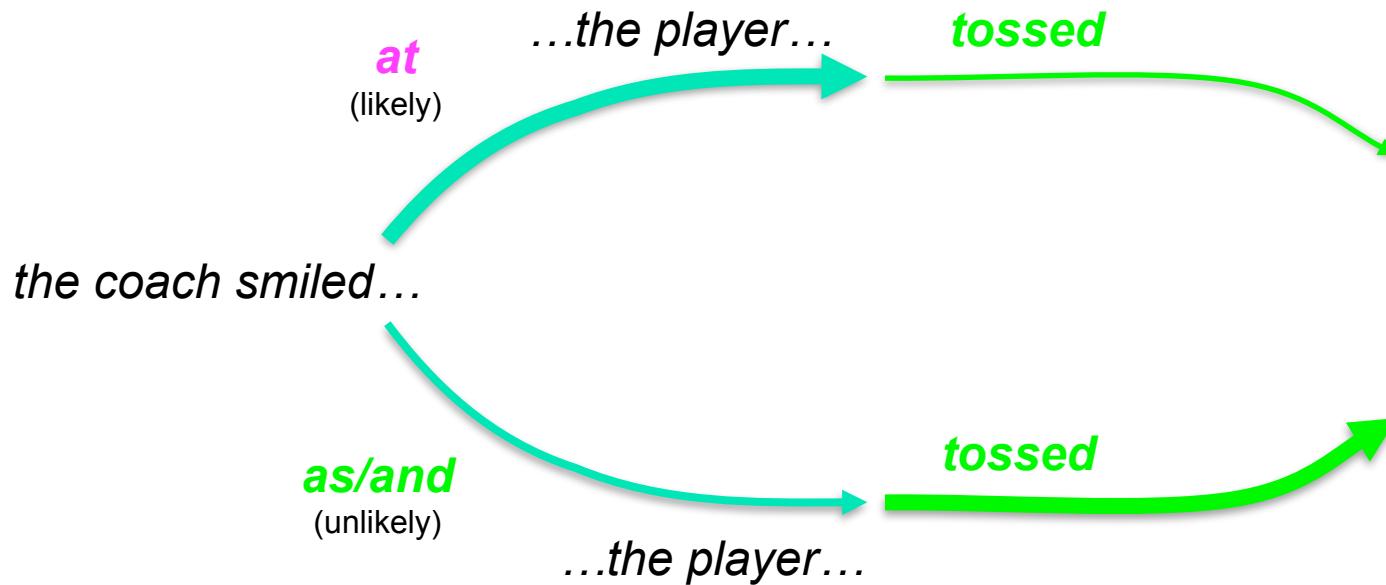
---

- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



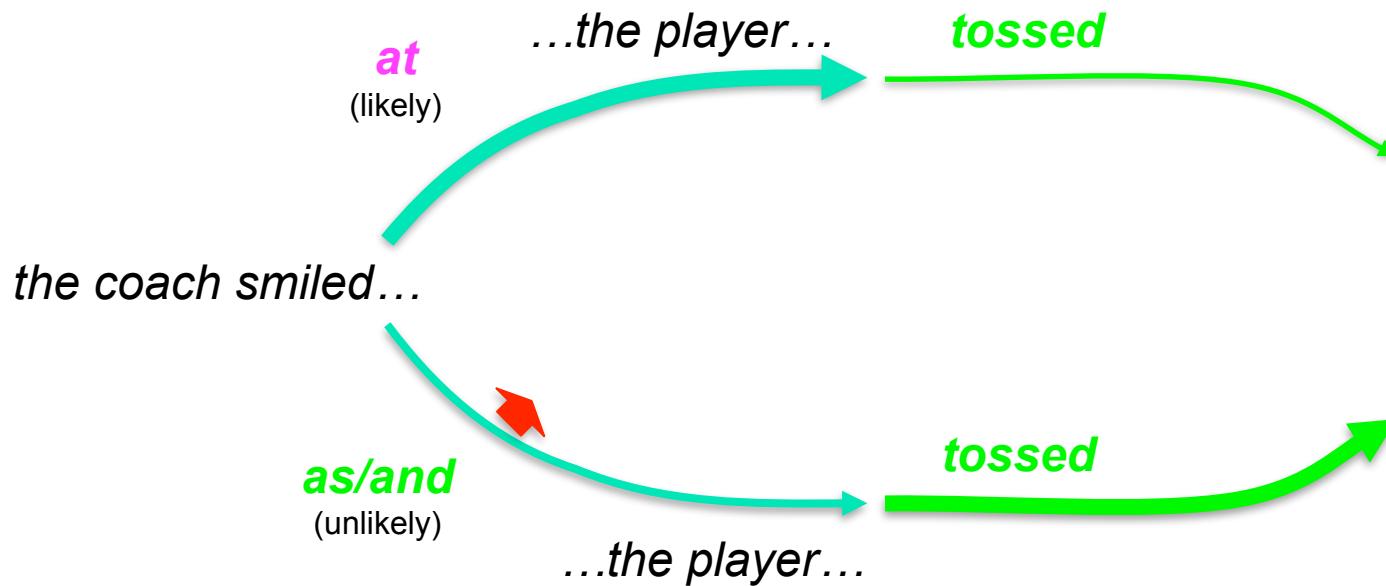
# The core of the intuition

- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



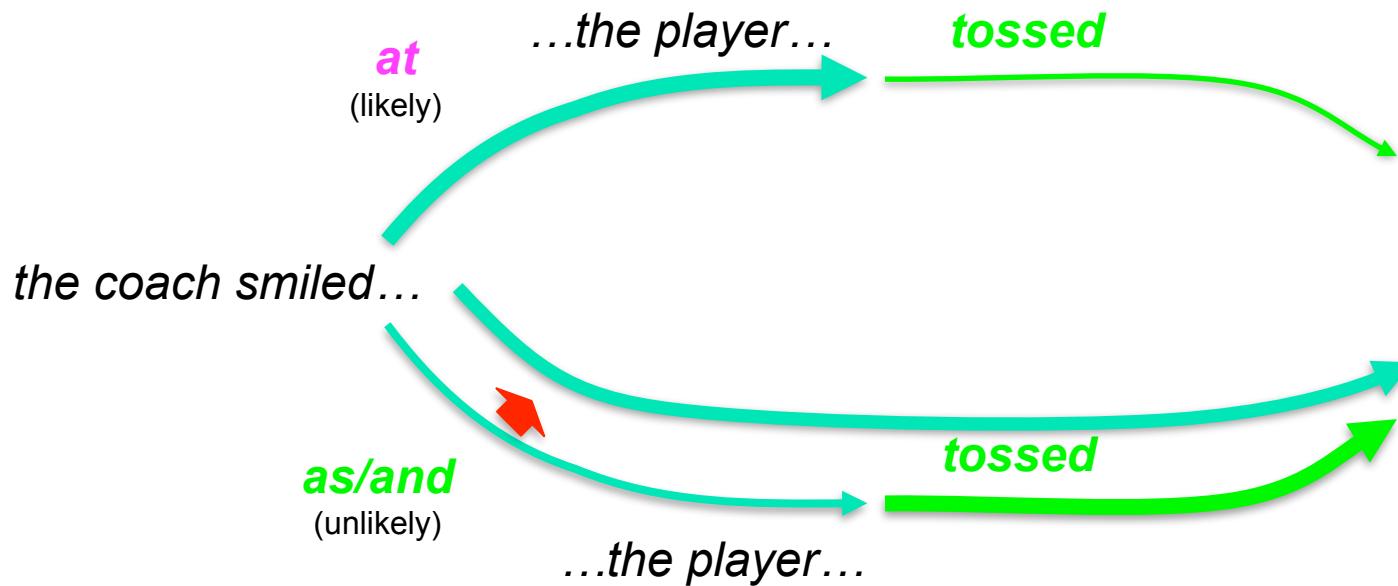
# The core of the intuition

- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



# The core of the intuition

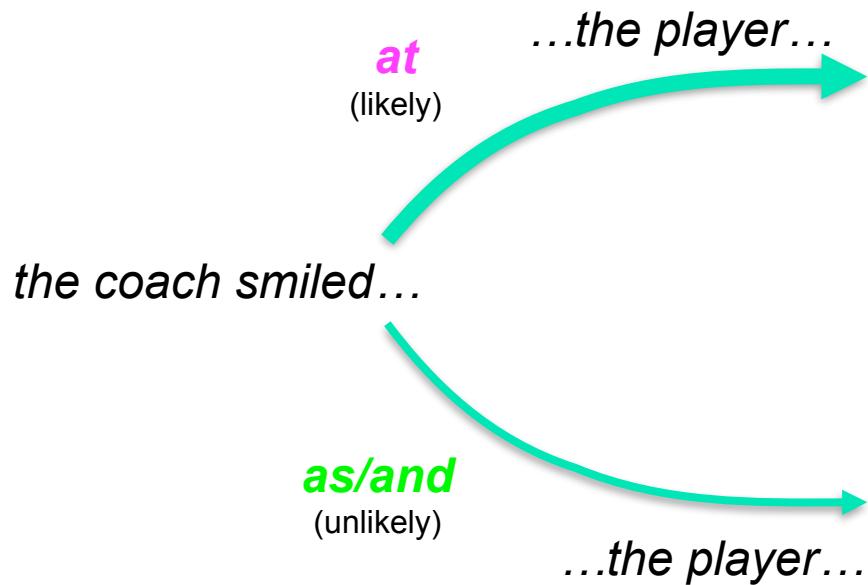
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



# The core of the intuition

---

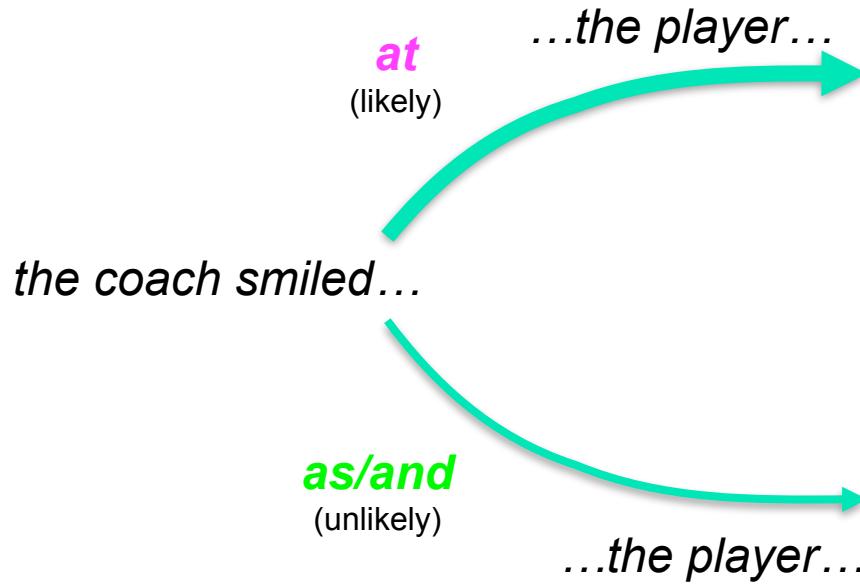
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



# The core of the intuition

---

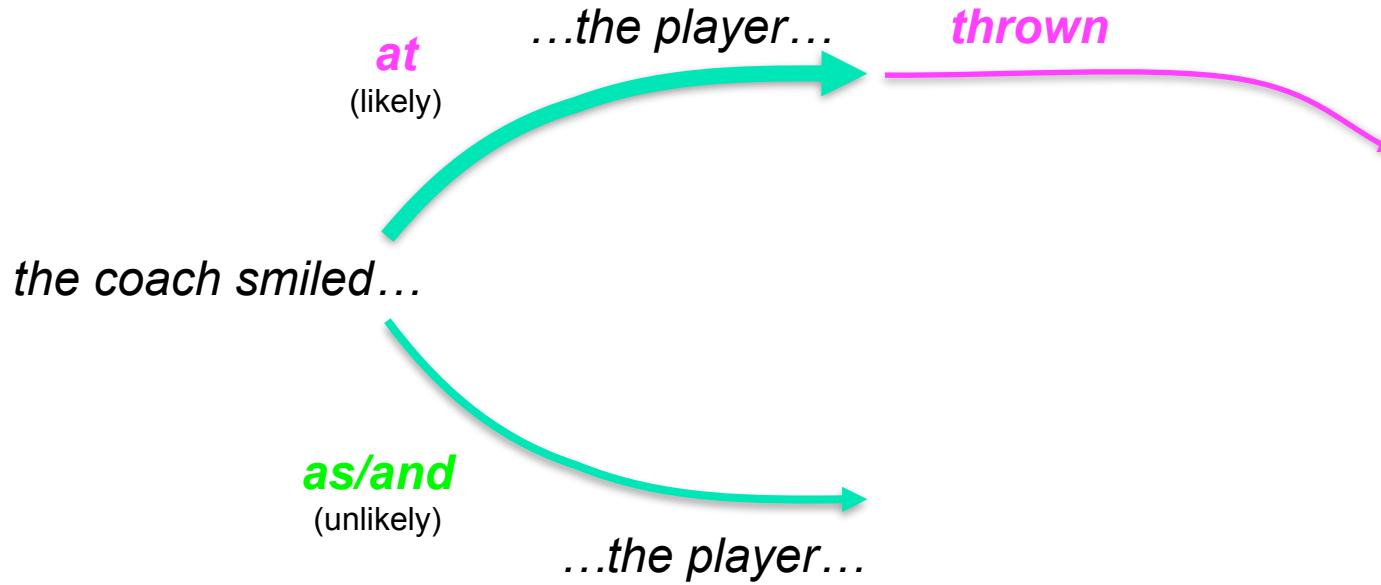
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

# The core of the intuition

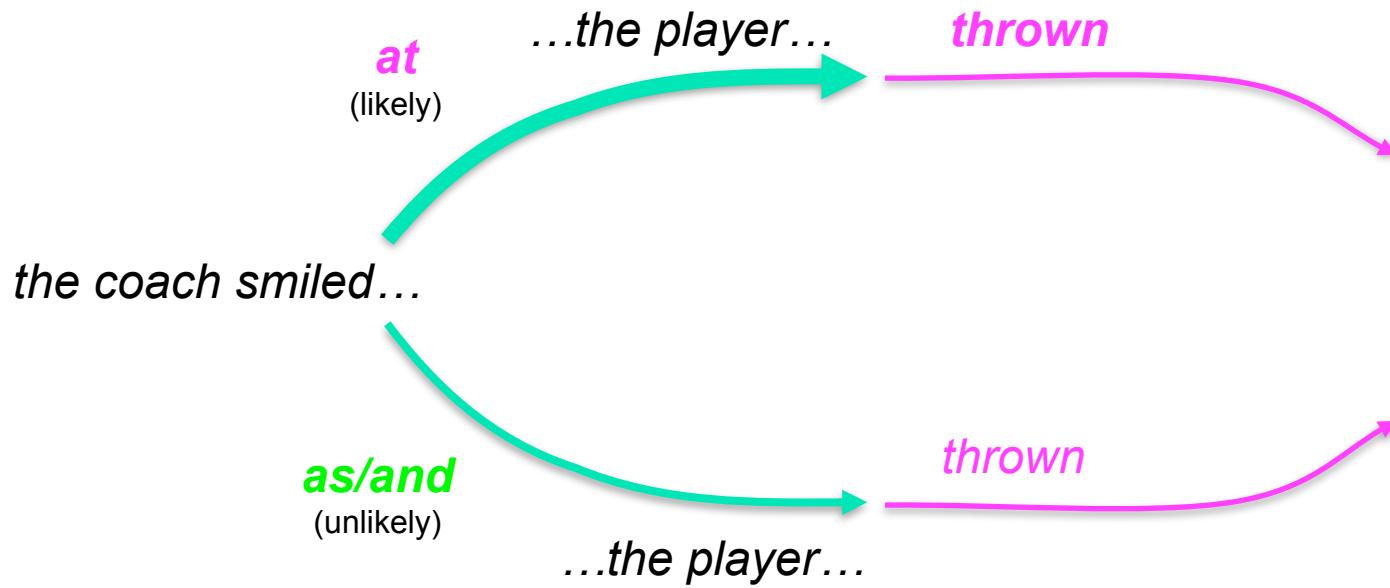
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

# The core of the intuition

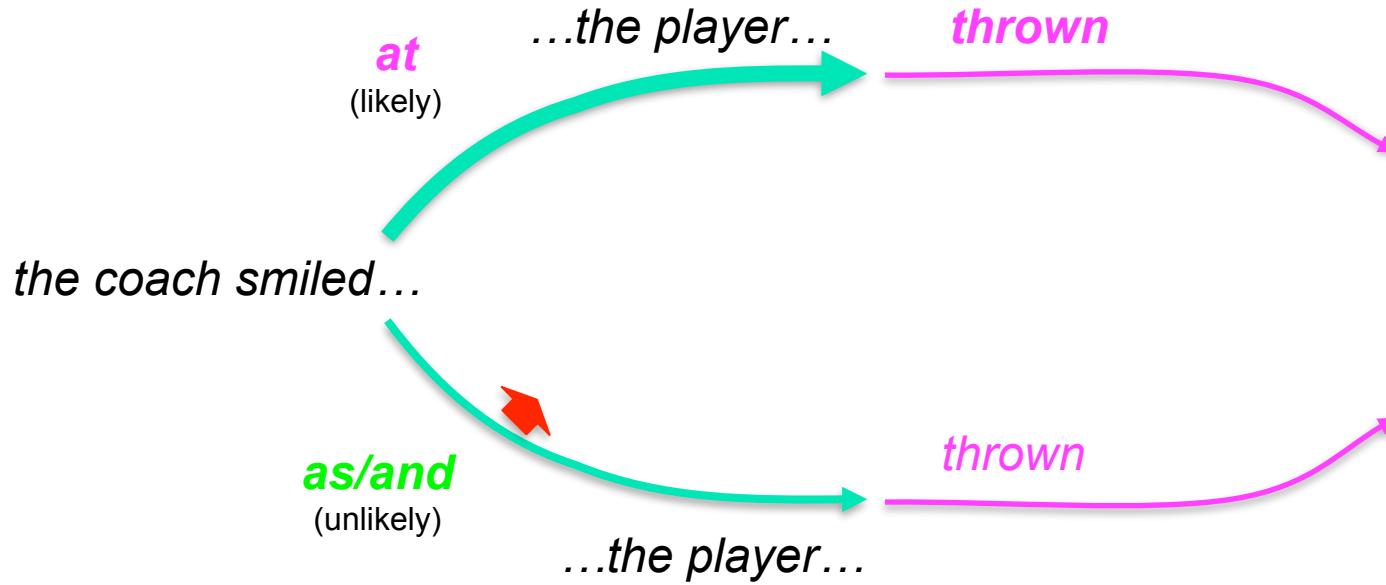
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

# The core of the intuition

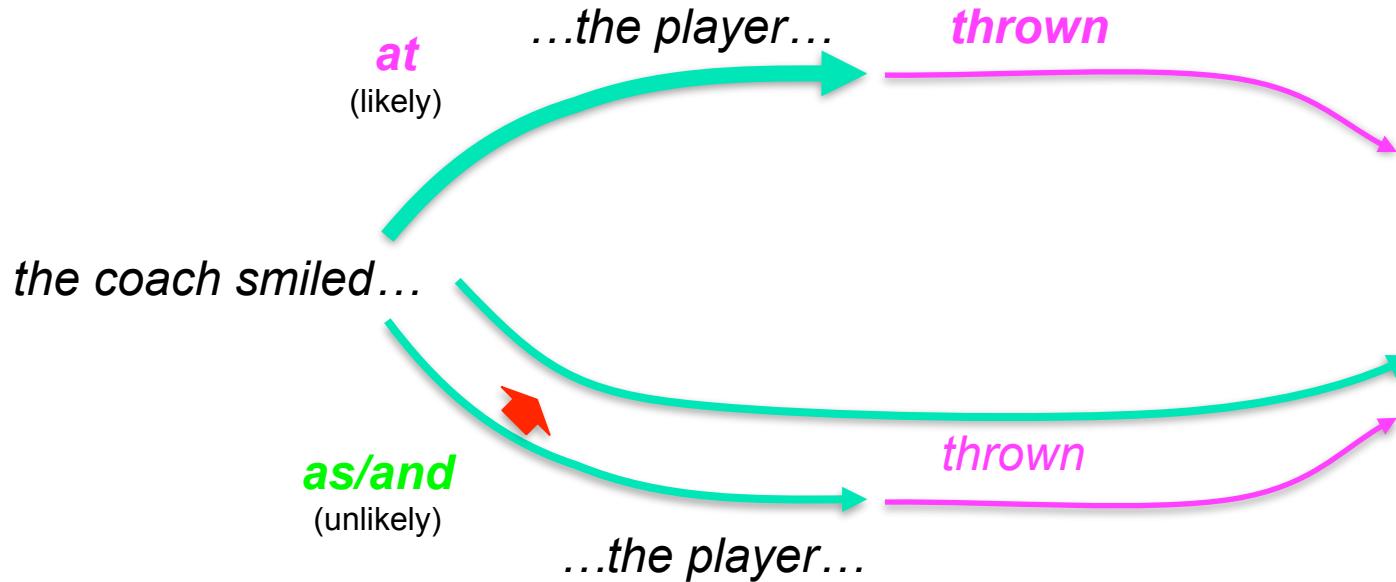
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

# The core of the intuition

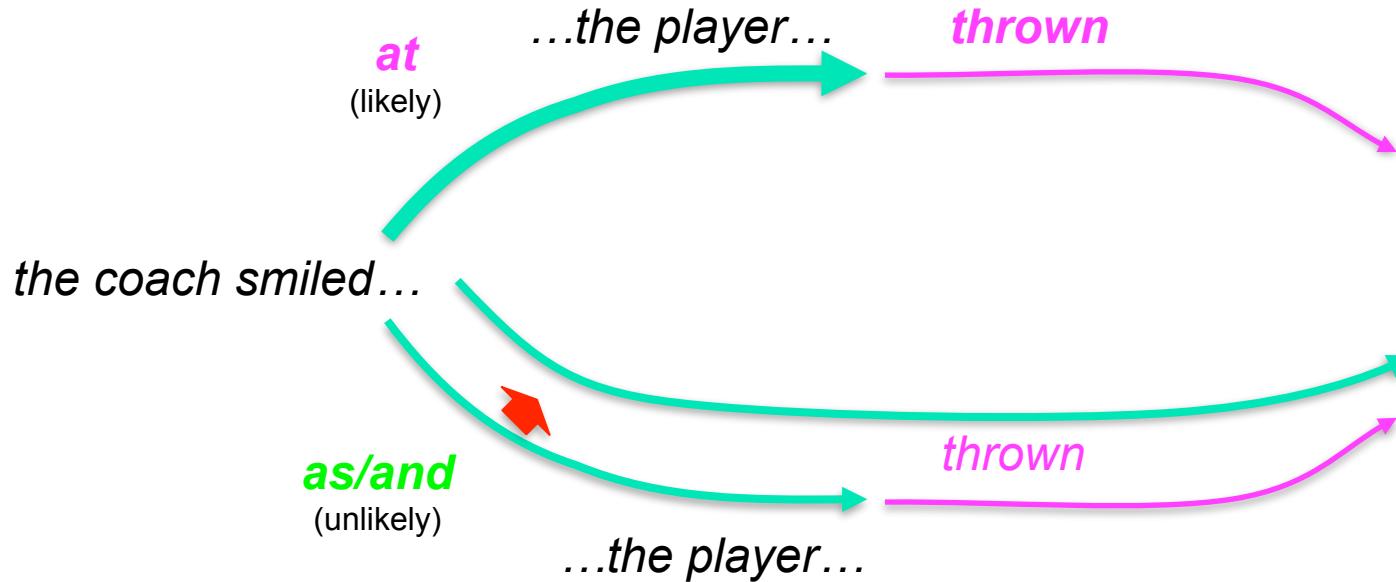
- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition

# The core of the intuition

- Grammar & input come together to determine two possible “paths” through the partial sentence: *(line thickness ≈ probability)*



- *tossed* is more likely to happen along the bottom path
  - This creates a large shift in belief in the *tossed* condition
- *thrown* is very unlikely to happen along the bottom path
  - As a result, there is no corresponding shift in belief

# Ingredients for the model

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto P_C(\mathbf{w}) Q(\mathbf{w}, \mathbf{w}^*)$$


Prior      Expected evidence

# Ingredients for the model

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto P_C(\mathbf{w}) Q(\mathbf{w}, \mathbf{w}^*)$$


Prior      Expected evidence

- $Q(\mathbf{w}, \mathbf{w}^*)$  comes from  $K_{LD}$  (with minor changes)

# Ingredients for the model

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto P_C(\mathbf{w}) Q(\mathbf{w}, \mathbf{w}^*)$$

The equation is shown with two curly braces underneath. The first brace, under  $P_C(\mathbf{w})$ , is labeled "Prior" in blue. The second brace, under  $Q(\mathbf{w}, \mathbf{w}^*)$ , is labeled "Expected evidence" in blue.

- $Q(\mathbf{w}, \mathbf{w}^*)$  comes from  $K_{LD}$  (with minor changes)
- $P_C(\mathbf{w})$  comes from a probabilistic grammar (this time finite-state)

# Ingredients for the model

---

$$P(\mathbf{w}|\mathbf{w}^*) \propto P_C(\mathbf{w}) Q(\mathbf{w}, \mathbf{w}^*)$$

The equation is shown with two curly braces underneath. The first brace covers the term  $P_C(\mathbf{w})$  and is labeled "Prior" below it. The second brace covers the term  $Q(\mathbf{w}, \mathbf{w}^*)$  and is labeled "Expected evidence" below it.

- $Q(\mathbf{w}, \mathbf{w}^*)$  comes from  $K_{LD}$  (with minor changes)
- $P_C(\mathbf{w})$  comes from a probabilistic grammar (this time finite-state)
- We need one more ingredient:
  - a **quantified signal** of the alarm induced by word  $w$ , about changes in beliefs about the past

# Quantifying alarm about the past

---

# Quantifying alarm about the past

---

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P || Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

# Quantifying alarm about the past

---

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*

# Quantifying alarm about the past

---

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$

# Quantifying alarm about the past

---

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$

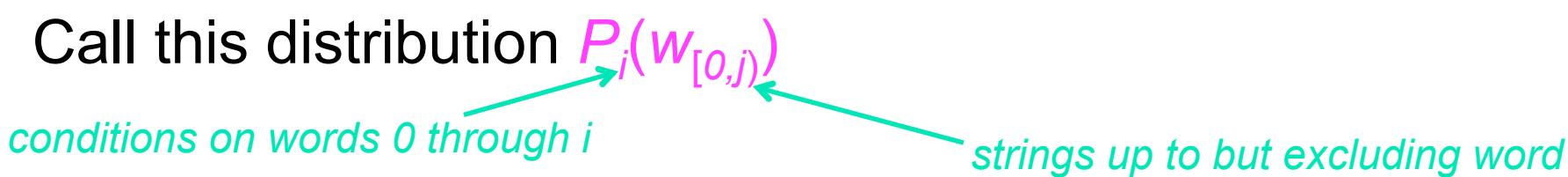
*conditions on words 0 through i*

# Quantifying alarm about the past

---

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$   


conditions on words 0 through  $i$

strings up to but excluding word  $j$

# Quantifying alarm about the past

---

- **Relative Entropy** (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$   

- The change induced by  $w_i$  is the **error identification signal EIS<sub>i</sub>**, defined as

$$D \left( P_i \left( w_{[0,i]} \right) \parallel P_{i-1} \left( w_{[0,i]} \right) \right)$$

# Quantifying alarm about the past

---

- **Relative Entropy** (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$   


*conditions on words 0 through i*      *strings up to but excluding word j*
- The change induced by  $w_i$  is the **error identification signal EIS<sub>i</sub>**, defined as

$$D \left( \underline{P_i(w_{[0,i]})} \parallel P_{i-1}(w_{[0,i]}) \right)$$

*new distribution*

# Quantifying alarm about the past

- **Relative Entropy** (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)

$$D(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Our distribution of interest is *probabilities over the previous words in the sentence*
- Call this distribution  $P_i(w_{[0,j]})$   

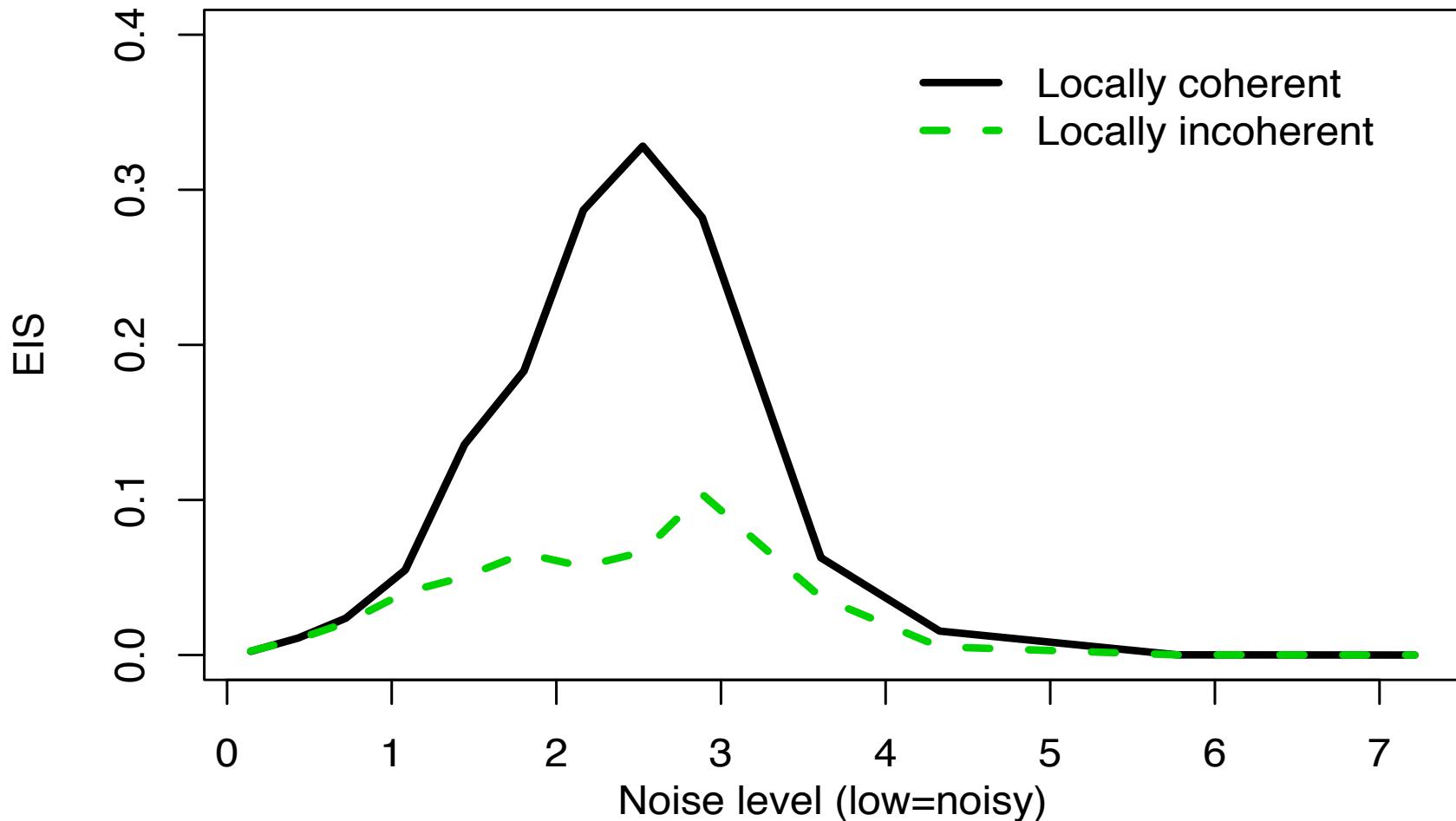

*conditions on words 0 through i*      *strings up to but excluding word j*
- The change induced by  $w_i$  is the **error identification signal EIS<sub>i</sub>**, defined as

$$D \left( \underline{P_i(w_{[0,i]})} \parallel \overline{P_{i-1}(w_{[0,i]})} \right)$$

*new distribution*      *old distribution*

# Results on local-coherence sentences

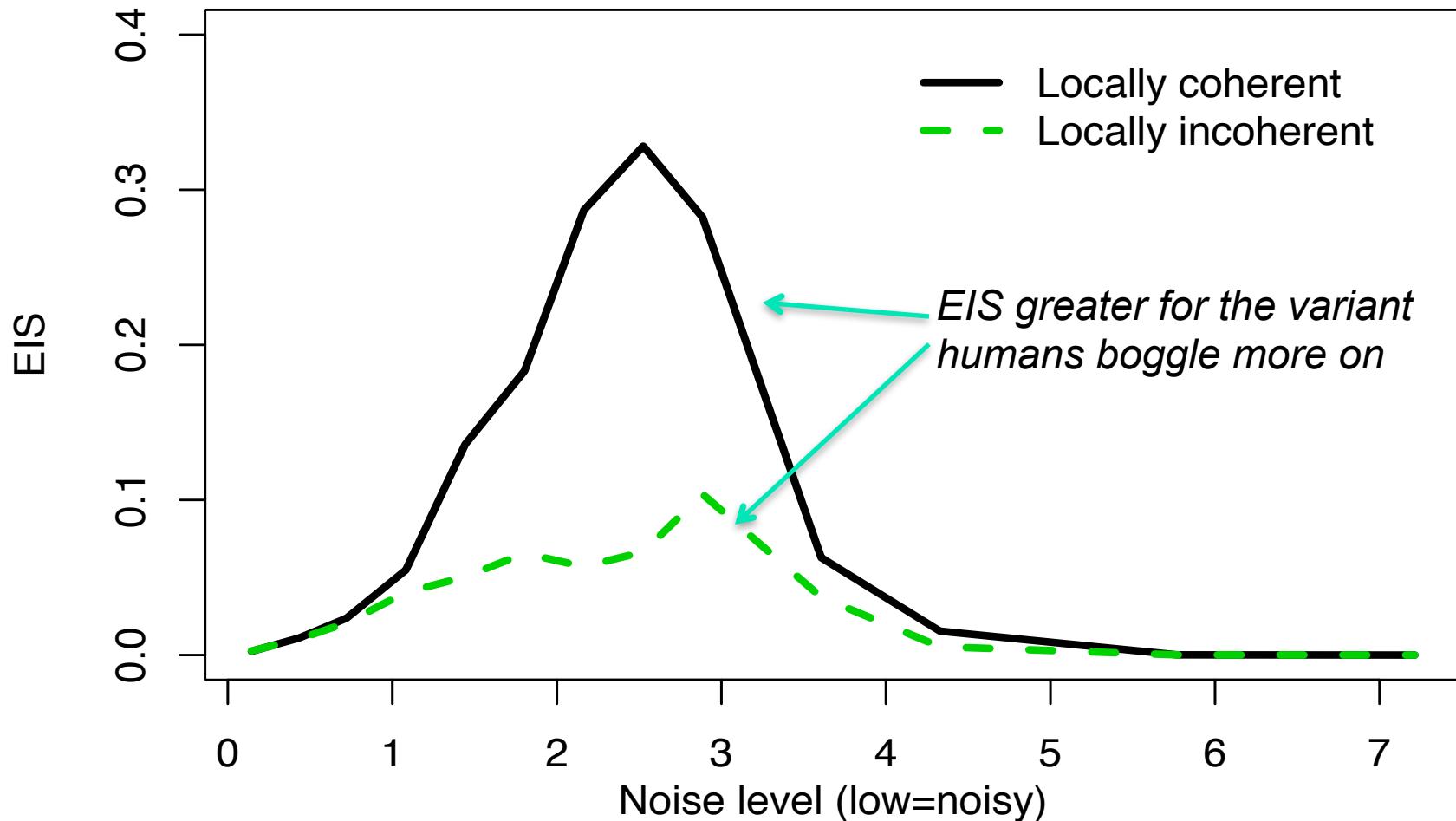
- Locally coherent: *The coach smiled at the player tossed the frisbee*
- Locally incoherent: *The coach smiled at the player thrown the frisbee*



(All sentences of Tabor et al. 2004 with lexical coverage in model)

# Results on local-coherence sentences

- Locally coherent: *The coach smiled at the player tossed the frisbee*
- Locally incoherent: *The coach smiled at the player thrown the frisbee*



# Novel prediction: neighborhood manipulation

---

(Levy, Bicknell, Slattery, & Rayner, 2009, PNAS)

# Novel prediction: neighborhood manipulation

---

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
- Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

# Novel prediction: neighborhood manipulation

---

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
- Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

*The coach smiled at the player tossed the frisbee*

(that?) (and?) (and?)  
(who?) (as?) (that?)  
                         (who?)

# Novel prediction: neighborhood manipulation

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
  - Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

*The coach smiled **toward** the player **tossed** the frisbee*

# Novel prediction: neighborhood manipulation

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
  - Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

*The coach smiled at the player tossed the frisbee*

(that?)  
(who?)

*the player tossed the frisbee*

(and?)  
(that?)  
(who?)

*The coach smiled toward the player tossed the frisbee*

# Novel prediction: neighborhood manipulation

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
  - Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

*The coach smiled at the player tossed the frisbee*

(that?)  
(who?)  (and?)  
(that?)  
(who?)

*The coach smiled **toward** the player **tossed** the frisbee*

- Substituting *toward* for *at* should reduce the EIS

# Novel prediction: neighborhood manipulation

- Uncertain-input effects should be *dependent on the perceptual neighborhood* of the sentence
  - Resulting novel prediction: changing neighborhood of the context can affect EIS & thus comprehension behavior

*The coach smiled at the player* **tossed** *the frisbee*

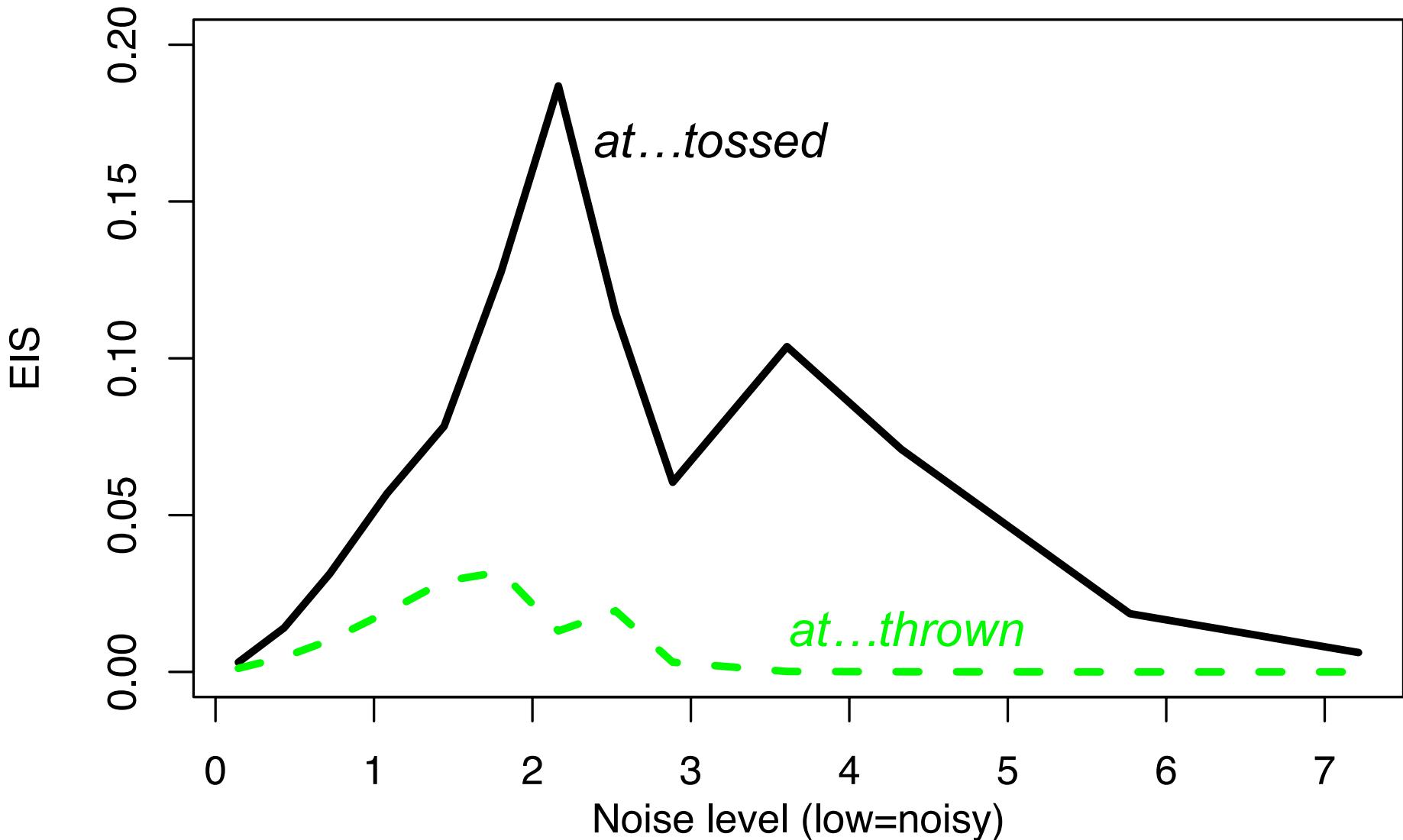
(that?)  
(who?)  (and?)  
(that?)  
(who?)

*The coach smiled toward the player tossed the frisbee*

- Substituting *toward* for *at* should reduce the EIS
  - In free reading, we should see less tendency to regress from *tossed* when the EIS is small

# Model predictions

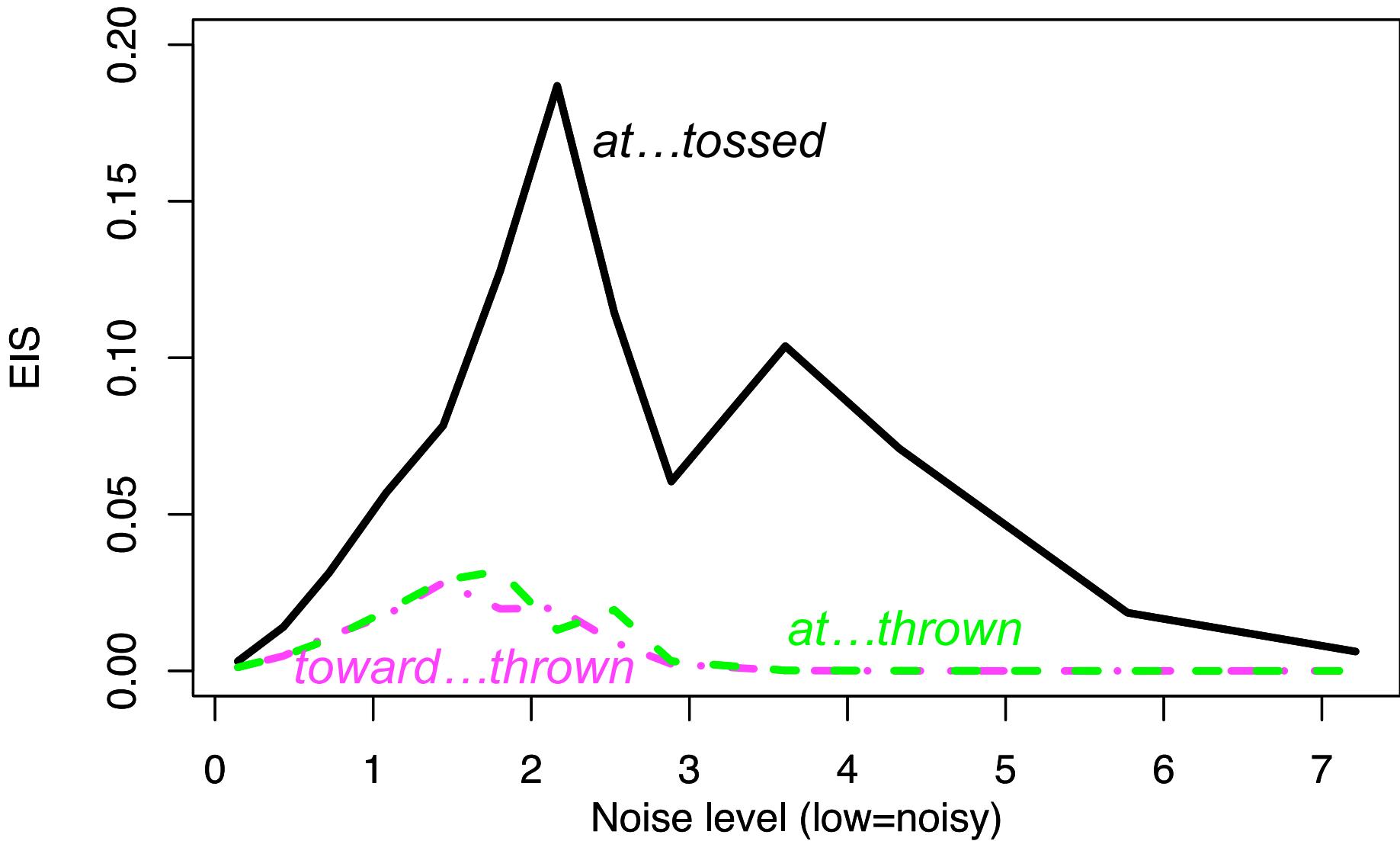
---



*(The coach smiled at/toward the player tossed/thrown the frisbee)*

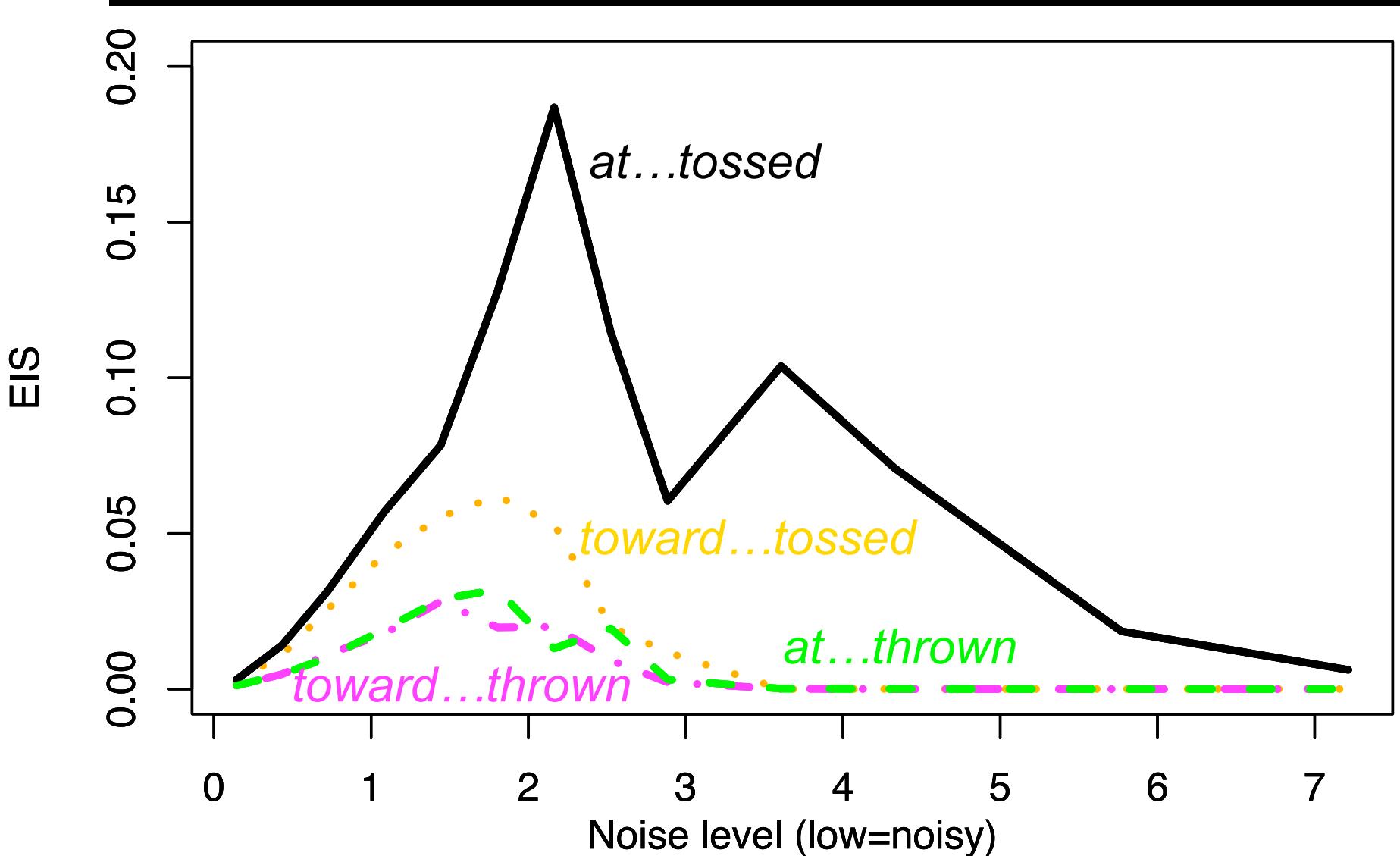
# Model predictions

---



(The coach smiled at/toward the player tossed/thrown the frisbee)

# Model predictions



(The coach smiled at/toward the player tossed/thrown the frisbee)

# Experimental design

---

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player **tossed** the frisbee  
The coach smiled at the player **thrown** the frisbee  
The coach smiled **toward** the player **tossed** the frisbee  
The coach smiled **toward** the player **thrown** the frisbee

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player **tossed** the frisbee  
The coach smiled at the player **thrown** the frisbee  
The coach smiled **toward** the player **tossed** the frisbee  
The coach smiled **toward** the player **thrown** the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player

*tossed*

the frisbee

The coach smiled at the player

*thrown*

the frisbee

The coach smiled toward the player

*tossed*

the frisbee

The coach smiled toward the player

*thrown*

the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player **tossed** the frisbee  
The coach smiled at the player **thrown** the frisbee  
The coach smiled **toward** the player **tossed** the frisbee  
The coach smiled **toward** the player **thrown** the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region **tossed/thrown**

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player **tossed** the frisbee  
The coach smiled at the player **thrown** the frisbee  
The coach smiled **toward** the player **tossed** the frisbee  
The coach smiled **toward** the player **thrown** the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region **tossed/thrown**
  - First-pass regressions out of critical region

# Experimental design

---

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player **tossed** the frisbee  
The coach smiled at the player **thrown** the frisbee  
The coach smiled **toward** the player **tossed** the frisbee  
The coach smiled **toward** the player **thrown** the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region **tossed/thrown**
  - First-pass regressions out of critical region
  - Go-past time for critical region

# Experimental design

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

The coach smiled at the player  
The coach smiled at the player  
The coach smiled toward the player  
The coach smiled toward the player

The frisbee  
the frisbee  
the frisbee  
the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region *tossed/thrown*
  - First-pass regressions out of critical region
  - Go-past time for critical region

# Experimental design

- In a free-reading eye-tracking study, we crossed *at/toward* with *tossed/thrown*:

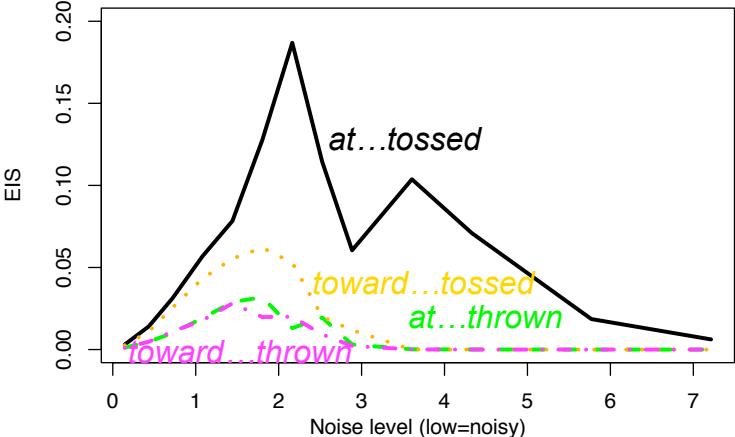
The coach smiled at the player  
The coach smiled at the player  
The coach smiled toward the player  
The coach smiled toward the player

The frisbee  
the frisbee  
the frisbee  
the frisbee

- Prediction: interaction between preposition & ambiguity in some subset of:
  - Early-measure RTs at critical region *tossed/thrown*
  - First-pass regressions out of critical region
  - Go-past time for critical region
  - Regressions into *at/toward*

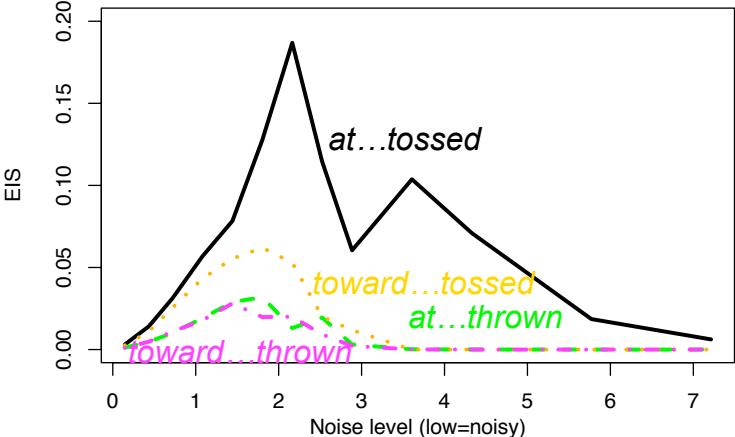
# Experimental results

*The coach smiled at the player tossed...*



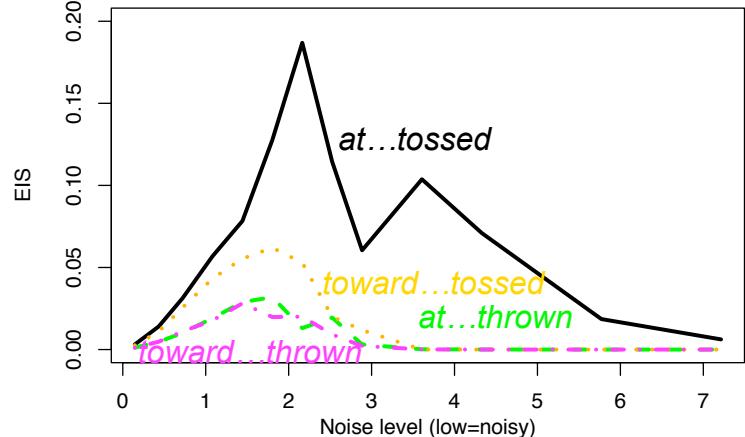
# Experimental results

*The coach smiled at the player tossed...*



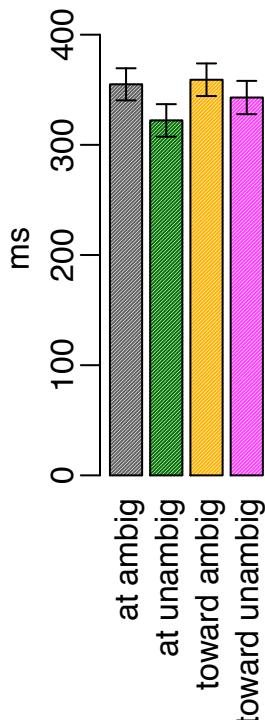
# Experimental results

*The coach smiled at the player tossed...*

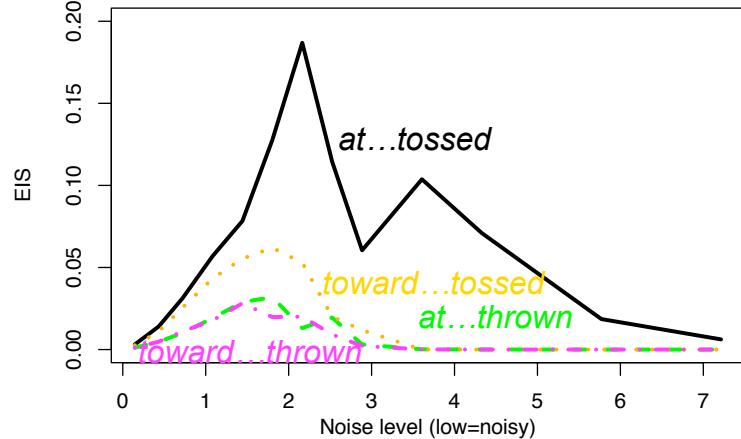


# Experimental results

*The coach smiled at the player tossed...*

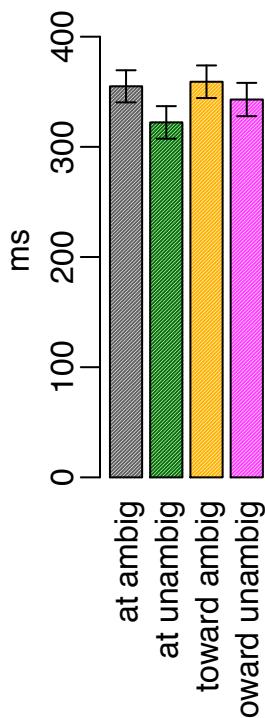


*First-pass  
RT*

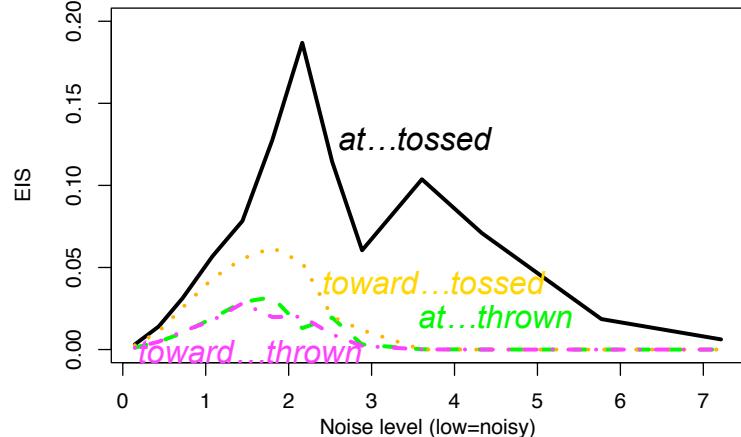


# Experimental results

*The coach smiled at the player tossed...*

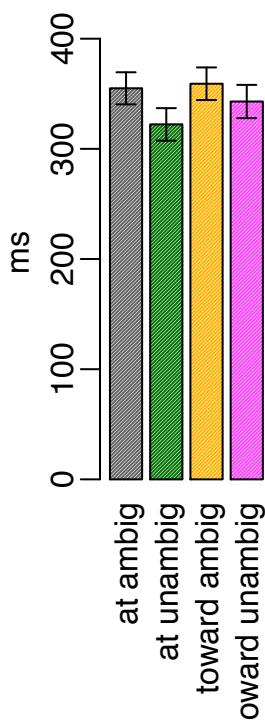


*First-pass  
RT*

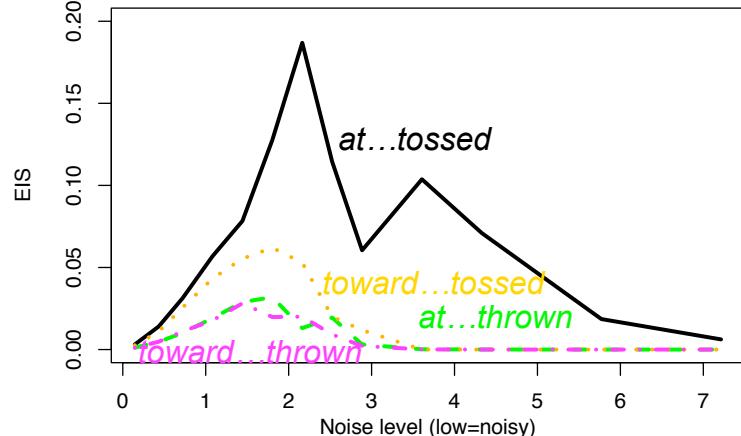


# Experimental results

*The coach smiled at the player tossed...*

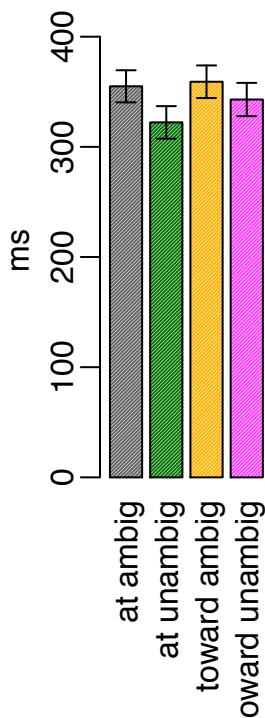


*First-pass  
RT*

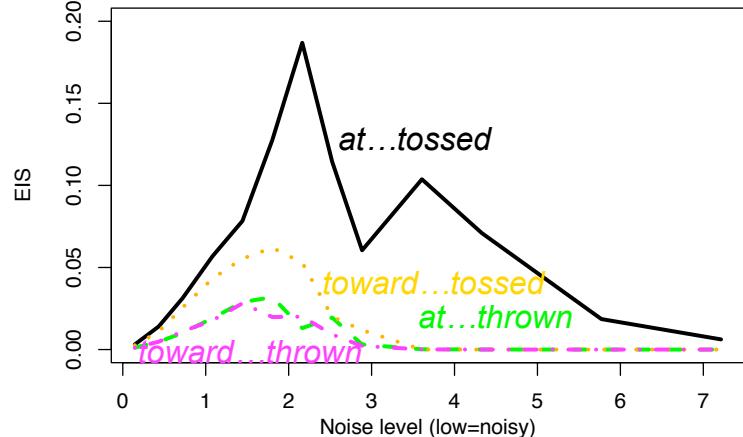


# Experimental results

*The coach smiled at the player tossed...*

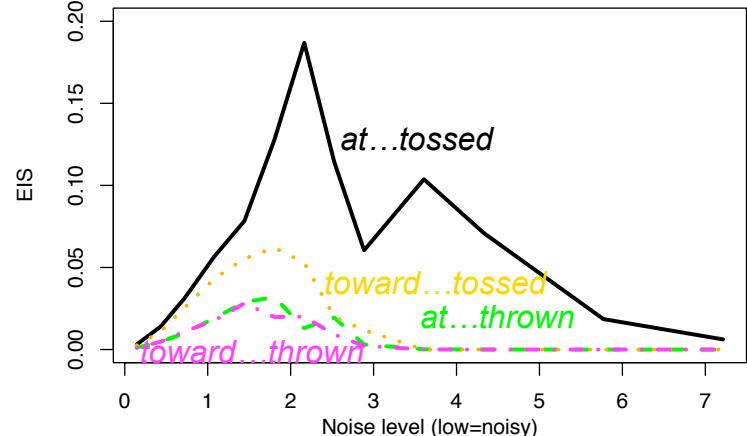
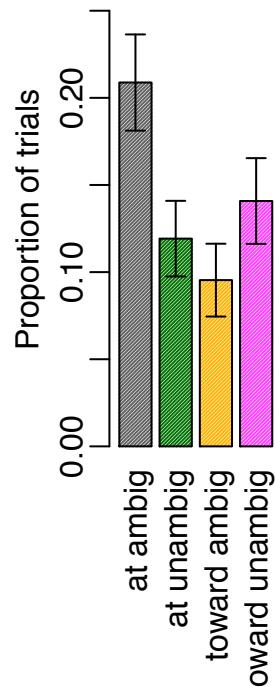
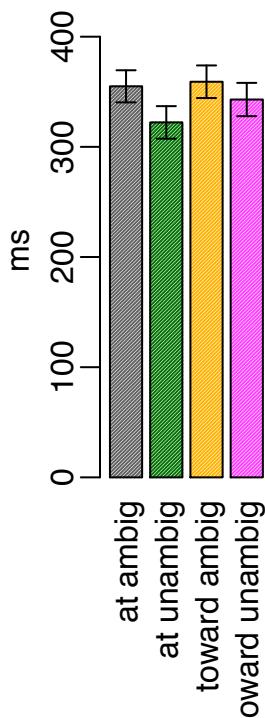


*First-pass  
RT*



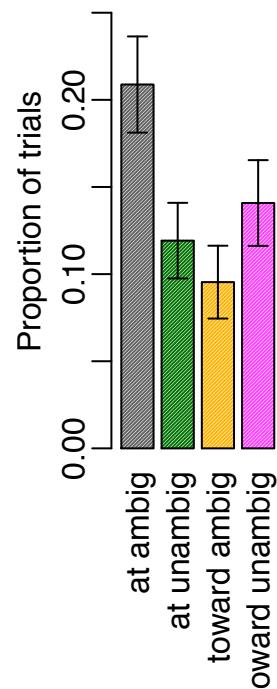
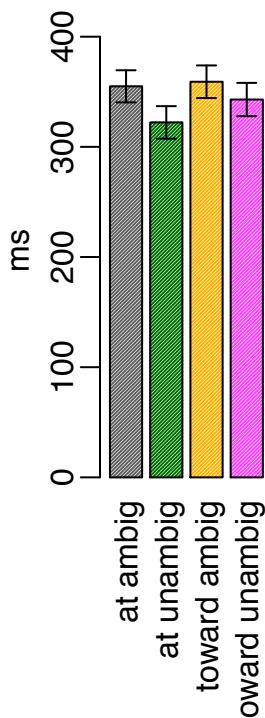
# Experimental results

*The coach smiled at the player tossed...*



# Experimental results

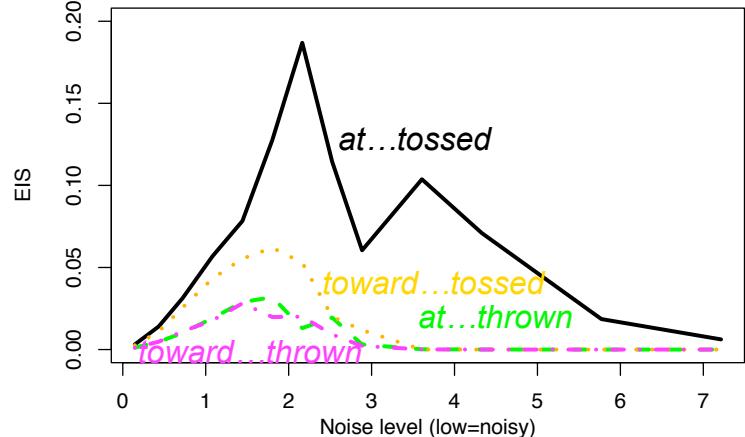
*The coach smiled at the player tossed...*



First-pass  
RT

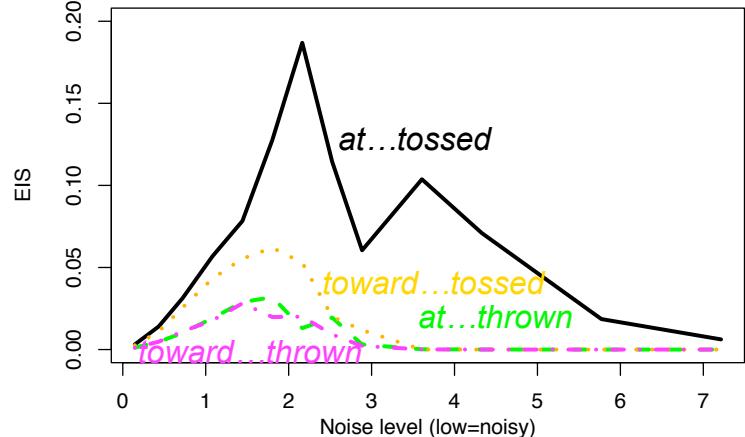
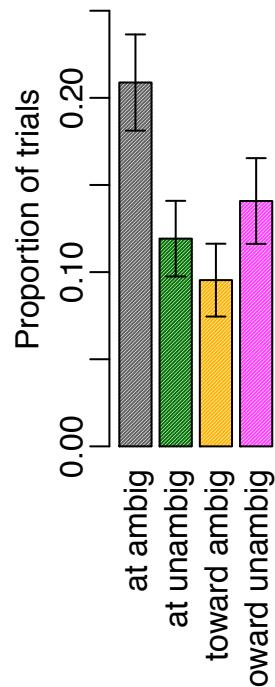
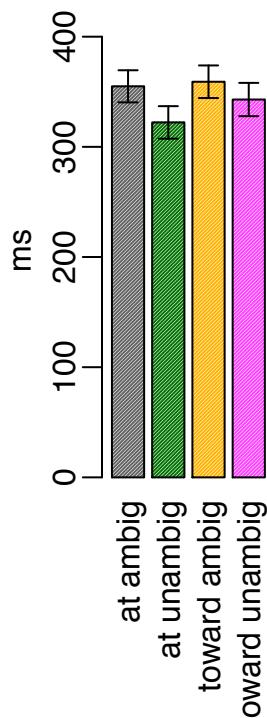


Regressions  
out



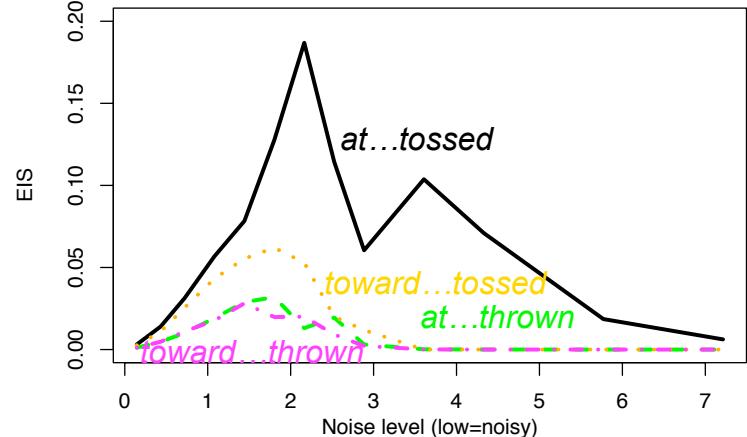
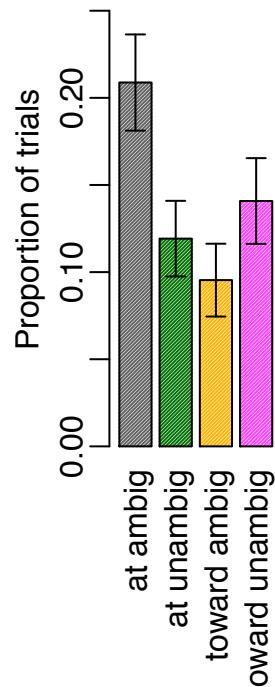
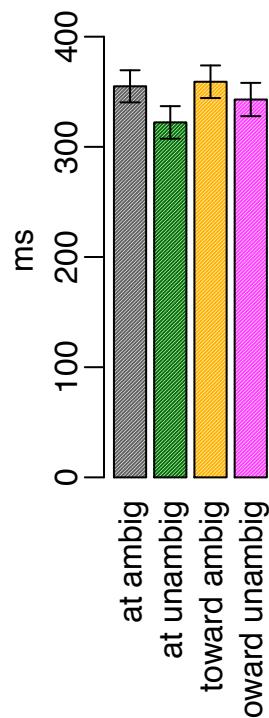
# Experimental results

*The coach smiled at the player tossed...*



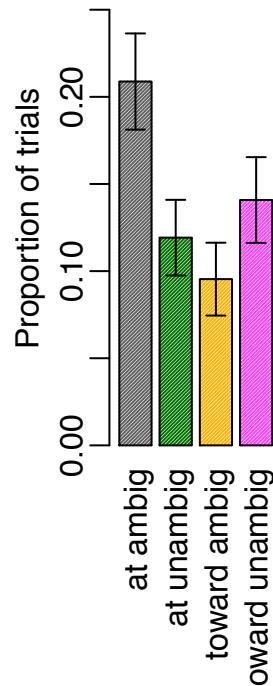
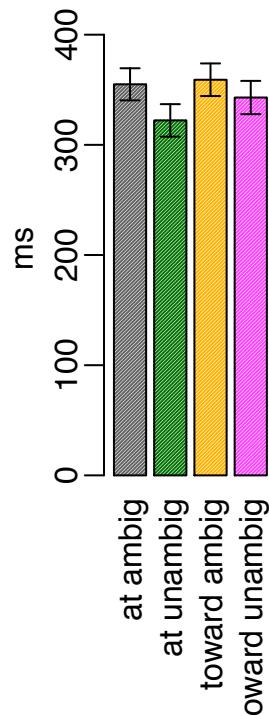
# Experimental results

*The coach smiled at the player tossed...*



# Experimental results

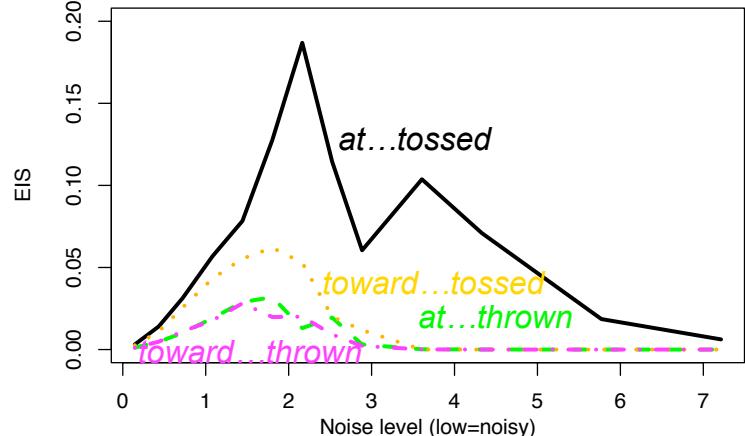
*The coach smiled at the player to see if he...  
at...tossed toward...tossed at...thrown toward...thrown*



First-pass  
RT

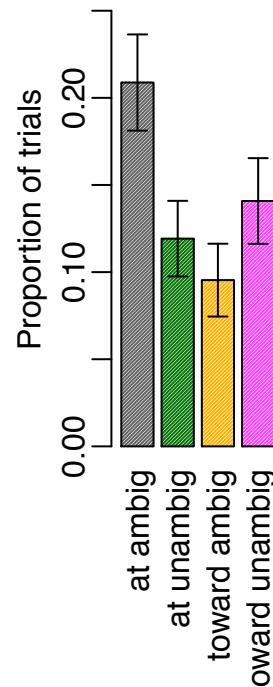
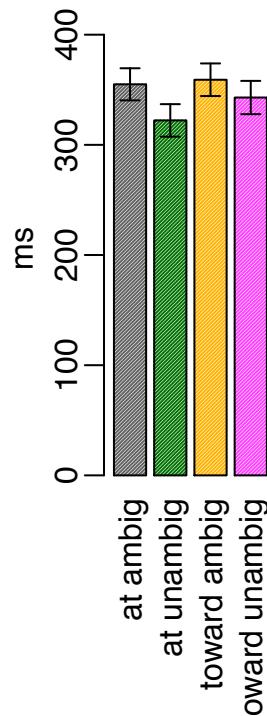


Regressions  
out



# Experimental results

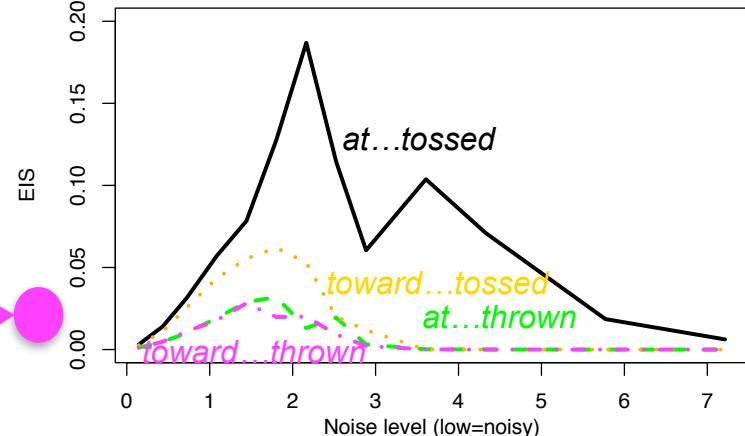
*The coach smiled at the player to score*



First-pass  
RT

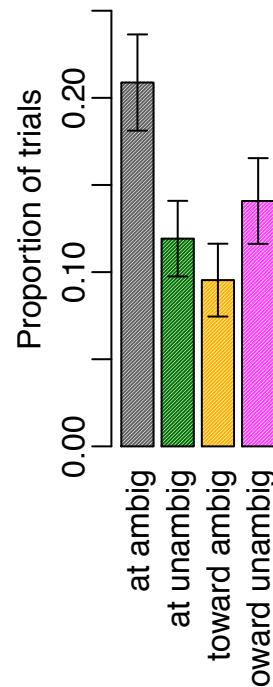
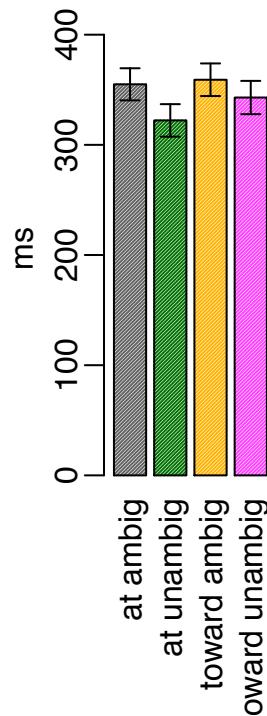


Regressions  
out



# Experimental results

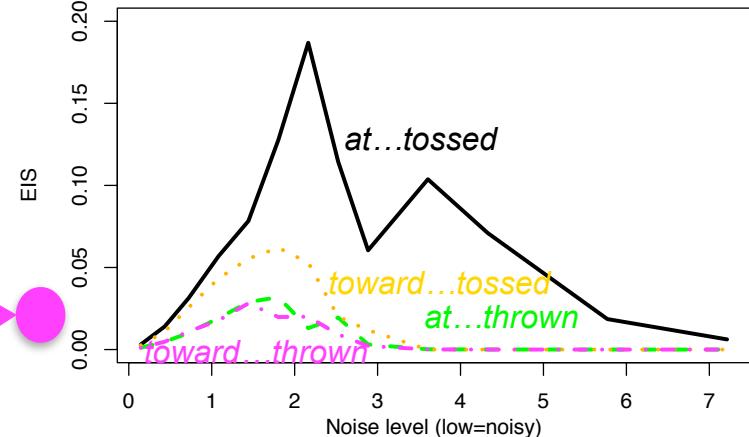
*The coach smiled at the player to score*



First-pass  
RT

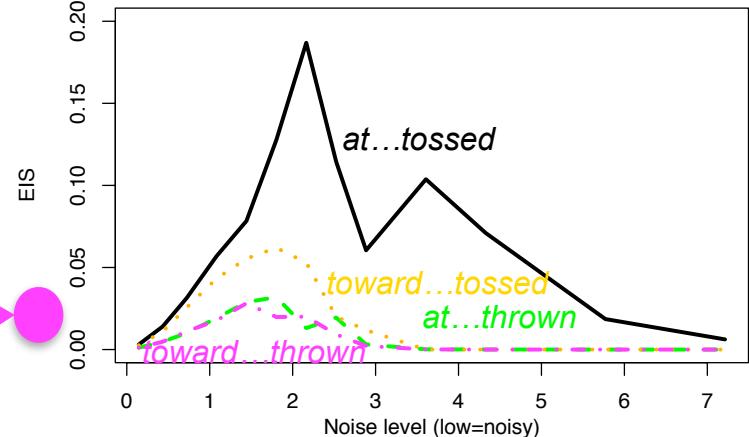
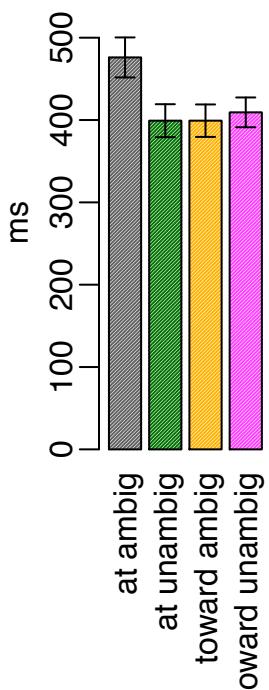
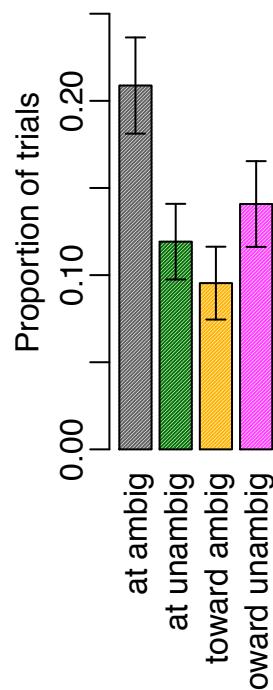
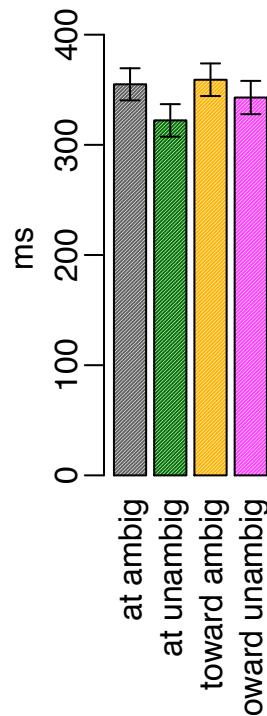


Regressions  
out



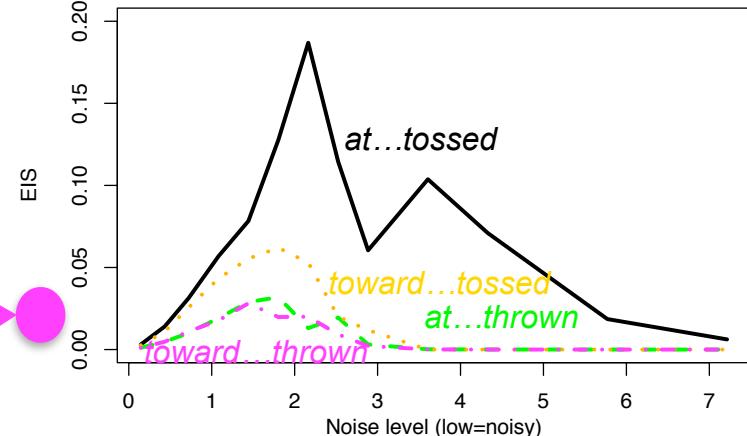
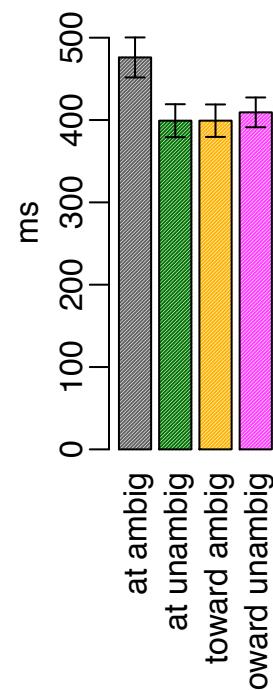
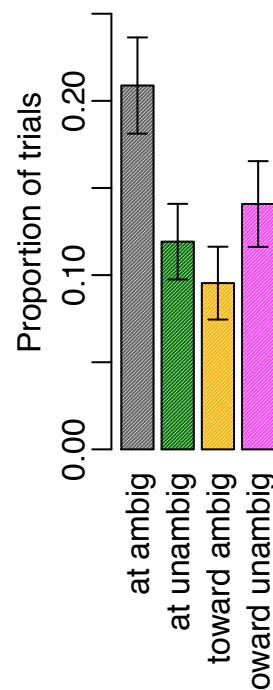
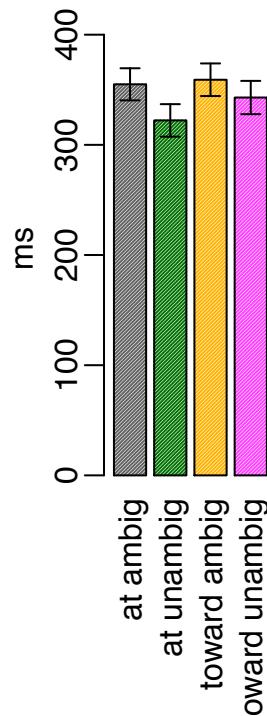
# Experimental results

*The coach smiled at the player to score*



# Experimental results

*The coach smiled at the player to score*



First-pass  
RT



Regressions  
out

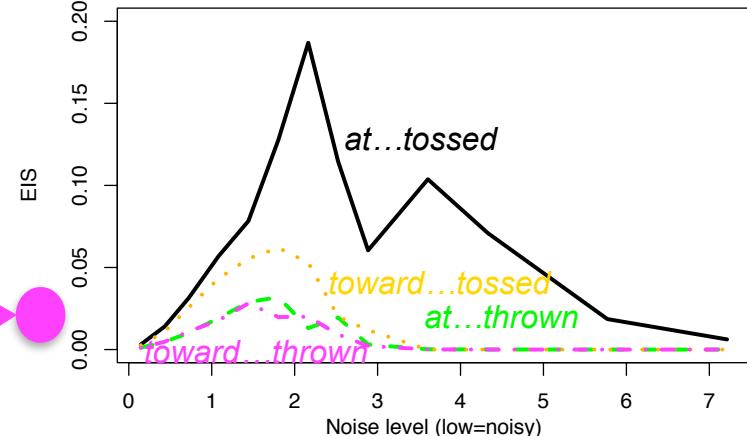
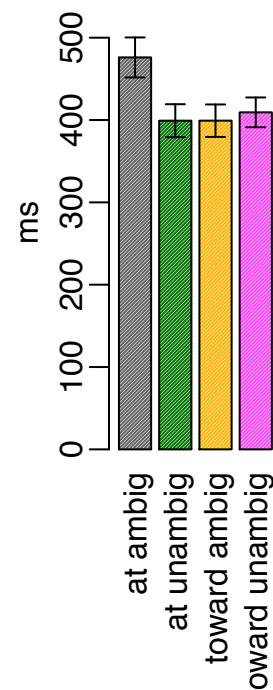
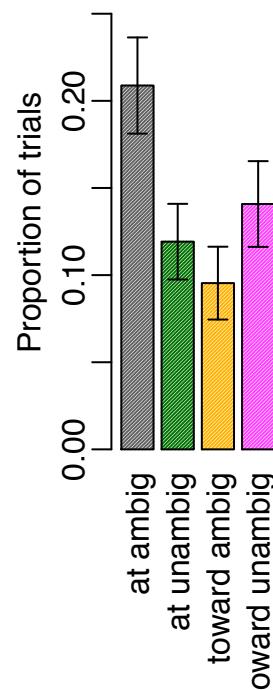
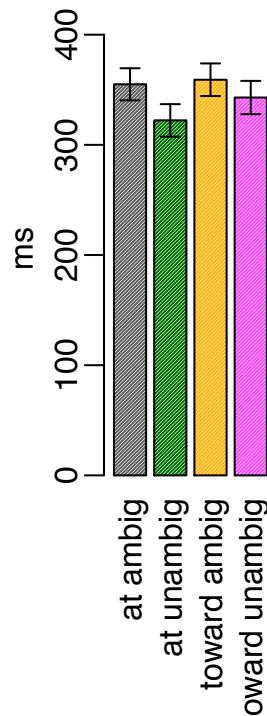


Go-past  
RT



# Experimental results

*The coach smiled at the player to...  
at...tossed toward...tossed at...thrown toward...thrown*



First-pass  
RT



Regressions  
out

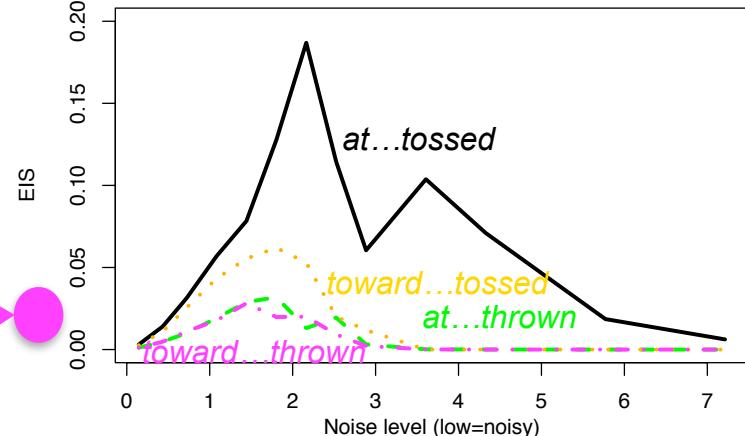
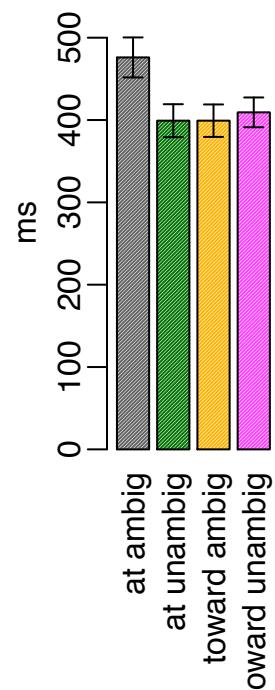
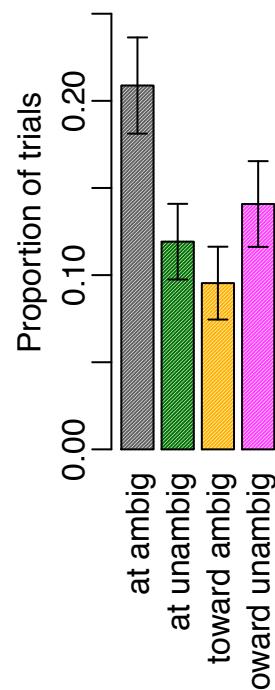
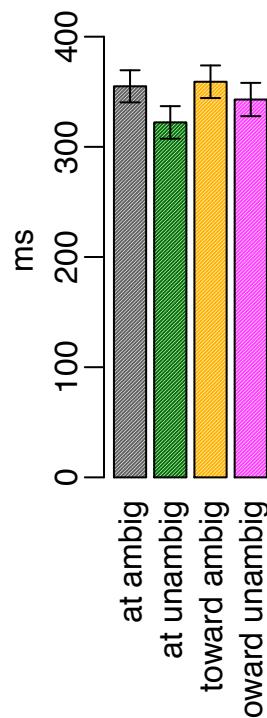


Go-past  
RT



# Experimental results

The coach smiled at the player to score?



First-pass  
RT



Regressions  
out

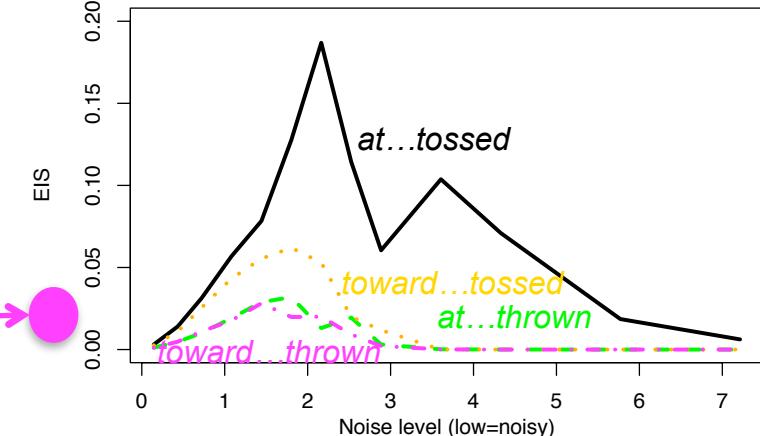
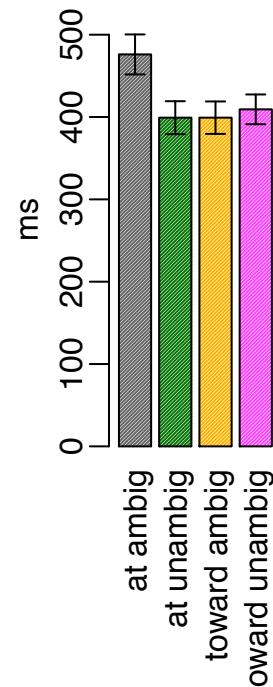
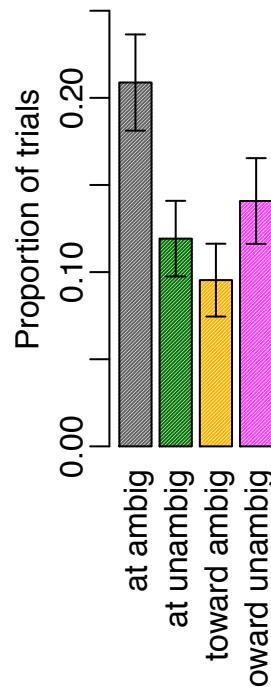
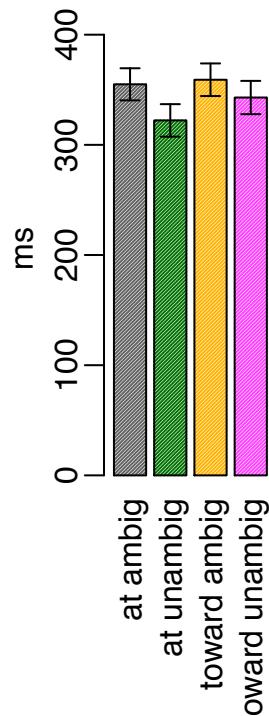


Go-past  
RT



# Experimental results

The coach smiled at the player to score?



First-pass  
RT



Regressions  
out



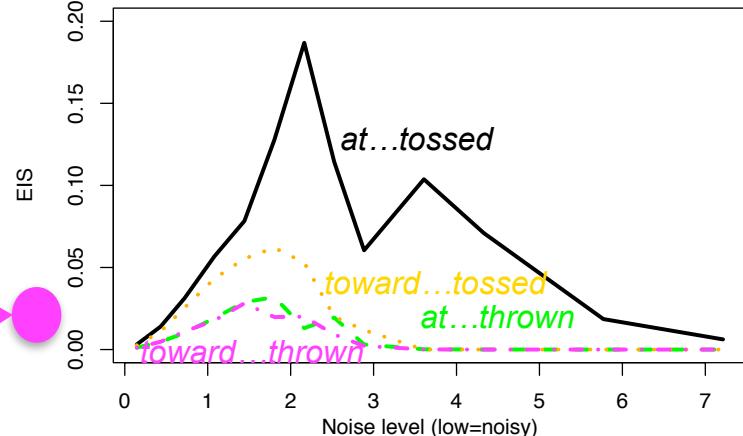
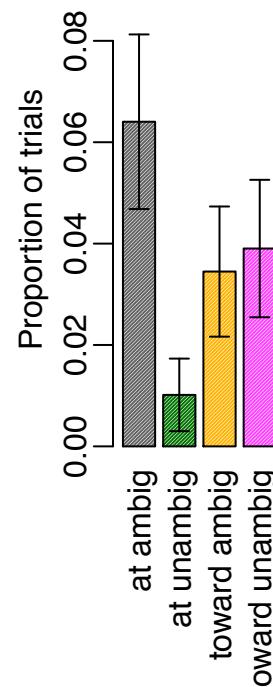
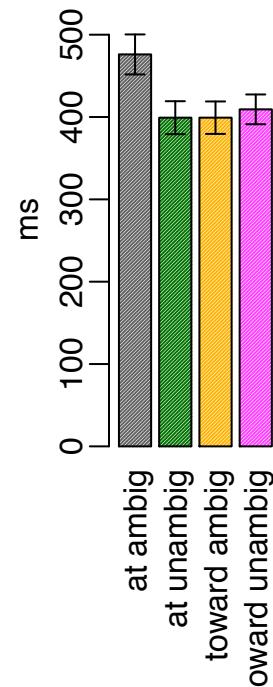
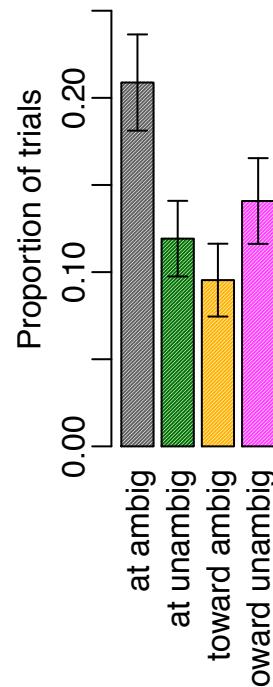
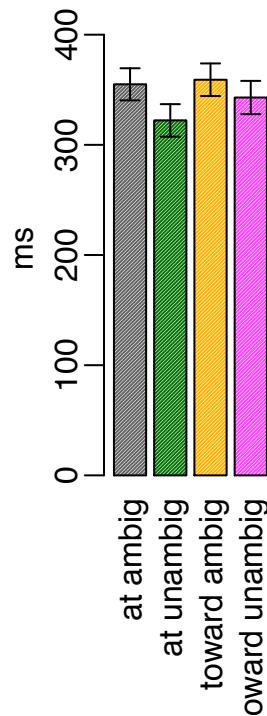
Go-past  
RT



Go-past  
regressions

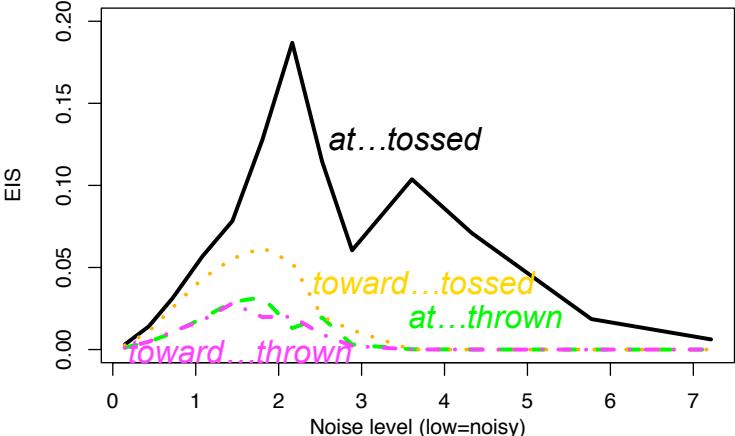
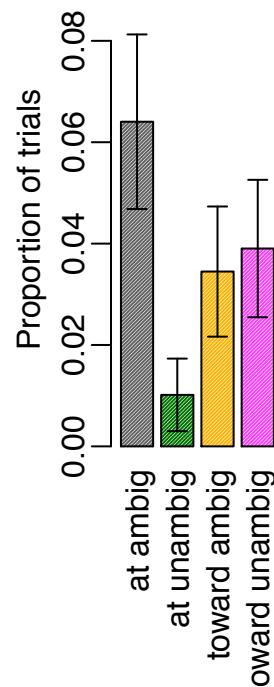
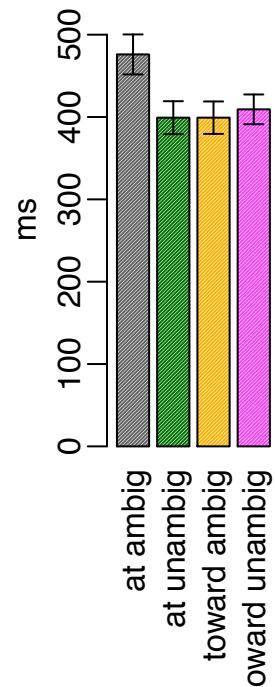
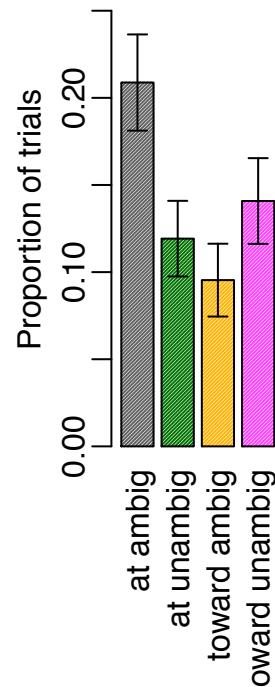
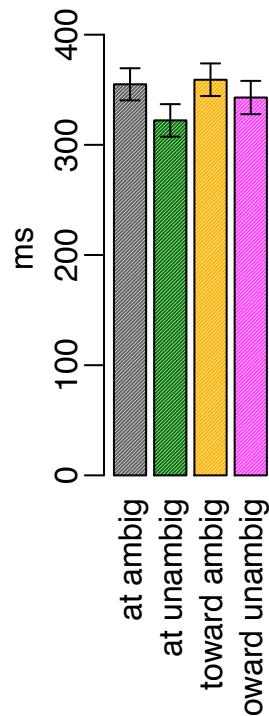
# Experimental results

The coach smiled at the player to score?



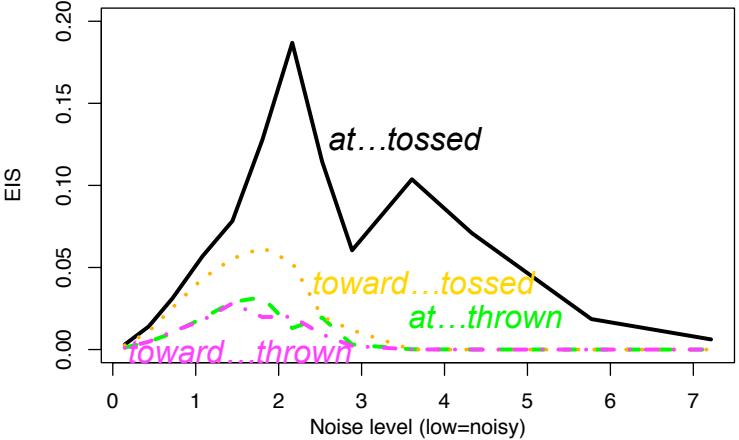
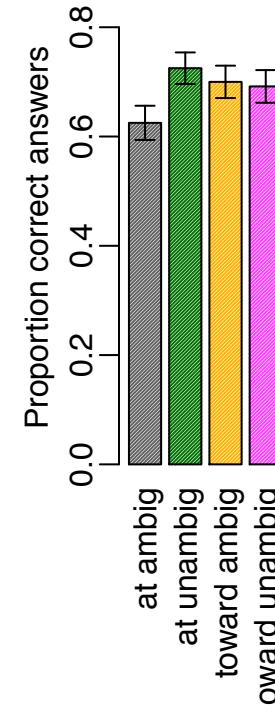
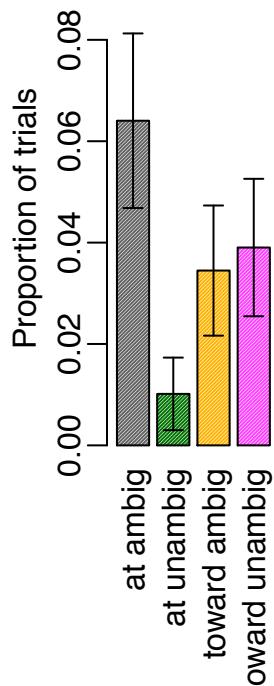
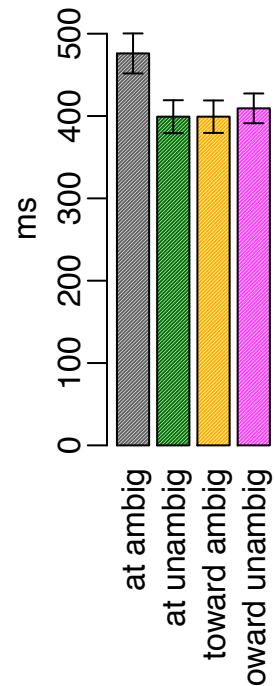
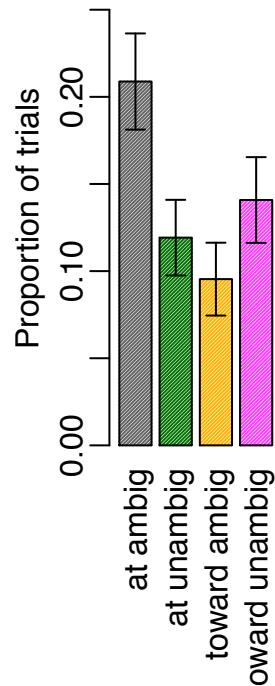
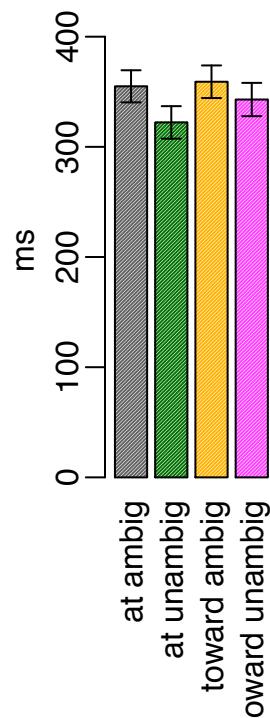
# Experimental results

*The coach smiled at the player tossed...*



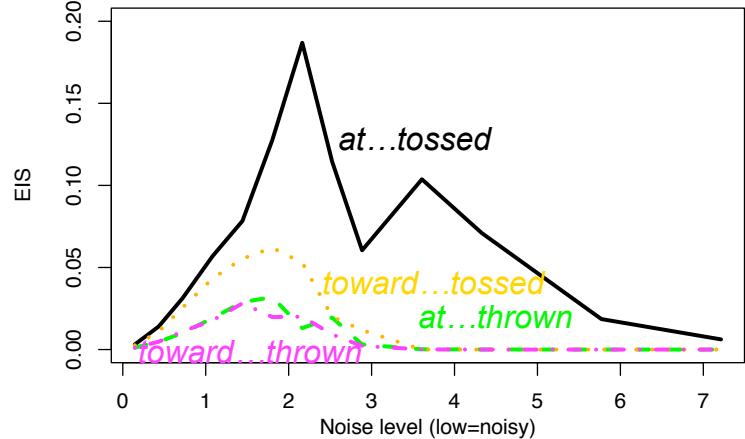
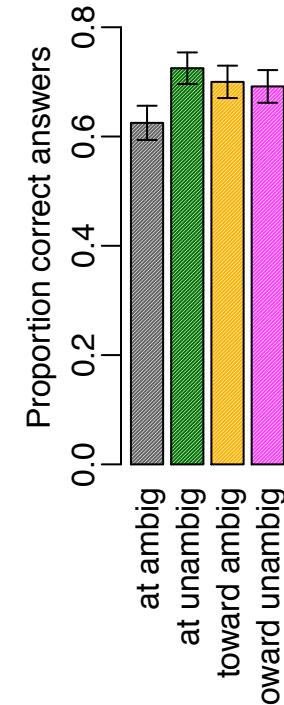
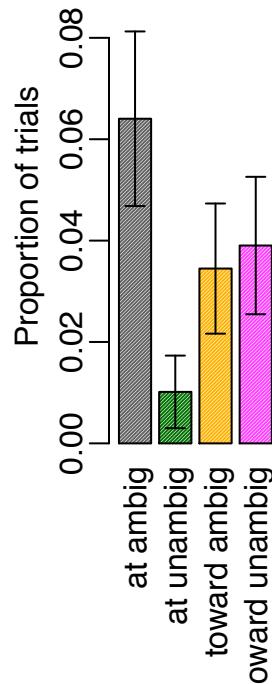
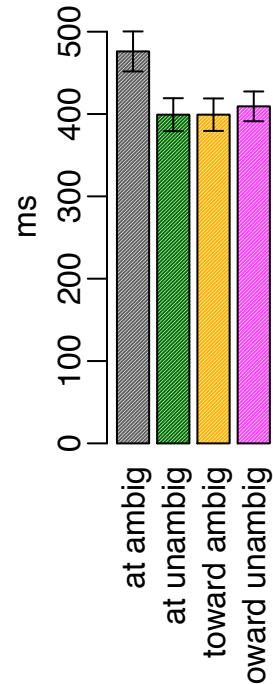
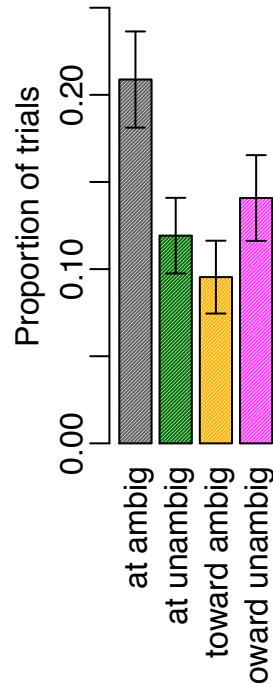
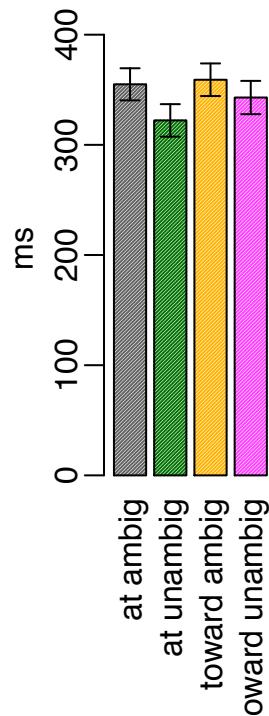
# Experimental results

*The coach smiled at the player tossed...*



# Experimental results

*The coach smiled at the player tossed...*



# Today's summary

- Reviewed principles of rational analysis and its application to theory of language comprehension
- Examined a phenomenon challenging for surprisal theory
- Proposed a noisy-channel processing theory, using information theory and probabilistic grammars
- Developed a hypothesis within the theory for the challenging phenomenon
- Empirically tested and confirmed a key prediction of the theory

# References

---

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409.
- Anderson, J. R. (1990). The adaptive character of human thought. Hillsdale, NJ: Lawrence Erlbaum.
- Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.
- Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 32–42).
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In Proceedings of the 13th conference on Empirical Methods in Natural Language Processing (pp. 234–243). Waikiki, Honolulu.
- Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086–21090.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27(4), 623–656.
- Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4), 355–370.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of Neural Information Processing Systems (pp. 5998–6008).
- Wilcox, E., Levy, R. P., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In Proceedings of the workshop on analyzing and interpreting neural networks for NLP.

# Prediction 2: hallucinated garden paths

---

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...
- ...but it's interrupted by a comma.

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...
- ...but it's interrupted by a comma.

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...
- ...but it's interrupted by a comma.

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...
- ...but it's interrupted by a comma.
- Readers are ordinarily very good at using commas to guide syntactic analysis:

*While the man hunted, the deer ran into the woods*

*While Mary was mending the sock fell off her lap*

- “With a comma after *mending* there would be no syntactic garden path left to be studied.” (Fodor, 2002)

# Prediction 2: hallucinated garden paths

---

- Try reading the sentence below:

While the clouds crackled, above the glider soared a magnificent eagle.

- There's a garden-path clause in this sentence...
- ...but it's interrupted by a comma.
- Readers are ordinarily very good at using commas to guide syntactic analysis:

*While the man hunted, the deer ran into the woods*

*While Mary was mending the sock fell off her lap*

- “With a comma after *mending* there would be no syntactic garden path left to be studied.” (Fodor, 2002)
- We'll see that the story is slightly more complicated.

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.

- This sentence is comprised of an initial intransitive subordinate clause...

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.

- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.



- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.



- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.



- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)
- Crucially, the main clause's initial PP would make a great dependent of the subordinate verb...

# Prediction 2: hallucinated garden paths

---

While the clouds crackled, above the glider soared a magnificent eagle.



- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)
- Crucially, the main clause's initial PP would make a great dependent of the subordinate verb...
- ...but doing that *would require the comma to be ignored*.

# Prediction 2: hallucinated garden paths

---

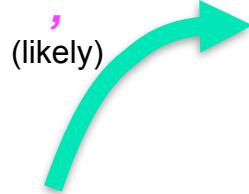
While the clouds crackled, above the glider soared a magnificent eagle.



- This sentence is comprised of an initial intransitive subordinate clause...
- ...and then a main clause with *locative inversion*.  
(c.f. a magnificent eagle soared above the glider)
- Crucially, the main clause's initial PP would make a great dependent of the subordinate verb...
- ...but doing that *would require the comma to be ignored*.
- Inferences through ...*glider* should thus involve a tradeoff between perceptual input and prior expectations

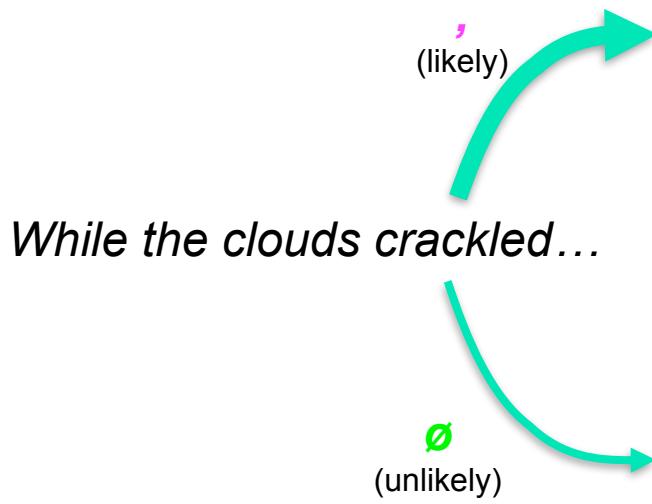
*While the clouds crackled...*

- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma

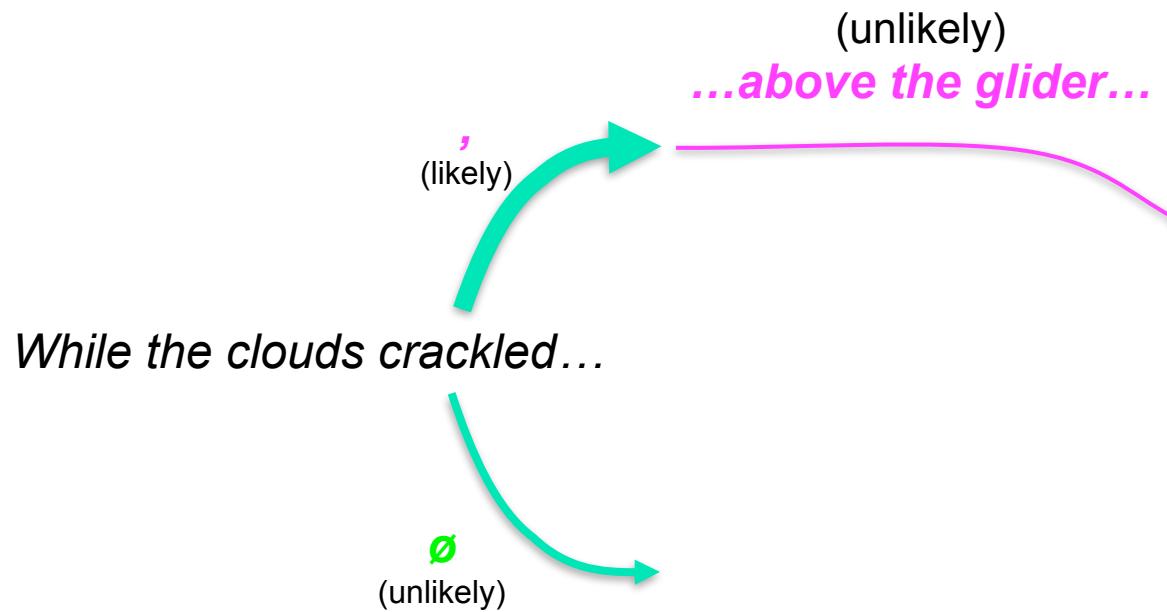


*While the clouds crackled...*

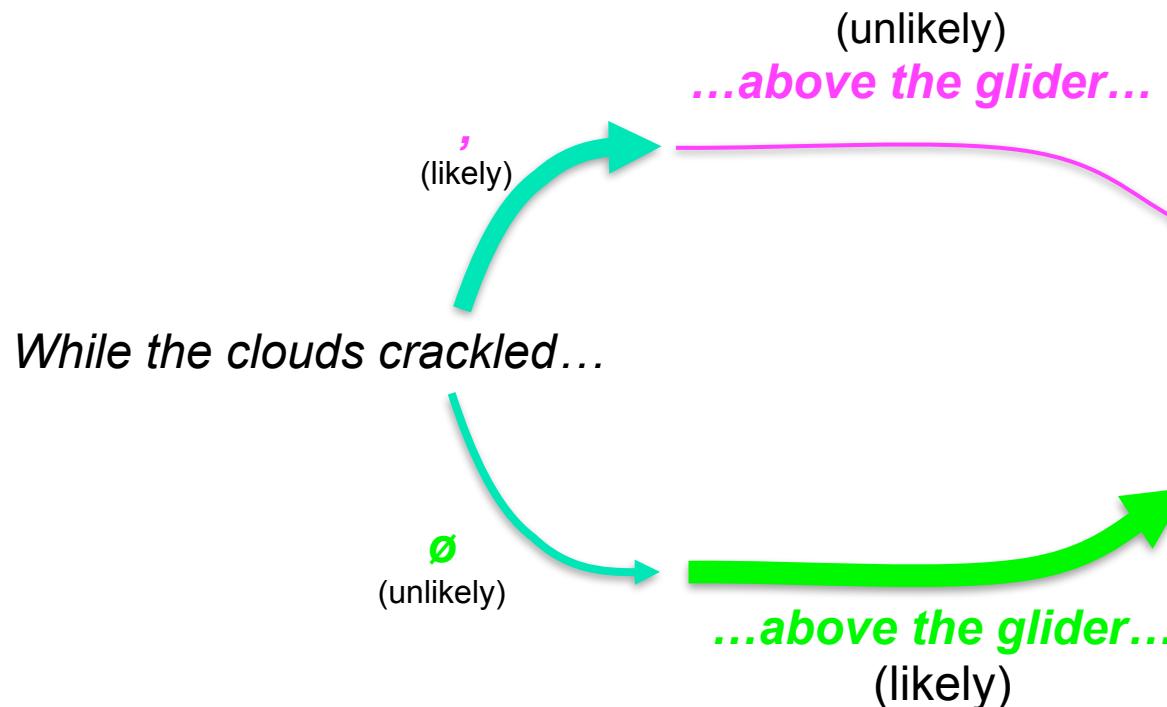
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma



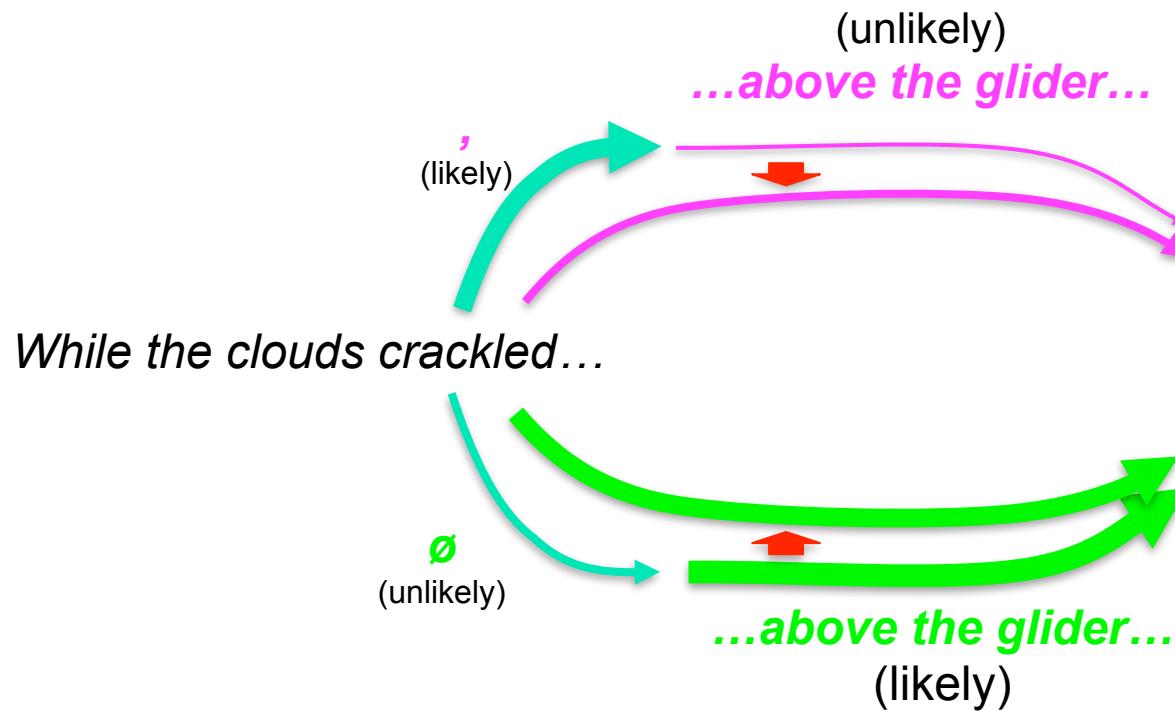
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma



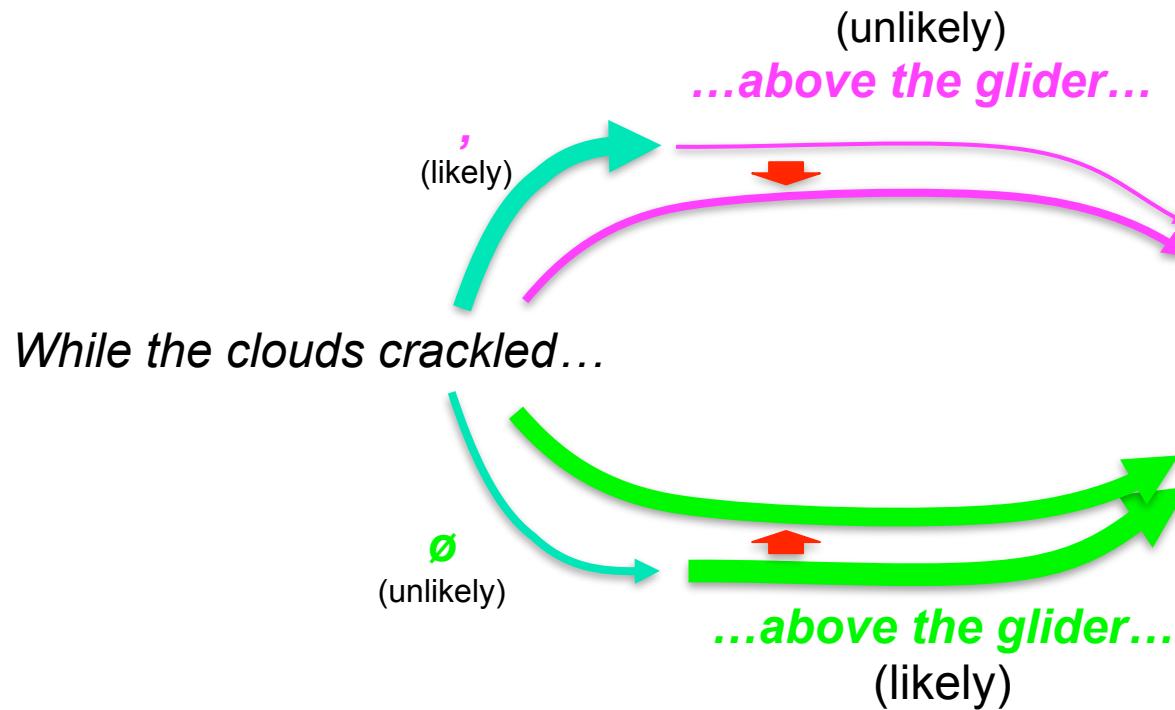
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma



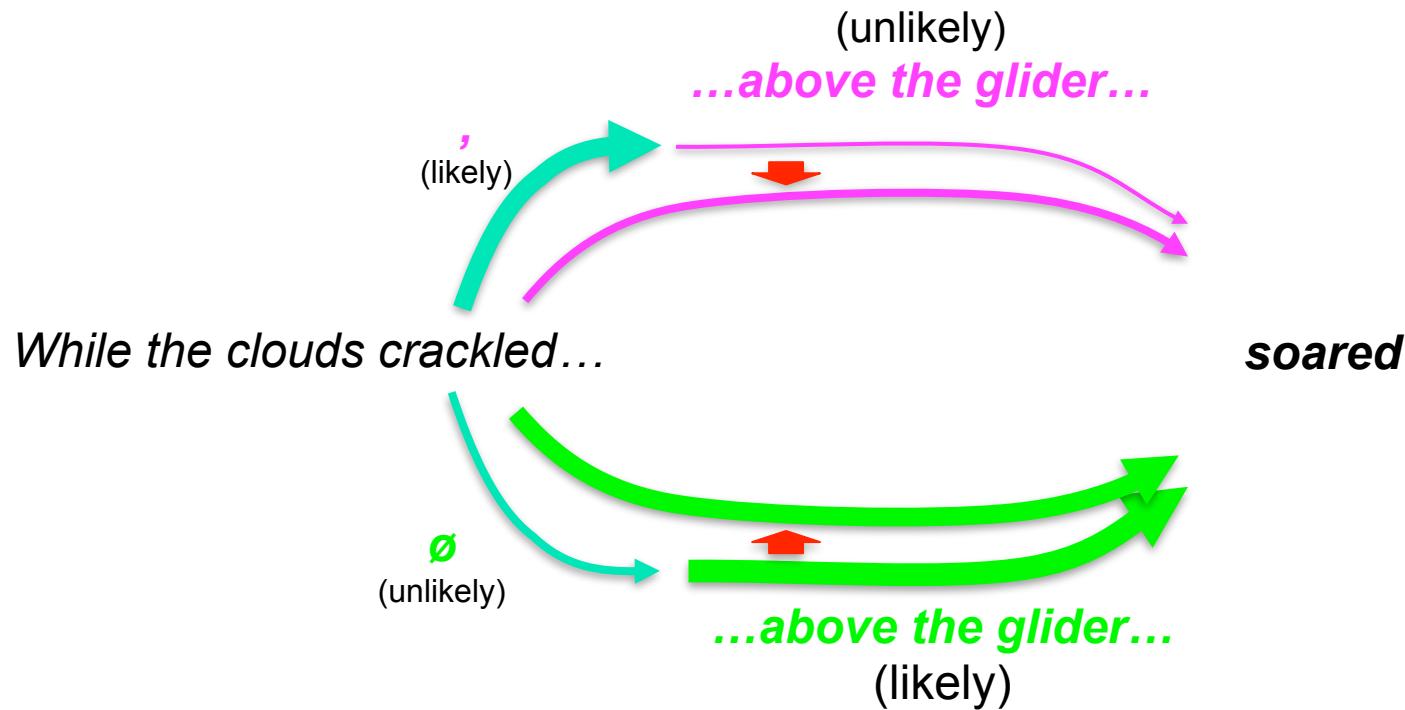
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma



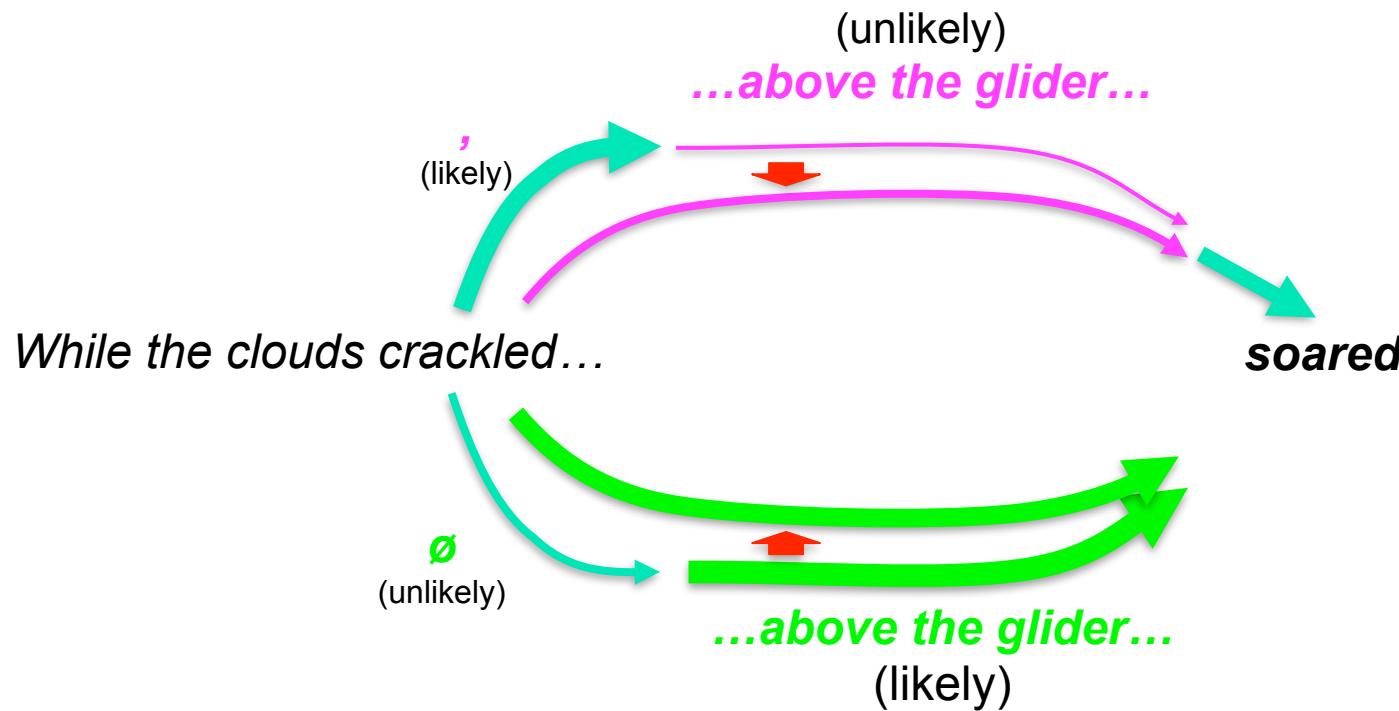
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma



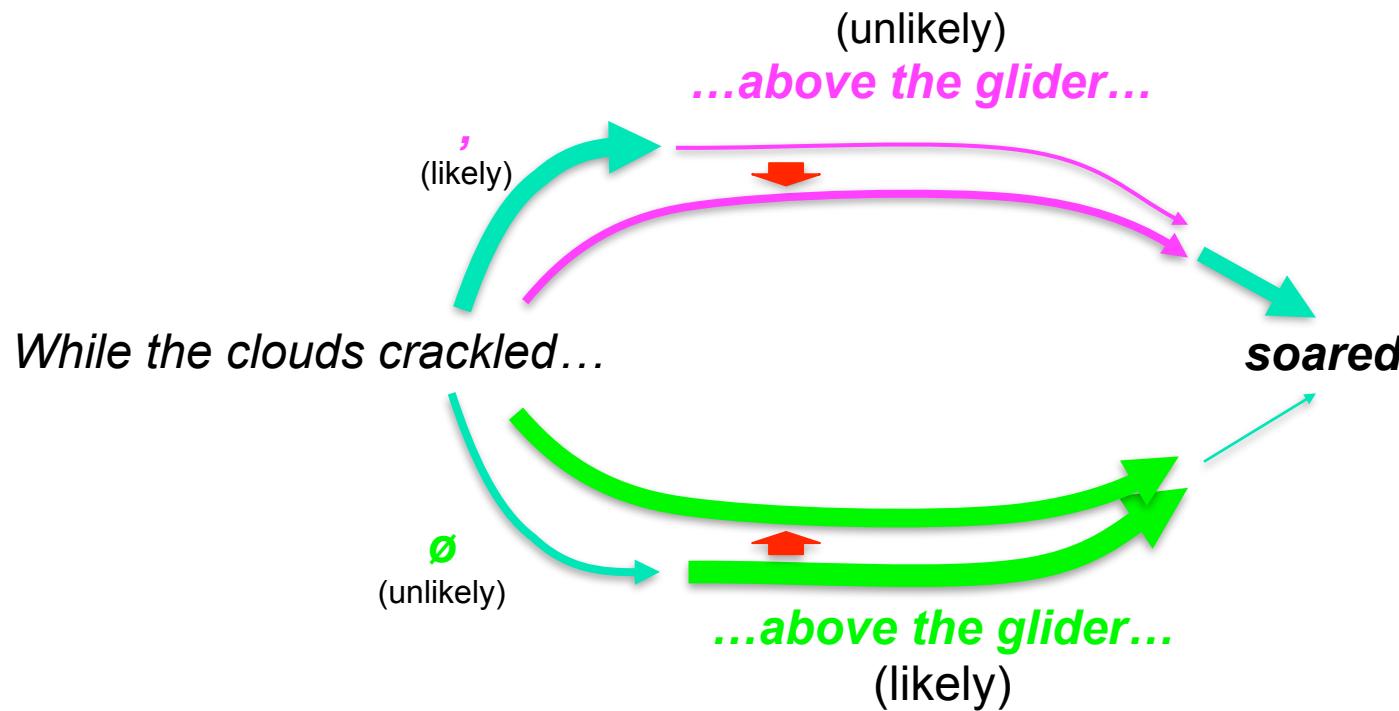
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!



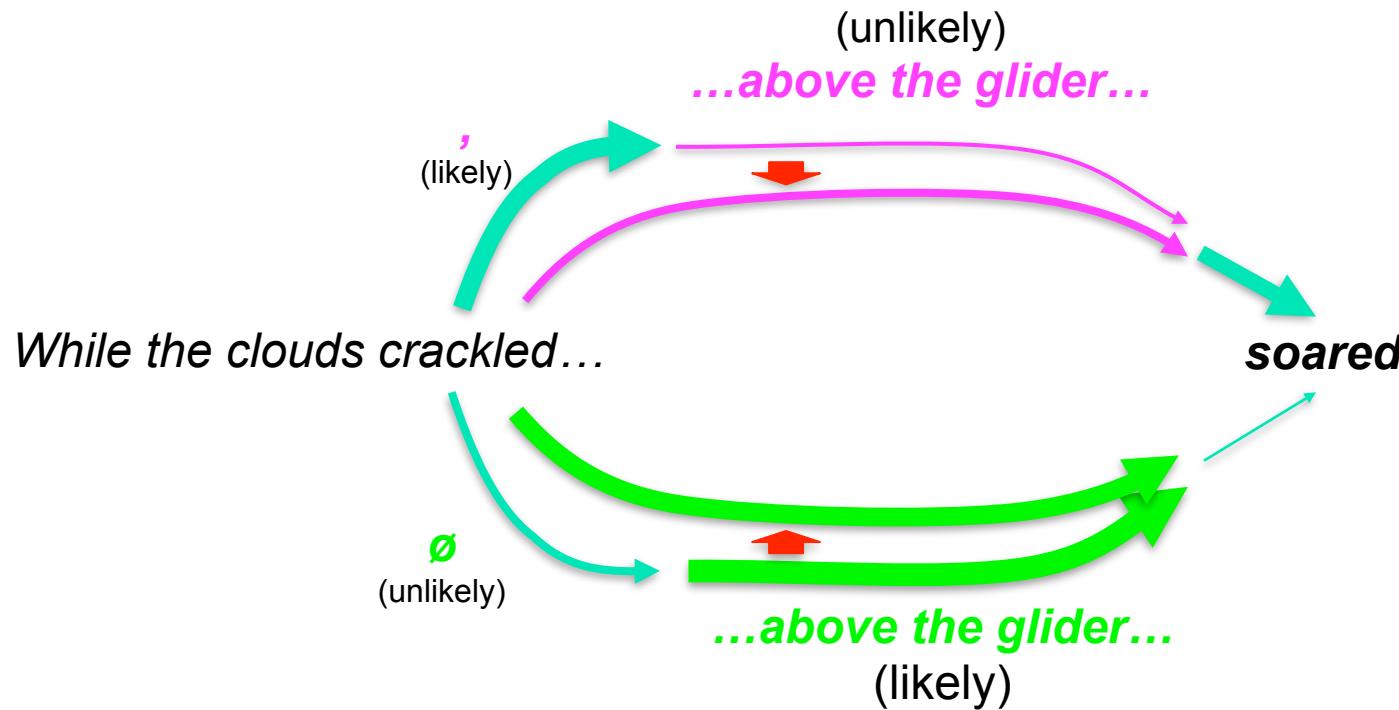
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!



- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

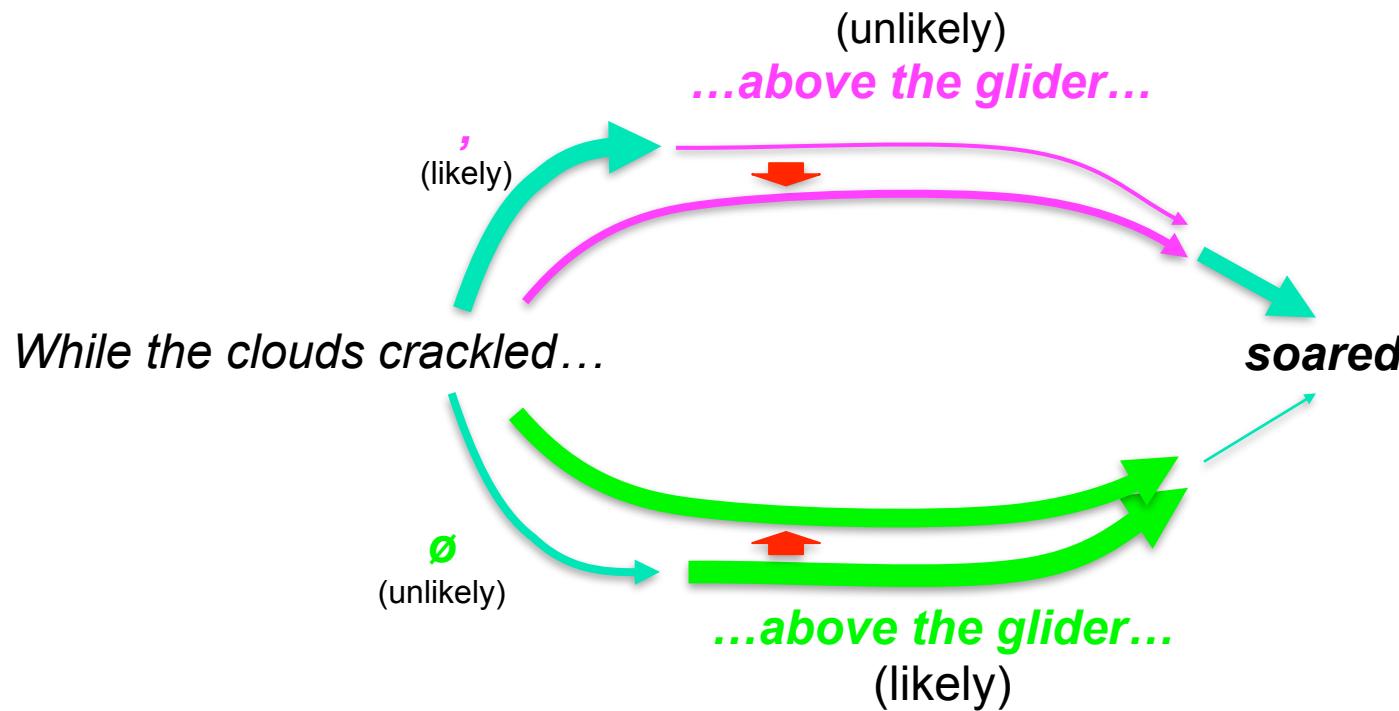


- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!



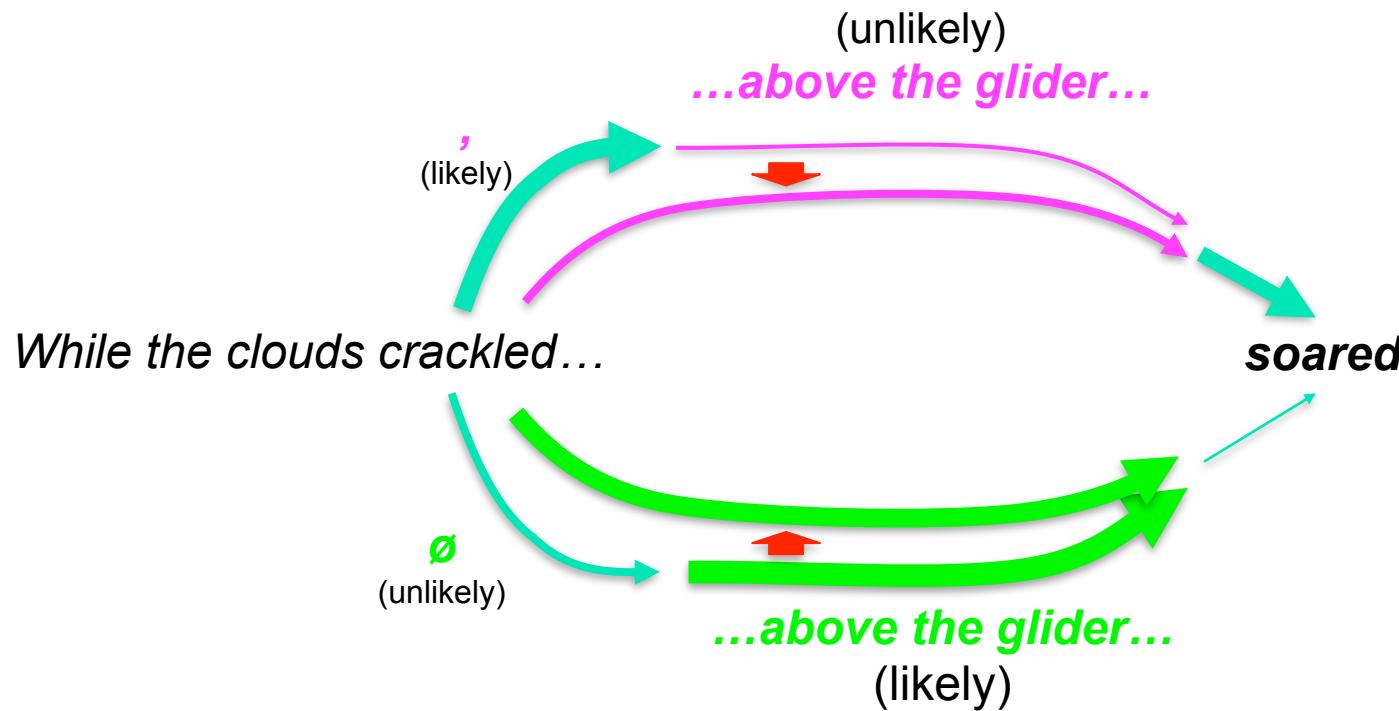
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

$$P(w_i|\text{Context}) = \sum_{\text{Path}} P(w_i|\text{Path, Context})P(\text{Path}|\text{Context})$$



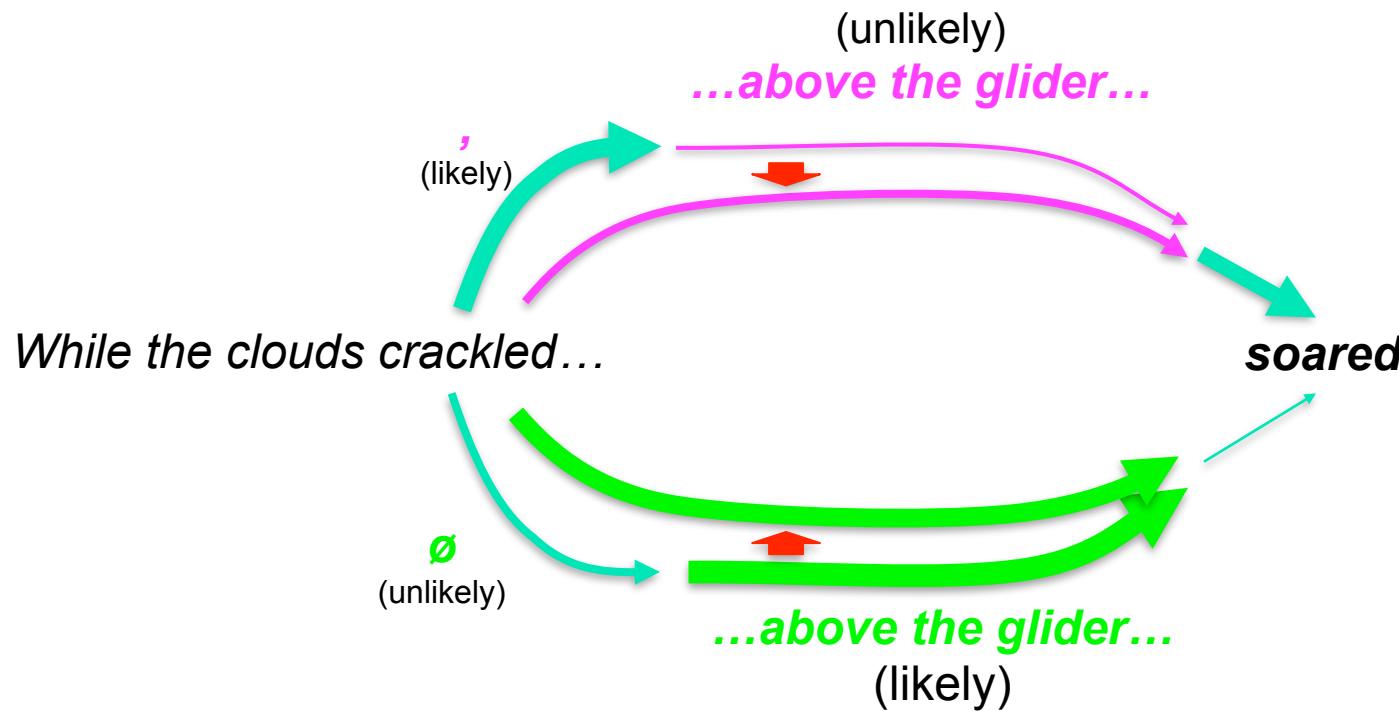
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

$$P(w_i|\text{Context}) = \sum_{\text{Path}} P(w_i|\text{Path, Context})P(\text{Path}|\text{Context})$$



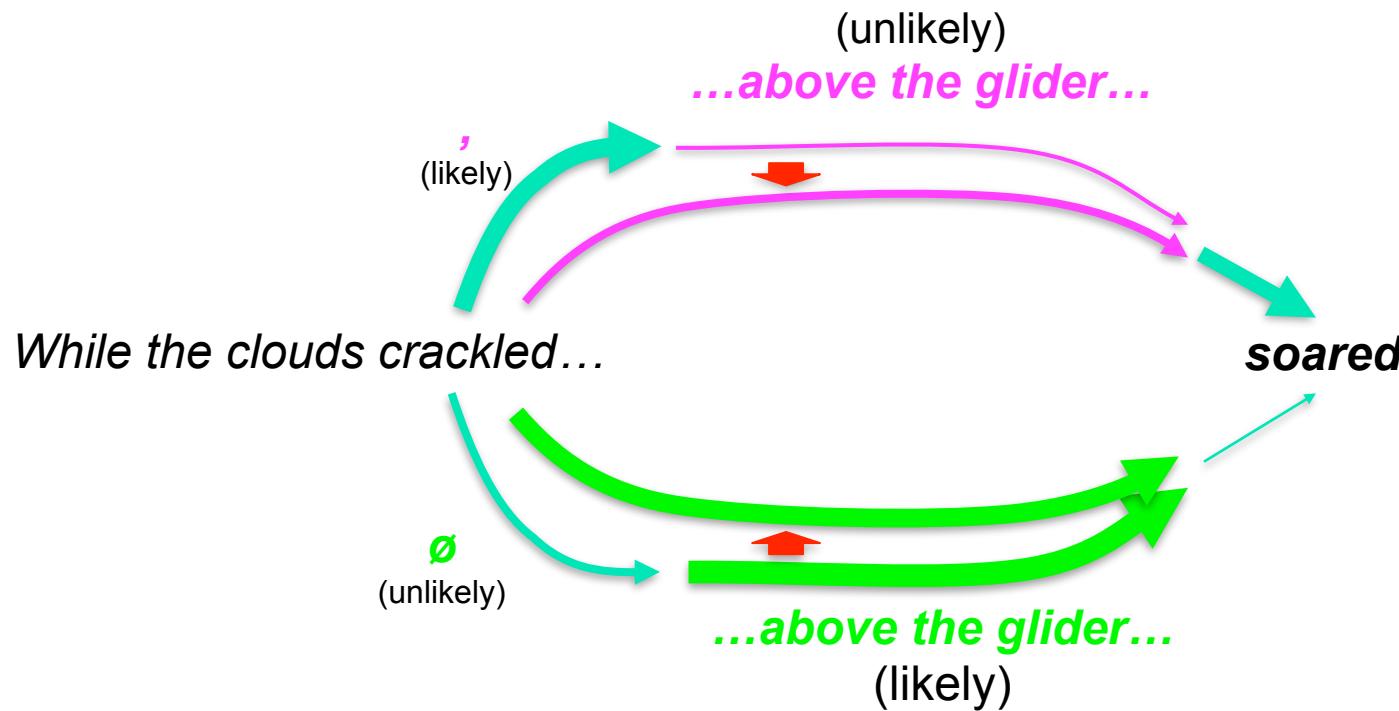
- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

$$P(w_i|\text{Context}) = \sum_{\text{Path}} P(w_i|\text{Path, Context})P(\text{Path}|\text{Context})$$



- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

$$P(w_i|\text{Context}) = \sum_{\text{Path}} P(w_i|\text{Path, Context}) P(\text{Path}|\text{Context})$$



- Inferences as probabilistic paths through the sentence:
  - Perceptual cost of ignoring the comma
  - Unlikeliness of main-clause continuation after comma
  - Likeliness of postverbal continuation without comma
- These inferences together make *soared* very surprising!

$$P(w_i|\text{Context}) = \sum_{\text{Path}} P(w_i|\text{Path, Context}) P(\text{Path}|\text{Context})$$

# Prediction 2: hallucinated garden paths

---

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion
- Experimental design: cross (1) and (2)

While the clouds crackled, above the glider soared a magnificent eagle.

While the clouds crackled, the glider soared above a magnificent eagle.

While the clouds crackled in the distance, above the glider soared a magnificent eagle.

While the clouds crackled in the distance, the glider soared above a magnificent eagle.

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion
- Experimental design: cross (1) and (2)

While the clouds crackled, above the glider soared a magnificent eagle.

While the clouds crackled, the glider soared above a magnificent eagle.

While the clouds crackled in the distance, above the glider soared a magnificent eagle.

While the clouds crackled in the distance, the glider soared above a magnificent eagle.

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion
- Experimental design: cross (1) and (2)

While the clouds crackled, above the glider soared a magnificent eagle.

While the clouds crackled, the glider soared above a magnificent eagle.

While the clouds crackled in the distance, above the glider soared a magnificent eagle.

While the clouds crackled in the distance, the glider soared above a magnificent eagle.

- The phrase *in the distance* fulfills a similar thematic role as above the glider for crackled

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion
- Experimental design: cross (1) and (2)

While the clouds crackled, above the glider soared a magnificent eagle.

While the clouds crackled, the glider soared above a magnificent eagle.

While the clouds crackled in the distance, above the glider soared a magnificent eagle.

While the clouds crackled in the distance, the glider soared above a magnificent eagle.

- The phrase *in the distance* fulfills a similar thematic role as above the glider for crackled
- Should reduce hallucinated garden-path effect

# Prediction 2: hallucinated garden paths

---

- Two properties come together to create “hallucinated garden path”
  1. Subordinate clause into which the main-clause inverted phrase would fit well
  2. Main clause with locative inversion
- Experimental design: cross (1) and (2)

While the clouds crackled, above the glider soared a magnificent eagle.

While the clouds crackled, the glider soared above a magnificent eagle.

While the clouds crackled in the distance, above the glider soared a magnificent eagle.

While the clouds crackled in the distance, the glider soared above a magnificent eagle.

- The phrase *in the distance* fulfills a similar thematic role as above the glider for crackled
- Should reduce hallucinated garden-path effect

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading
- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading
- 
- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-eleuds-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-clouds-crackled,-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-eleuds-erackled,-above-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-eleuds-crackled,-above-the-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-eleuds-crackled,-above-the-glider-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-eleuds-crackled,-above-the-glider-seared-----

- Readers aren't allowed to backtrack

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-clouds-crackled,-above-the-glider-soared-----

- Readers aren't allowed to backtrack
- So the comma is visually *gone* by the time the inverted main clause appears

# Prediction 2: Hallucinated garden paths

---

- Methodology: word-by-word self-paced reading

white-the-~~the~~-heads-crackled,-above-the-glider-soared-----

- Readers aren't allowed to backtrack
- So the comma is visually *gone* by the time the inverted main clause appears
- Simple test of whether beliefs about previous input can be revised

# Model predictions

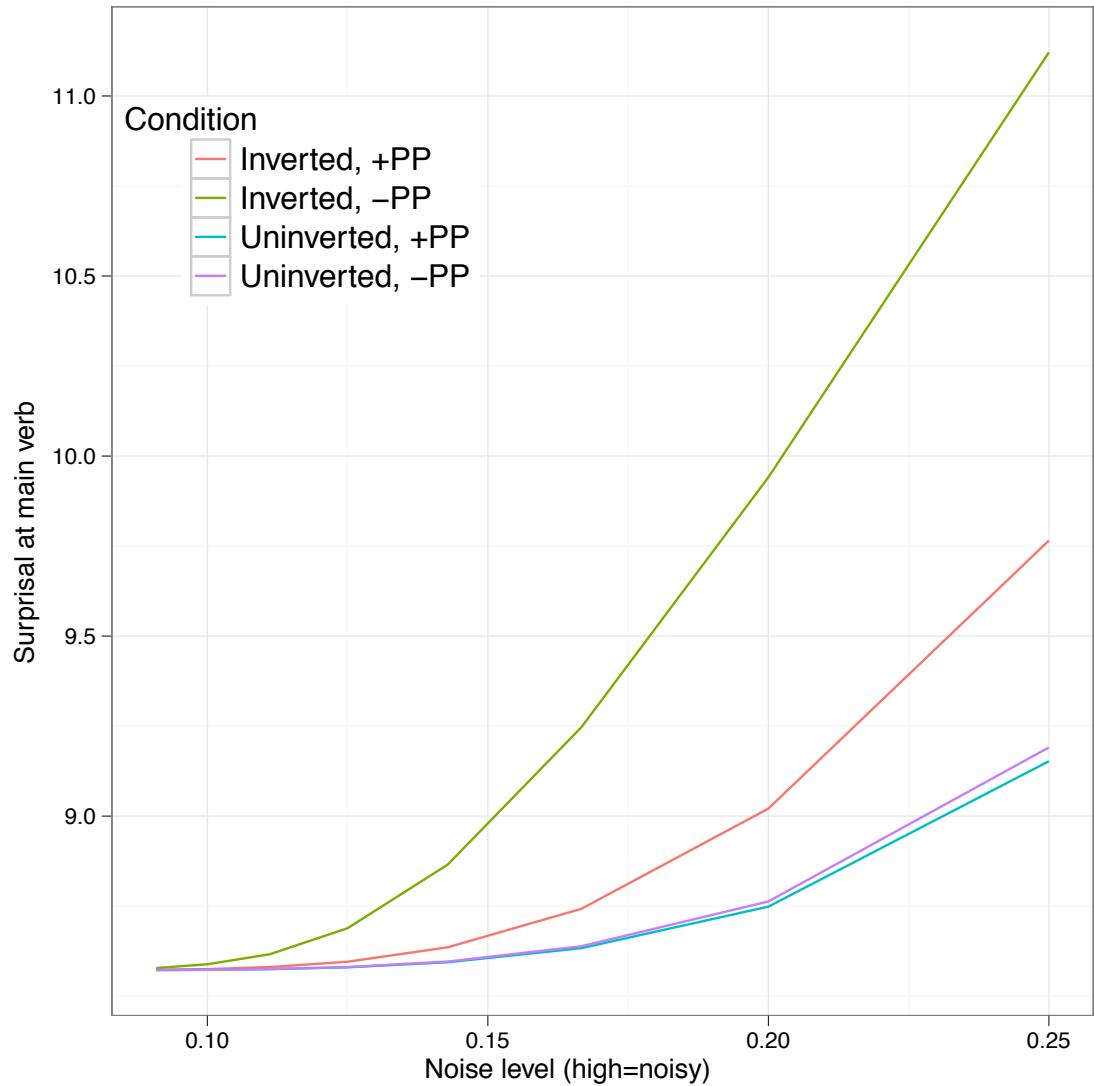
---

While the clouds  
**crackled**, **above** the  
glider soared a  
magnificent eagle.

While the clouds **crackled**  
**in the distance**, **above**  
the glider soared a  
magnificent eagle.

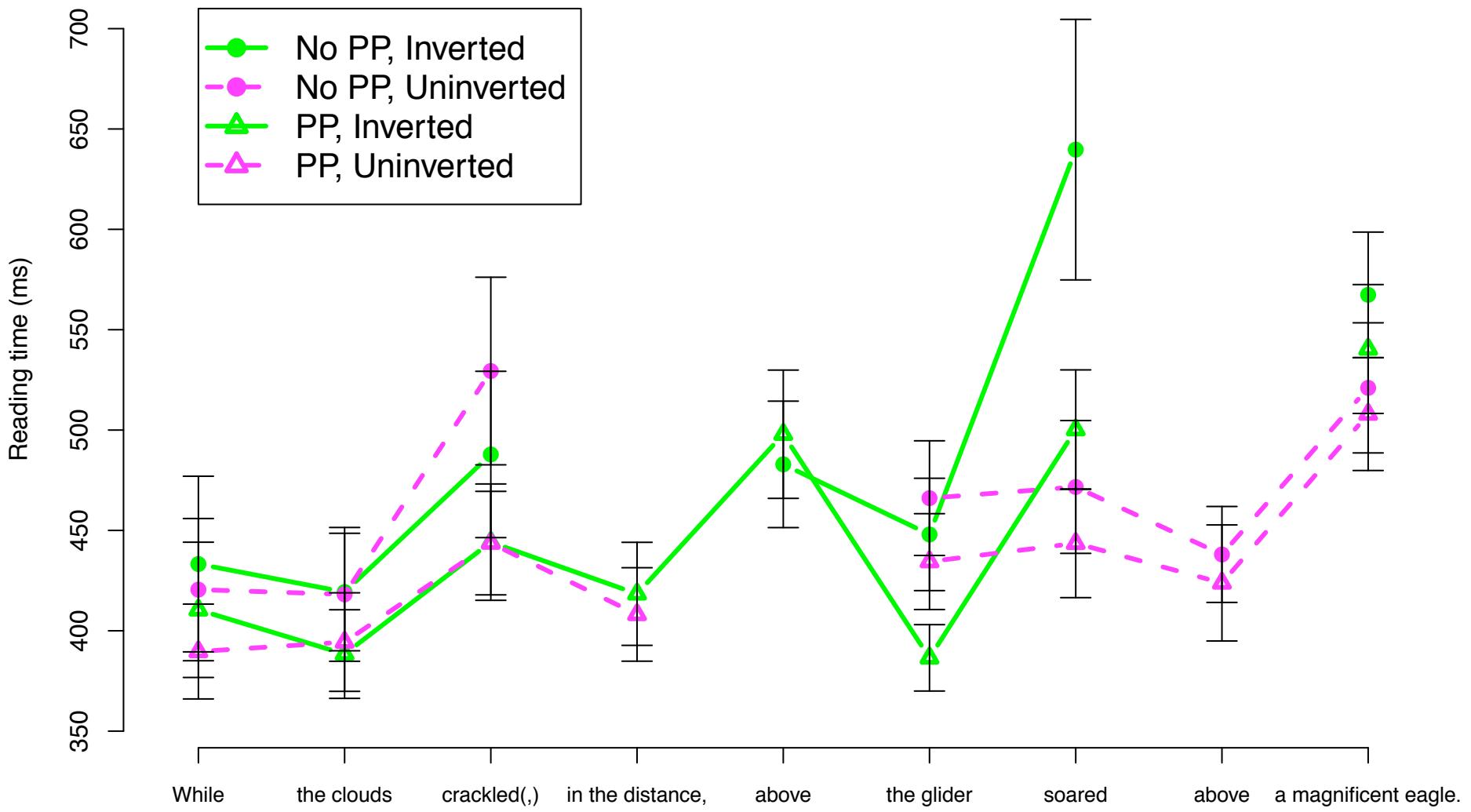
While the clouds  
**crackled**, the glider  
soared **above** a  
magnificent eagle.

While the clouds **crackled**  
**in the distance**, the  
glider soared **above** a  
magnificent eagle.

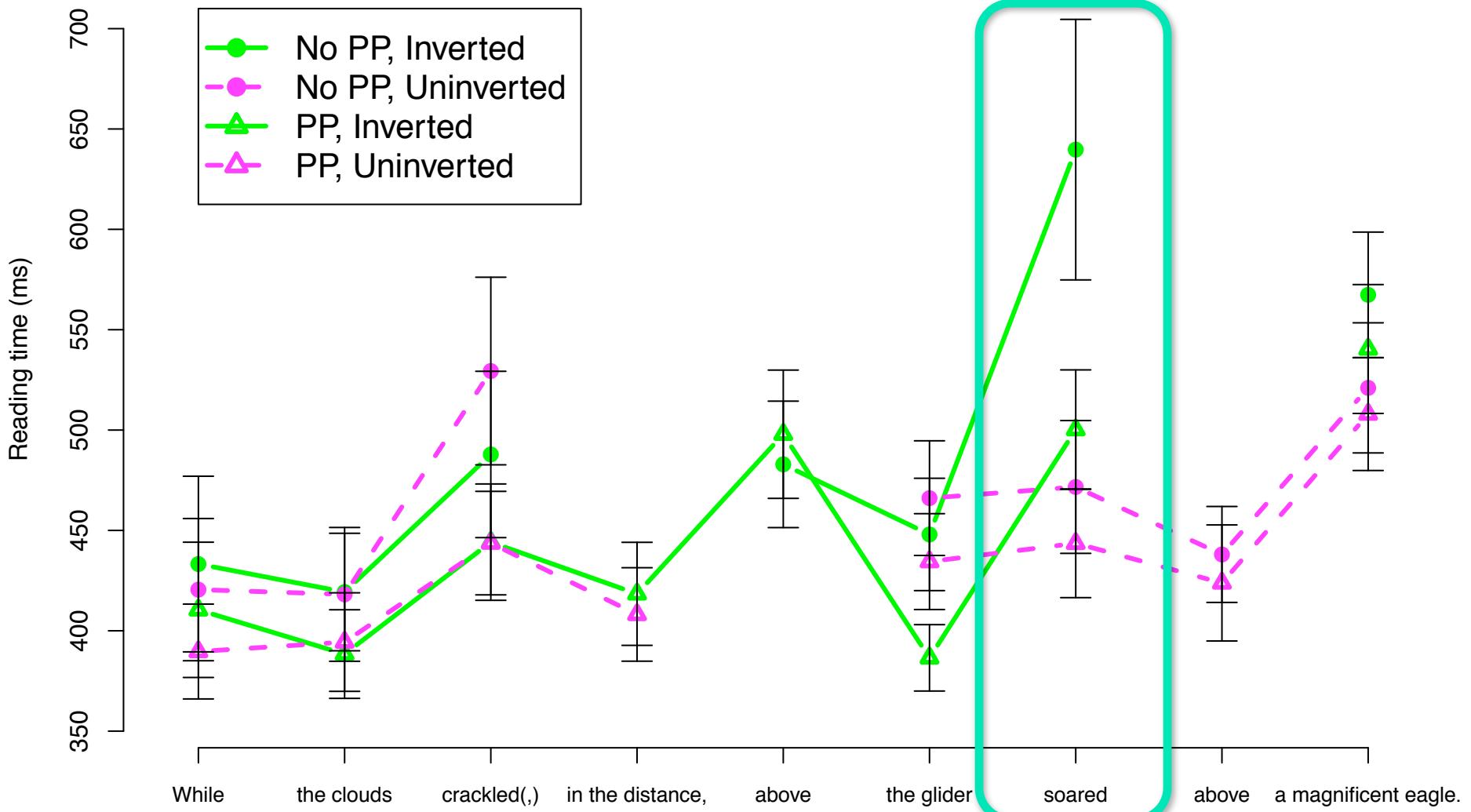


# Results: whole sentence reading times

---

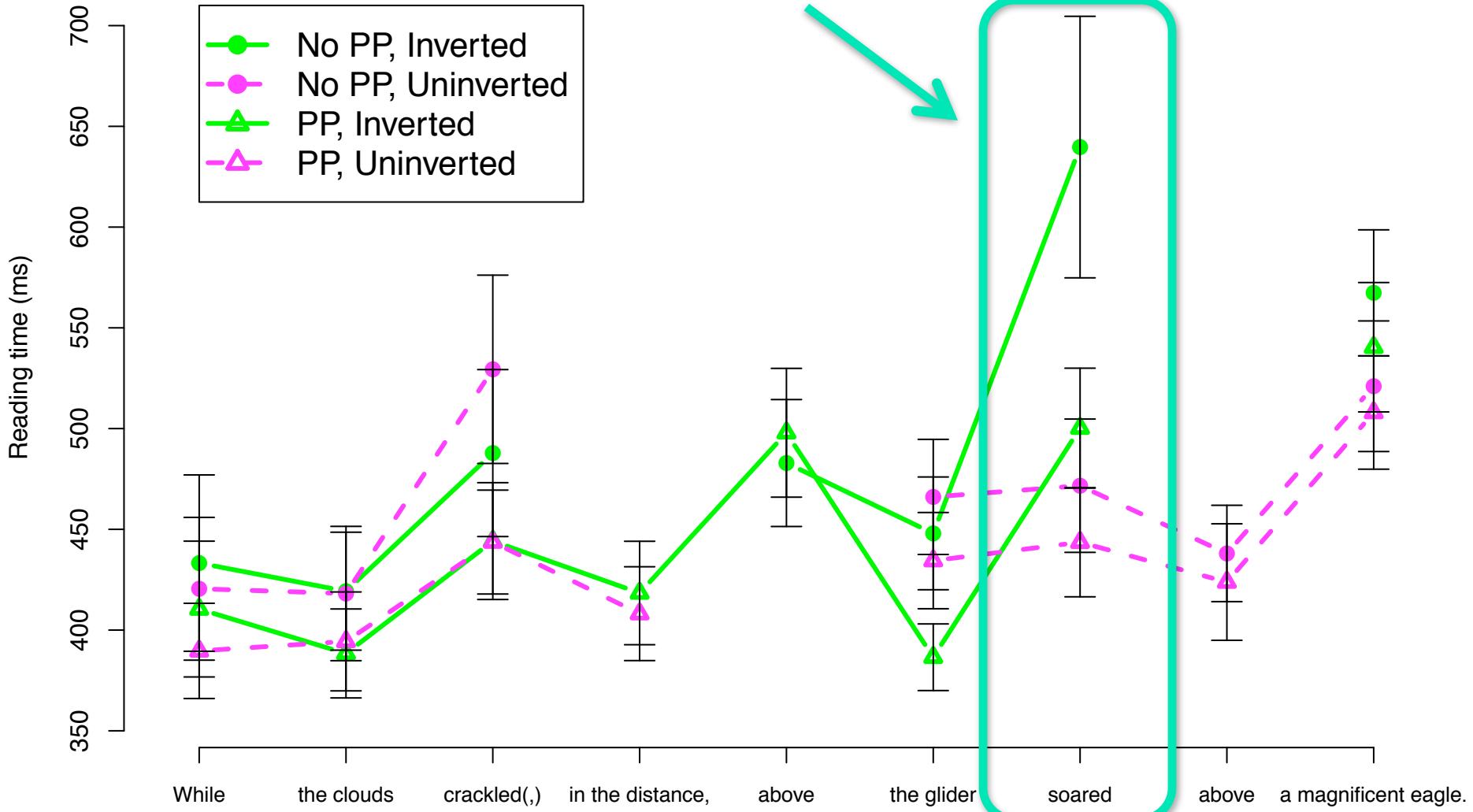


# Results: whole sentence reading times



# Results: whole sentence reading times

*Processing boggle occurs exactly where predicted*



# Hallucinated garden-path summary

---

- The *at/toward* study showed that comprehenders *note the possibility of alternative strings and act on it*
- This study showed that comprehenders can actually *devote resources to grammatical analyses inconsistent with the surface string*

# Hallucinated garden paths cont'd

---

- Sure, but punctuation's weird stuff
  - What about *real words*?
- 
- At least sometimes, bias *against N N interpretation*

# Hallucinated garden paths cont'd

---

- Sure, but punctuation's weird stuff
- What about *real words*?

*I know that the desert trains could resupply the camp.*

- At least sometimes, bias *against N N interpretation*

# Hallucinated garden paths cont'd

---

- Sure, but punctuation's weird stuff
- What about *real words*?

*I know that the desert trains could resupply the camp.*

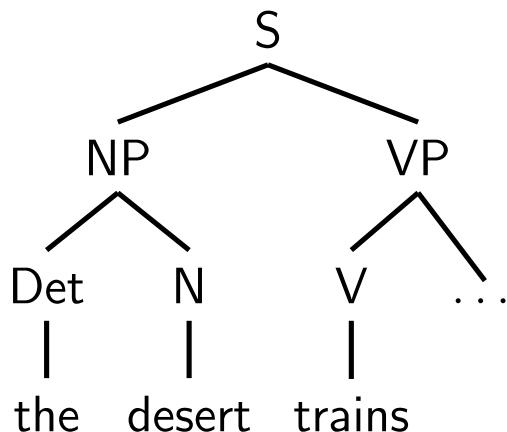
- At least sometimes, bias *against N N interpretation*

# Hallucinated garden paths cont'd

---

- Sure, but punctuation's weird stuff
- What about *real words*?

*I know that the desert trains could resupply the camp.*

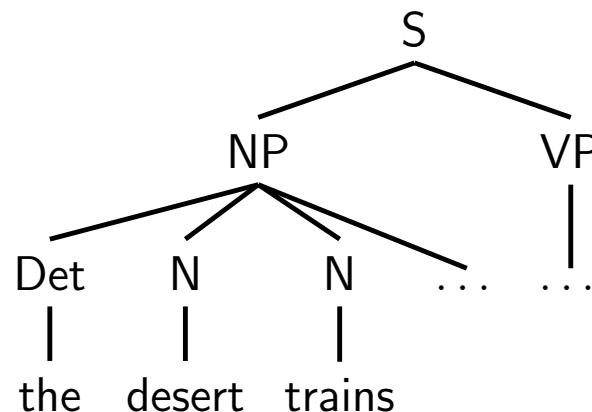
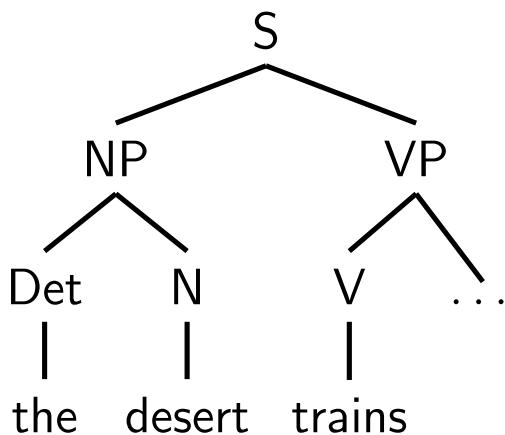


- At least sometimes, bias *against* N N interpretation

# Hallucinated garden paths cont'd

- Sure, but punctuation's weird stuff
- What about *real words*?

*I know that the desert trains could resupply the camp.*



- At least sometimes, bias against N N interpretation

# Hallucinated GPs with words

---

*Could be “intern chauffeured”*

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*The intern chauffeured for the governor but hoped for more interesting work.*  
[NV, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*The intern chauffeured for the governor but hoped for more interesting work.*  
[NV, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

*The inexperienced chauffeur for the governor hoped for more interesting work.*  
[NN, “sparse” neighborhood]

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*The intern chauffeured for the governor but hoped for more interesting work.*  
[NV, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

*The inexperienced chauffeur for the governor hoped for more interesting work.*  
[NN, “sparse” neighborhood]

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*The intern chauffeured for the governor but hoped for more interesting work.*  
[NV, “dense” neighborhood]

*Could NOT be “inexperienced chauffeured”*

*The inexperienced chauffeur for the governor hoped for more interesting work.*  
[NN, “sparse” neighborhood]

# Hallucinated GPs with words

---

- We use a contextual bias against NN and toward NV to test for GP hallucinations involving wordform change

*Could be “intern chauffeured”*

*The intern chauffeur for the governor hoped for more interesting work.*  
[NN, “dense” neighborhood]

*The intern chauffeured for the governor but hoped for more interesting work.*  
[NV, “dense” neighborhood]

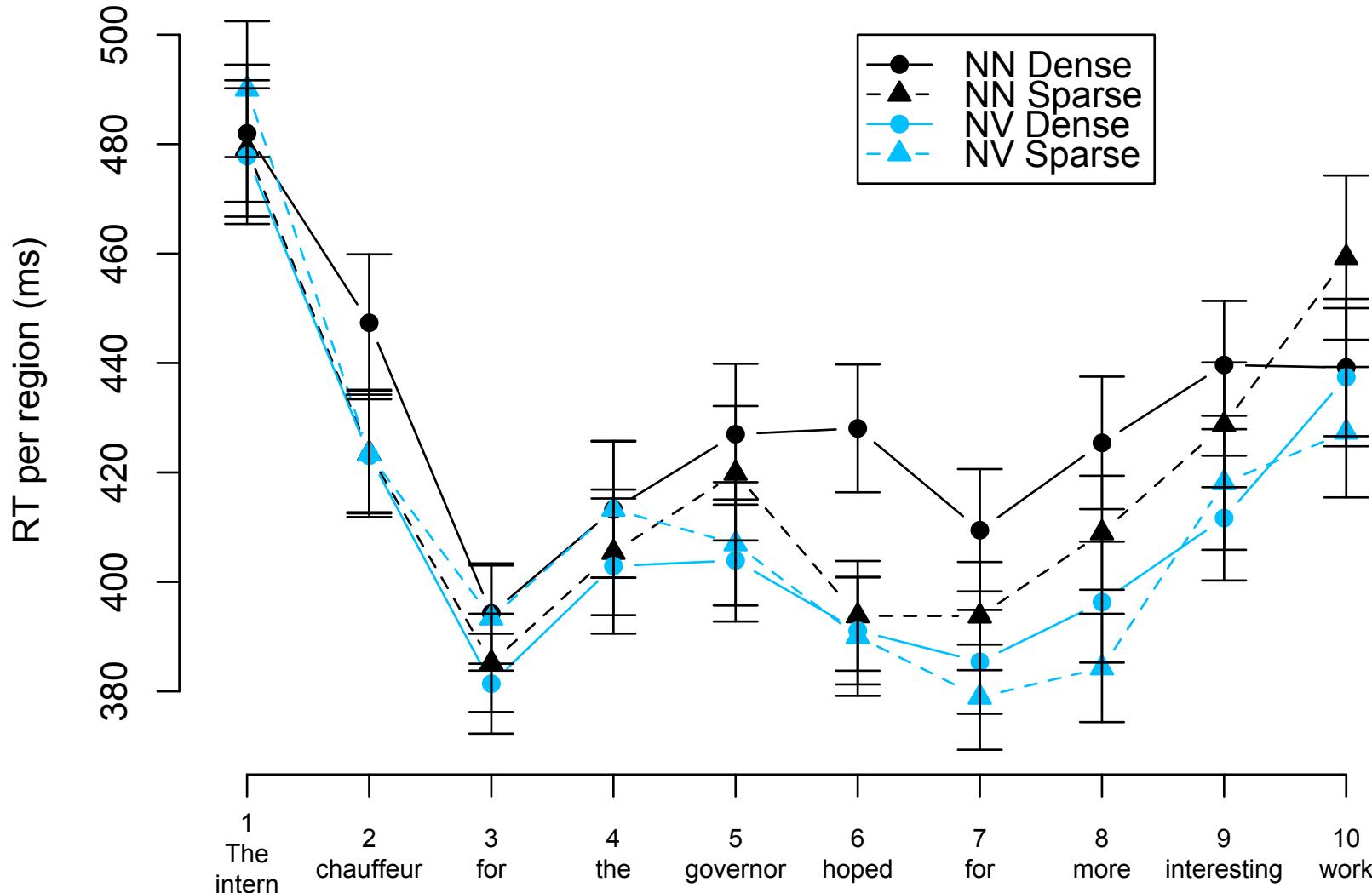
*Could NOT be “inexperienced chauffeured”*

*The inexperienced chauffeur for the governor hoped for more interesting work.*  
[NN, “sparse” neighborhood]

*Some interns chauffeured for the governor but hoped for more interesting work.*  
[NV, “sparse” neighborhood]

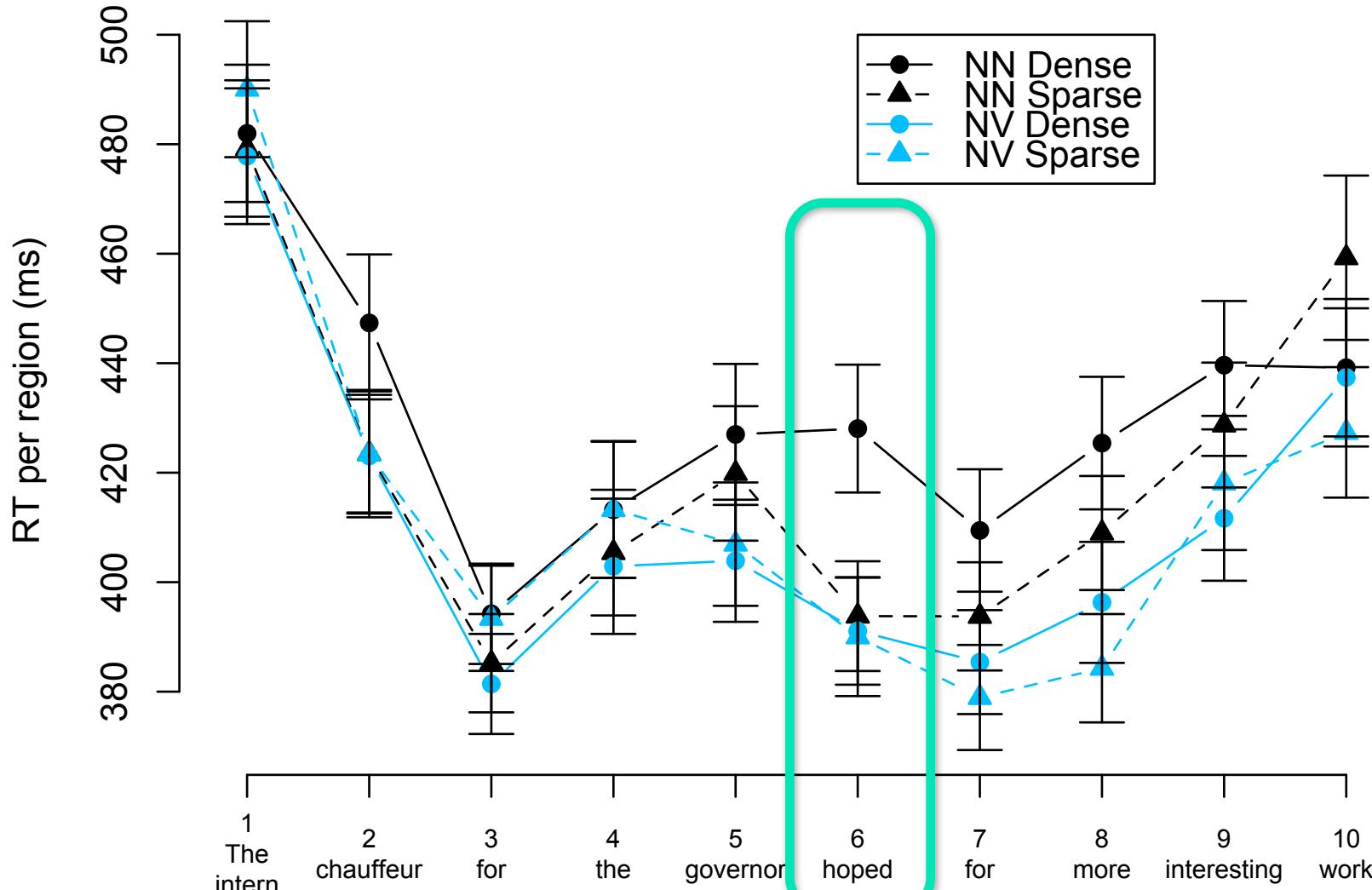
# Results

- RT spike at disambiguating region for NN Dense



# Results

- RT spike at disambiguating region for NN Dense



# Structure of the noise model

- Gibson et al. (2013) explored noise model with restricted noise operations: word *insertions* and *deletions*

Sentence	Plausibility	Insertions	Deletions
The cook baked Lucy a cake.	Plausible	0	1
The cook baked Lucy <i>for</i> a cake.	Implausible	1	0
The cook baked a cake <i>for</i> Lucy.	Plausible	1	0
The cook baked a cake Lucy.	Implausible	0	1

# Structure of the noise model

---

Sentence	Construction	Edits
The girl <u>was</u> kicked <u>by</u> the ball.	passive	2I
The ball kicked the girl.	active	2D
The tax law benefited <u>from</u> the businessman.	intransitive	1I
The businessman benefited the tax law.	transitive	1D
The cook baked Lucy <u>for</u> a cake.	Prepositional Object (PO) benefactive	1I
The cook baked a cake Lucy.	Double Object (DO) benefactive	1D

# Structure of the noise model

Sentence	Construction	Edits
The girl <u>was</u> kicked <u>by</u> the ball.	passive	2I
The ball kicked the girl.	active	2D
The tax law benefited <u>from</u> the businessman.	intransitive	1I
The businessman benefited the tax law.	transitive	1D
The cook baked Lucy <u>for</u> a cake.	Prepositional Object (PO) benefactive	1I
The cook baked a cake Lucy.	Double Object (DO) benefactive	1D



*Often “corrected” to plausible interpretation inconsistent with literal meaning*

# Structure of the noise model

*Consistently given literal interpretation*

## Sentence

The girl was kicked by the ball.

The ball kicked the girl.

The tax law benefited from the businessman.

The businessman benefited the tax law.

The cook baked Lucy for a cake.

The cook baked a cake Lucy.

## Construction

passive

active

intransitive

transitive

Prepositional Object  
(PO) benefactive

Double Object (DO)  
benefactive

## Edits

2I

2D

1I

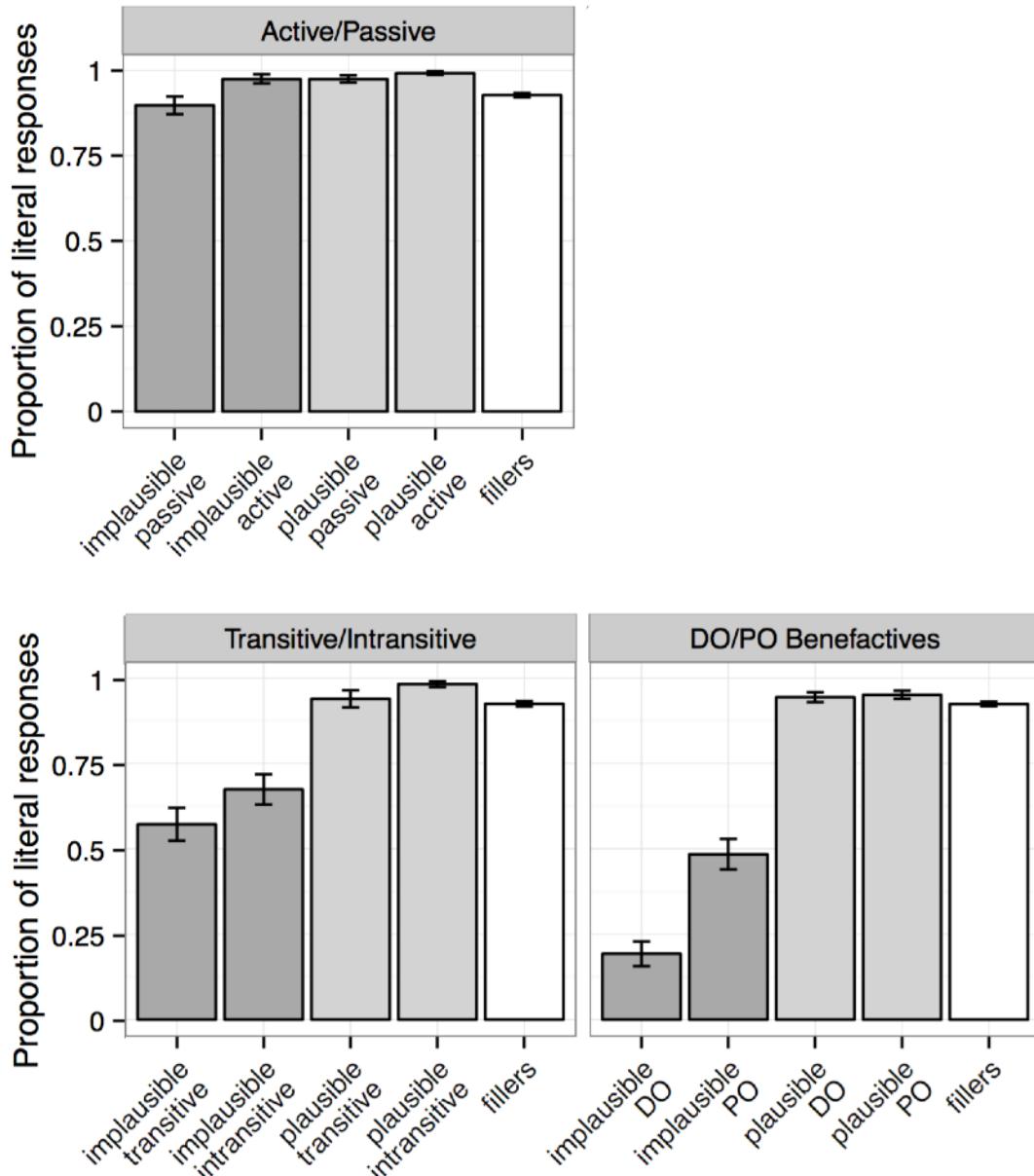
1D

1I

1D

*Often “corrected” to plausible interpretation inconsistent with literal meaning*

# Noisy-channel inference results



(Poppels & Levy 2015,  
replication of Gibson et al., 2013)

# Exchanges in the noise model?

---

*This is a problem that I need to talk about Joe with.*

# Exchanges in the noise model?

---

*This is a problem that I need to talk about Joe with.*

- An occasional speech error of mine that I've noticed for years, but that no one ever notices me make

# Exchanges in the noise model?

---

*This is a problem that I need to talk about Joe with.*

- An occasional speech error of mine that I've noticed for years, but that no one ever notices me make
- Extraordinarily unlikely under the Gibson noise model

# Exchanges in the noise model?

---

*This is a problem that I need to talk about Joe with.*

- An occasional speech error of mine that I've noticed for years, but that no one ever notices me make
- Extraordinarily unlikely under the Gibson noise model
- But reasonably likely if word **exchanges** are admitted

# Exchanges in the noise model?

---

*This is a problem that I need to talk about Joe with.*

- An occasional speech error of mine that I've noticed for years, but that no one ever notices me make
- Extraordinarily unlikely under the Gibson noise model
- But reasonably likely if word **exchanges** are admitted

The package fell from the table to the floor. [plausible; canonical]

The package fell to the floor from the table. [plausible; non-canonical]

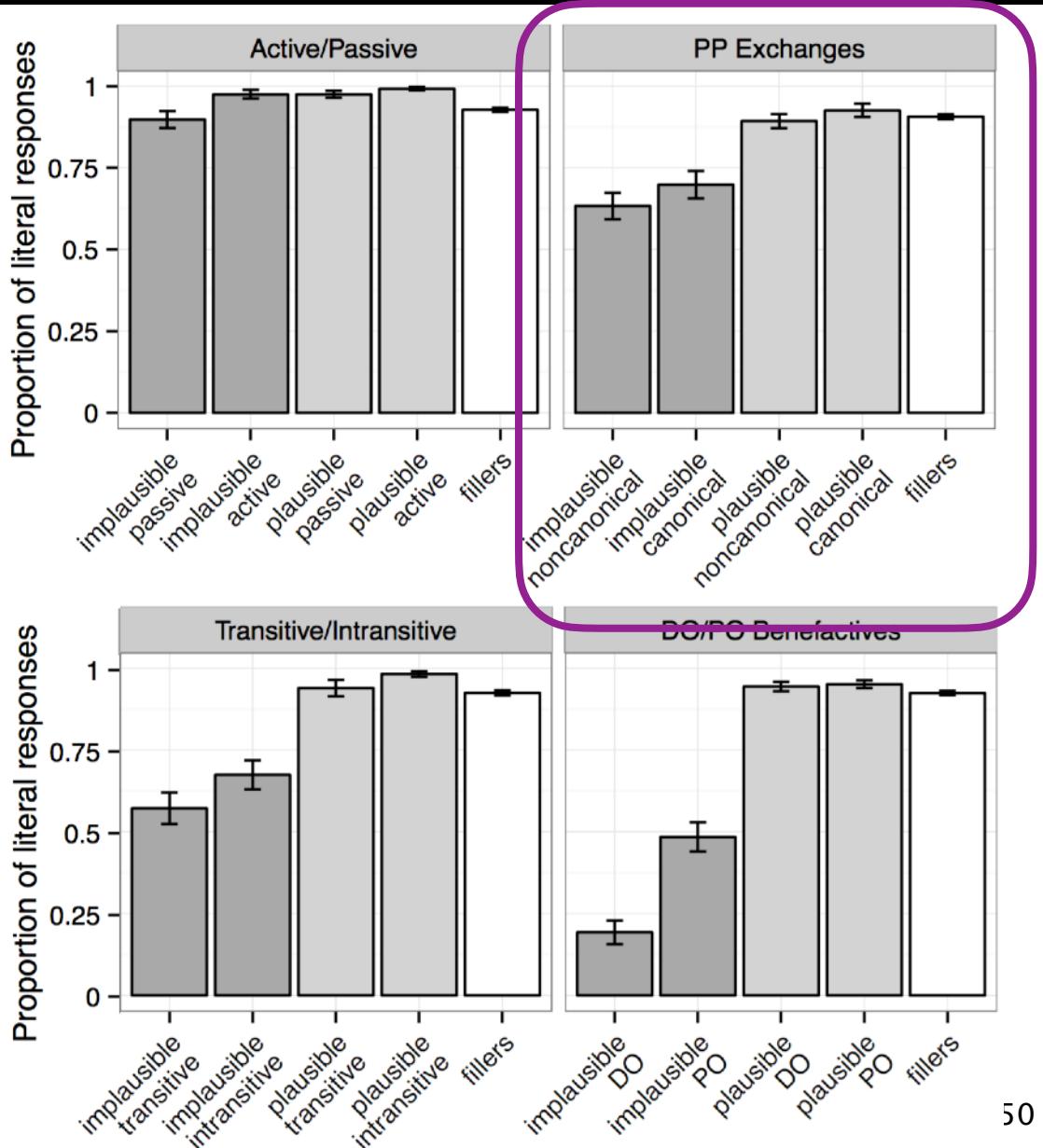
The package fell from the floor to the table. [implausible; canonical]

The package fell to the table from the floor. [implausible; non-canonical]

*Till Poppels*



# Noisy-channel inference results



# Structural Forgetting and the Noisy Channel



# Structural Forgetting and the Noisy Channel

1. The apartment that the maid who the cleaning service sent over was well-decorated.



# Structural Forgetting and the Noisy Channel

1. The apartment that the maid who the cleaning service sent over was well-decorated.
2. The apartment that the maid who the cleaning service sent over cleaned was well-decorated.

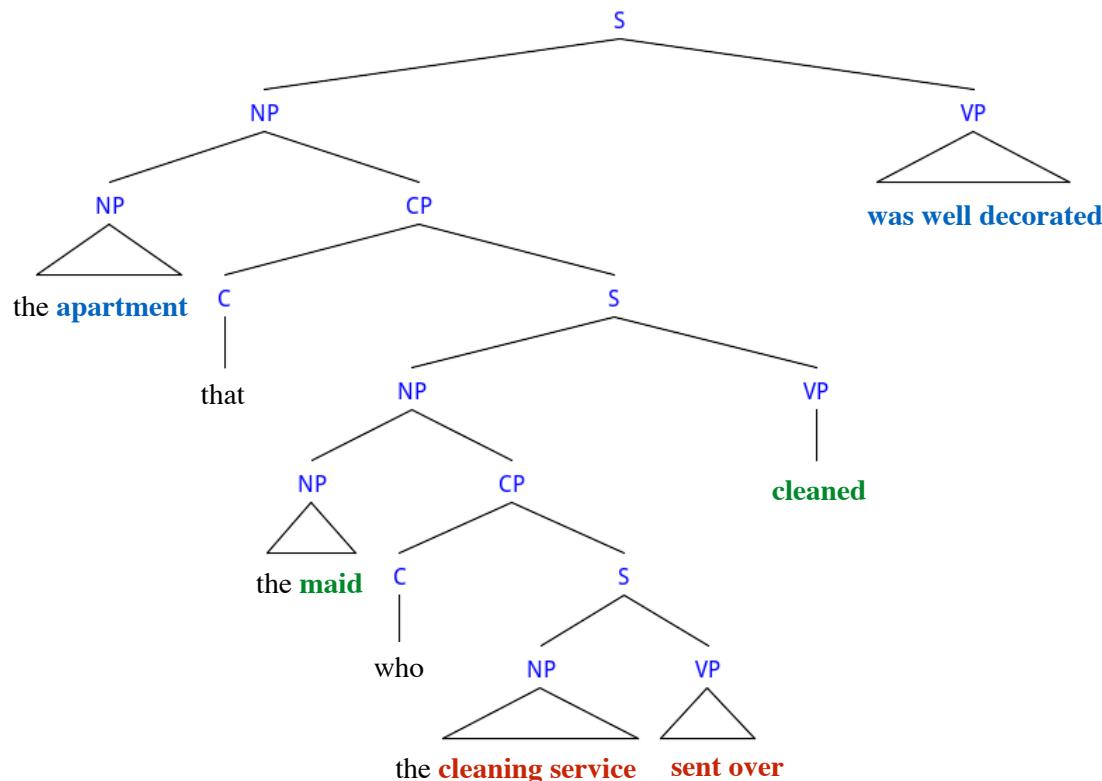


## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 

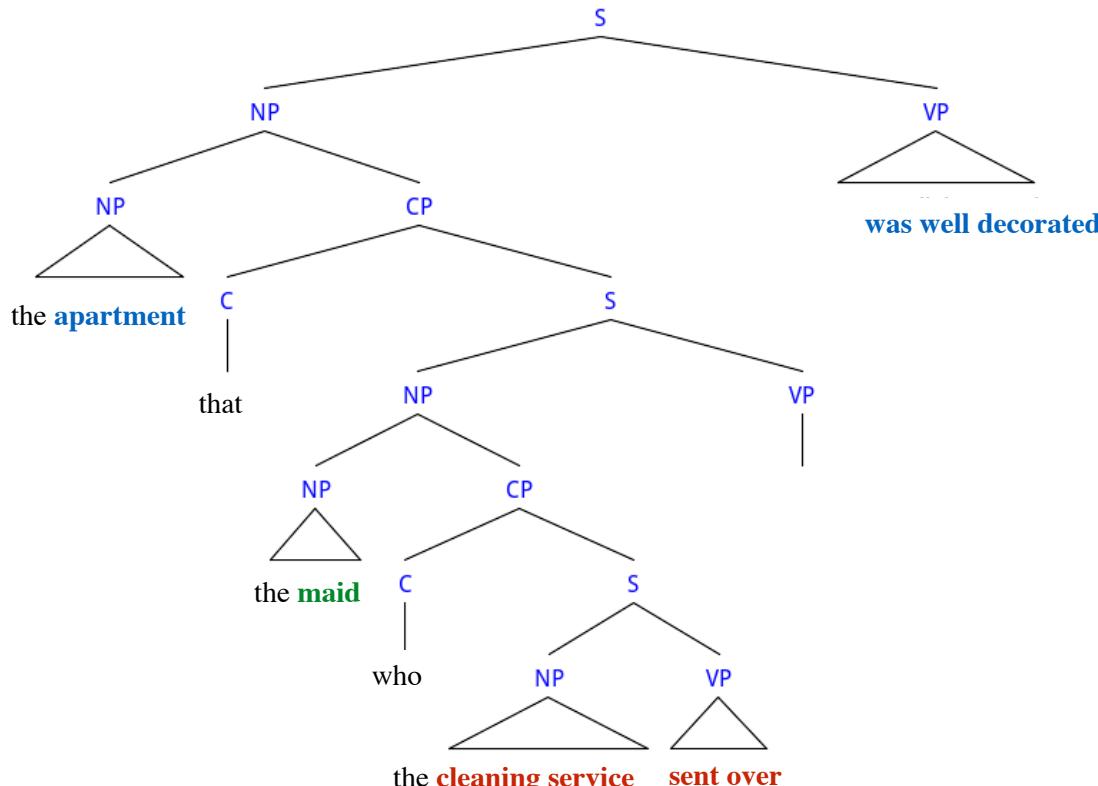
# Structural Forgetting

- \*The **apartment** that the **maid** who the **cleaning service sent over** was well-decorated. 
  - The **apartment** that the **maid** who the **cleaning service sent over cleaned** was well-decorated. 



# Structural Forgetting

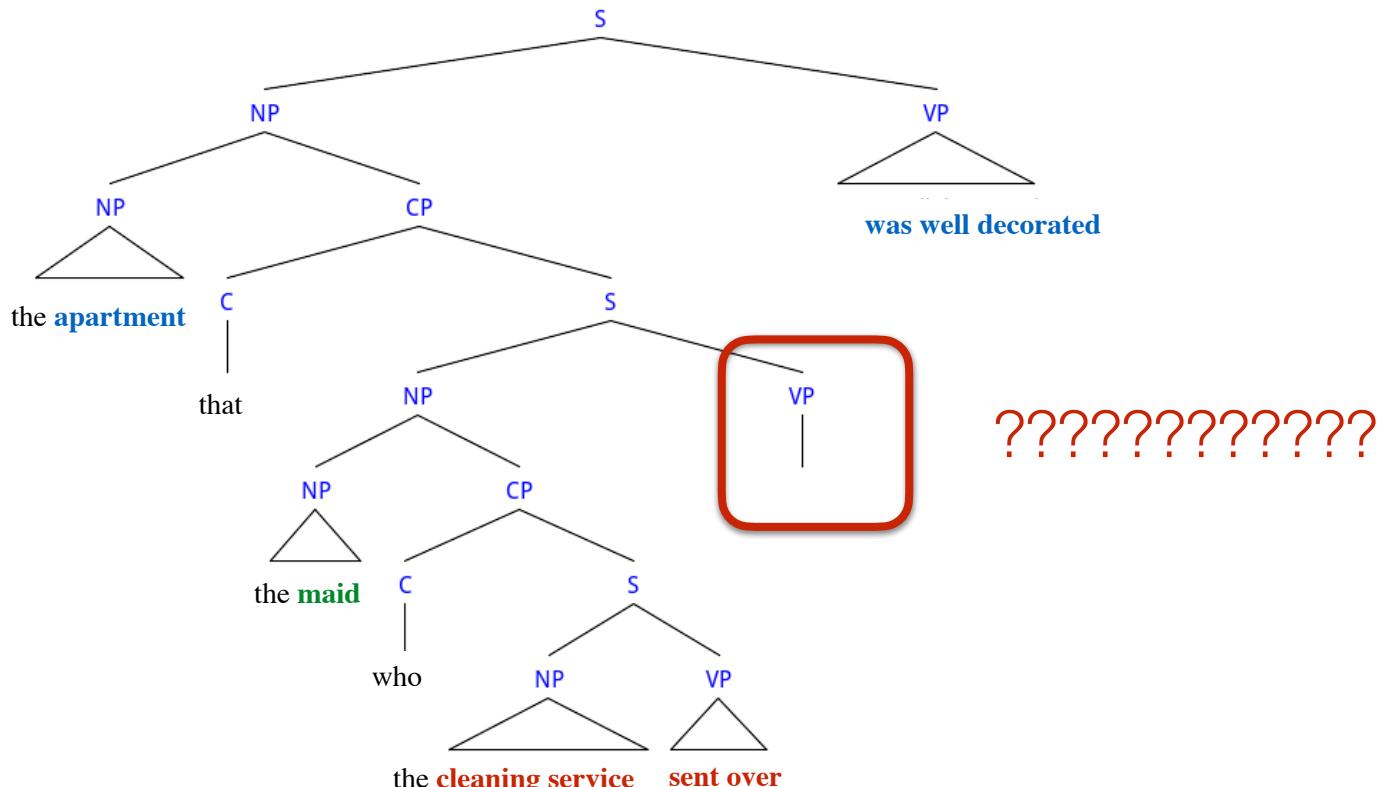
1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated**. 
2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated**. 



# Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 

2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 



## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- **Structural forgetting effect:** part of the sentence is forgotten by the time you get to the end (Gibson & Thomas, 1999; Frazier, 1985; Fodor, p.c.)

# Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- **Structural forgetting effect:** part of the sentence is forgotten by the time you get to the end (Gibson & Thomas, 1999; Frazier, 1985; Fodor, p.c.)
  - The ungrammatical sentence seems better than the grammatical one.
    - A "**grammaticality illusion**": how could we define grammaticality in this case?

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).

## Structural Forgetting

1. \*Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, war gut eingerichtet.** 
  2. Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, reinigte, war gut eingerichtet.** 
- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).
    - In German (and Dutch), people prefer 2 over 1.

## Structural Forgetting

1. \*Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, war gut eingerichtet.** 

2. Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, reinigte, war gut eingerichtet.** 

- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).
  - In German (and Dutch), people prefer 2 over 1.
  - What is the difference between English and German?

# Structural Forgetting

1. \*Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, war gut eingerichtet.** 
2. Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, reinigte, war gut eingerichtet.** 

- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).
  - In German (and Dutch), people prefer 2 over 1.
  - What is the difference between English and German?
  - Frank et al. (2016) show that at recurrent neural network gives higher probability to (1) in English, but (2) in German.

# Structural Forgetting

1. \*Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, war gut eingerichtet.** 
2. Die **Wohnung**, die das **Zimmermädchen**, das der **Reinigungsdienst übersandte, reinigte, war gut eingerichtet.** 

- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).
  - In German (and Dutch), people prefer 2 over 1.
  - What is the difference between English and German?
  - Frank et al. (2016) show that at recurrent neural network gives higher probability to (1) in English, but (2) in German.
    - But why?

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- These contexts are more common in German than English (Roland et al., 2007).

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- These contexts are more common in German than English (Roland et al., 2007).
    - English: the maid [that cleaned the apartment]  
the apartment [that the maid cleaned]

## Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- These contexts are more common in German than English (Roland et al., 2007).
    - English: the maid [that cleaned the apartment]      **80%**  
         the apartment [that the maid cleaned]

# Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- These contexts are more common in German than English (Roland et al., 2007).
    - English: the maid [that cleaned the apartment] **80%**  
the apartment [that the maid cleaned] **20%**

# Structural Forgetting

1. \*The **apartment** that the **maid** who the **cleaning service sent over was well-decorated.** 
  2. The **apartment** that the **maid** who the **cleaning service sent over cleaned was well-decorated.** 
- These contexts are more common in German than English (Roland et al., 2007).
    - English: the maid [that cleaned the apartment] **80%**  
the apartment [that the maid cleaned] **20%**
    - German: das Dienstmädchen, [das die Wohnung reinigte] die Wohnung, [die das Dienstmädchen reinigte]

# Noisy-Context Surprisal Account of Structural Forgetting

# Noisy-Context Surprisal Account of Structural Forgetting

- Structural forgetting means the ungrammatical sentence with two verbs is **easier to process** than the grammatical sentence with three verbs:

# Noisy-Context Surprisal Account of Structural Forgetting

- Structural forgetting means the ungrammatical sentence with two verbs is **easier to process** than the grammatical sentence with three verbs:

$C($  The **apartment** that the **maid** who the **cleaning service**  
**sent over was well-decorated.** ) <

$C($ The **apartment** that the **maid** who the **cleaning service**  
**sent over cleaned was well-decorated.**)

# Noisy-Context Surprisal Account of Structural Forgetting

- Structural forgetting means the ungrammatical sentence with two verbs is **easier to process** than the grammatical sentence with three verbs:

$$C(\text{NOUN THAT NOUN THAT NOUN VERB VERB}) < \\ C(\text{NOUN THAT NOUN THAT NOUN VERB VERB VERB})$$

# Noisy-Context Surprisal Account of Structural Forgetting

- Structural forgetting means the ungrammatical sentence with two verbs is **easier to process** than the grammatical sentence with three verbs:

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$

context

key word

NOUN THAT NOUN THAT VERB VERB

VERB  
#

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



# Noisy-Context Surprisal Account of Structural Forgetting

$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$



- Correct noise based on prior about the language.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

$$C(2 \text{ VERBS}) < C(3 \text{ VERBS})$$



- Correct noise based on prior about the language.
- Higher probability for verb-final RCs in German,
  - so more likely to make the right prediction.

# Noisy-Context Surprisal Account of Structural Forgetting

## Noisy-Context Surprisal Account of Structural Forgetting

- We demonstrate that this works for toy grammars of English and German.

# Noisy-Context Surprisal Account of Structural Forgetting

- We demonstrate that this works for toy grammars of English and German.

Rule	Probability
S -> NP v <sub>erb</sub>	1
NP -> n <sub>oun</sub>	$1-m$
NP -> n <sub>oun</sub> RC	$mr$
NP -> n <sub>oun</sub> PP	$m(1-r)$
PP -> p <sub>rep</sub> NP	1
RC -> t <sub>HAT</sub> v <sub>erb</sub> NP	$s$
RC -> t <sub>HAT</sub> NP v <sub>erb</sub>	$1-s$

# Noisy-Context Surprisal Account of Structural Forgetting

- We demonstrate that this works for toy grammars of English and German.

Rule	Probability					
S -> NP VERB	1	NOUN	VERB			
NP -> NOUN	$1-m$	NOUN	PREP	NOUN	VERB	
NP -> NOUN RC	$mr$	NOUN	THAT	VERB	NOUN	VERB
NP -> NOUN PP	$m(1-r)$	NOUN	THAT	NOUN	VERB	VERB
PP -> PREP NP	1	NOUN	THAT	NOUN	THAT	NOUN...
RC -> THAT VERB NP	$s$					
RC -> THAT NP VERB	$1-s$					

# Noisy-Context Surprisal Account of Structural Forgetting

## Noisy-Context Surprisal Account of Structural Forgetting

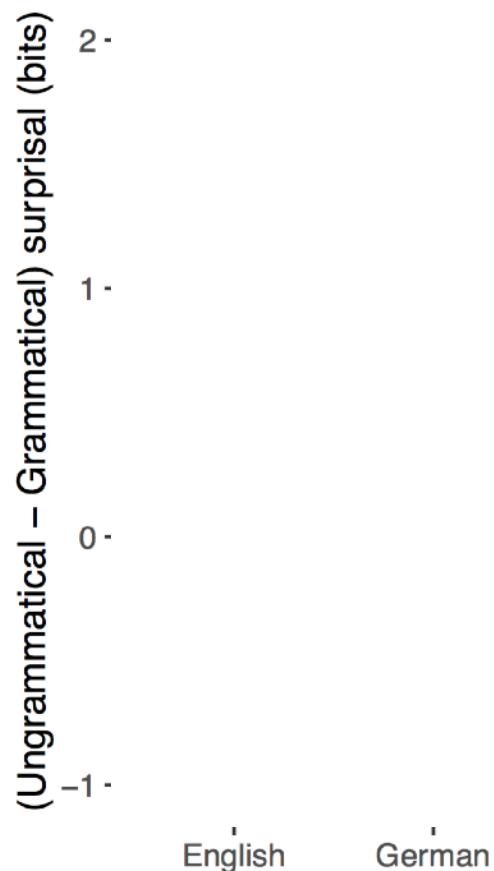
- Setting the verb-final RC rate to 100% for German and 20% for English (Roland et al., 2007),

## Noisy-Context Surprisal Account of Structural Forgetting

- Setting the verb-final RC rate to 100% for German and 20% for English (Roland et al., 2007),
- we find surprisal differences matching the forgetting effect:

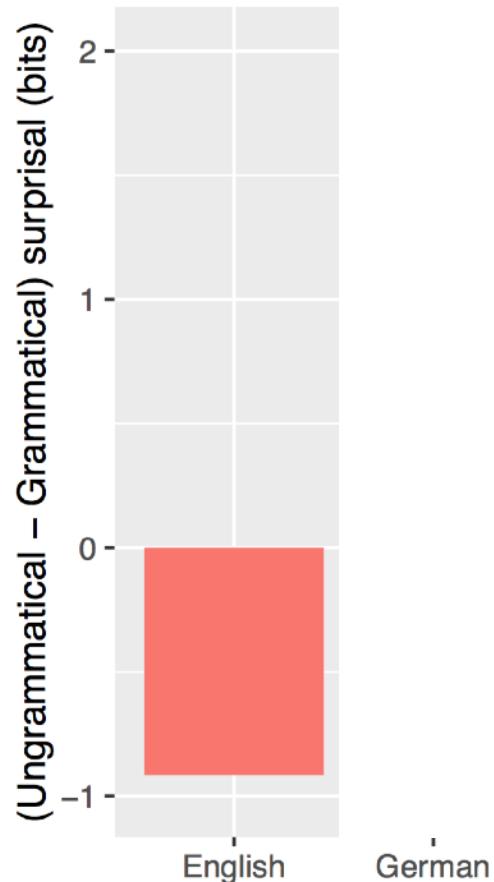
# Noisy-Context Surprisal Account of Structural Forgetting

- Setting the verb-final RC rate to 100% for German and 20% for English (Roland et al., 2007),
- we find surprisal differences matching the forgetting effect:



# Noisy-Context Surprisal Account of Structural Forgetting

- Setting the verb-final RC rate to 100% for German and 20% for English (Roland et al., 2007),
- we find surprisal differences matching the forgetting effect:

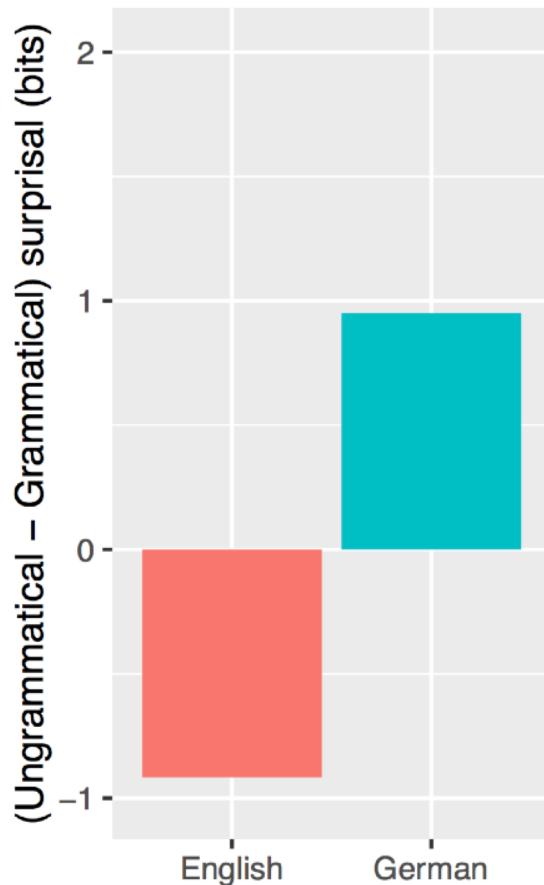


# Noisy-Context Surprisal Account of Structural Forgetting

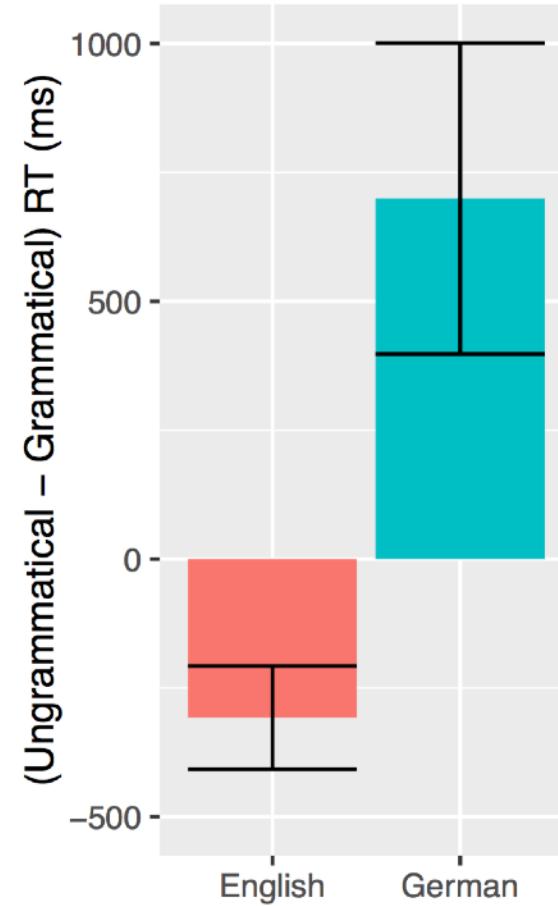
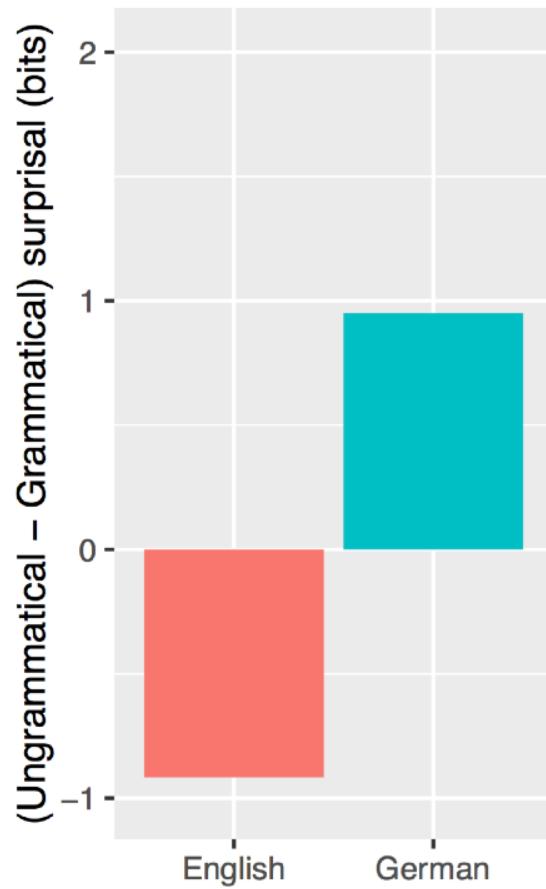
- Setting the verb-final RC rate to 100% for German and 20% for English (Roland et al., 2007),
- we find surprisal differences matching the forgetting effect:



# Noisy-Context Surprisal Account of Structural Forgetting



# Noisy-Context Surprisal Account of Structural Forgetting



Vasishth et al. (2010)

# Noisy-Context Surprisal Account of Structural Forgetting

## Noisy-Context Surprisal Account of Structural Forgetting

- Probability that a context is remembered depends on its prior probability.
  - Noisy-context surprisal *explains* the behavior of the RNN in Frank et al. (2016): the RNN is using a lossily compressed / noisy representation of context.

## Noisy-Context Surprisal Account of Structural Forgetting

- Probability that a context is remembered depends on its prior probability.
  - Noisy-context surprisal *explains* the behavior of the RNN in Frank et al. (2016): the RNN is using a lossily compressed / noisy representation of context.
- The model has an explicit grammar (competence), but cannot apply it correctly (performance).

# Dependency length and noisy-channel surprisal

---

# Dependency length and noisy-channel surprisal

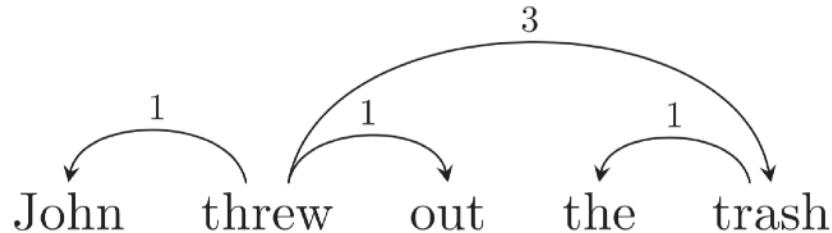
---

- Syntactic dependencies vary in linear distance

# Dependency length and noisy-channel surprisal

---

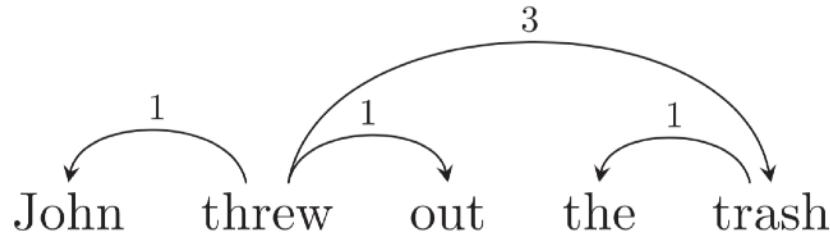
- Syntactic dependencies vary in linear distance



# Dependency length and noisy-channel surprisal

---

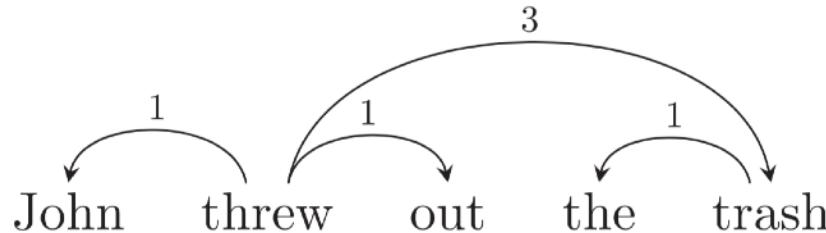
- Syntactic dependencies vary in linear distance



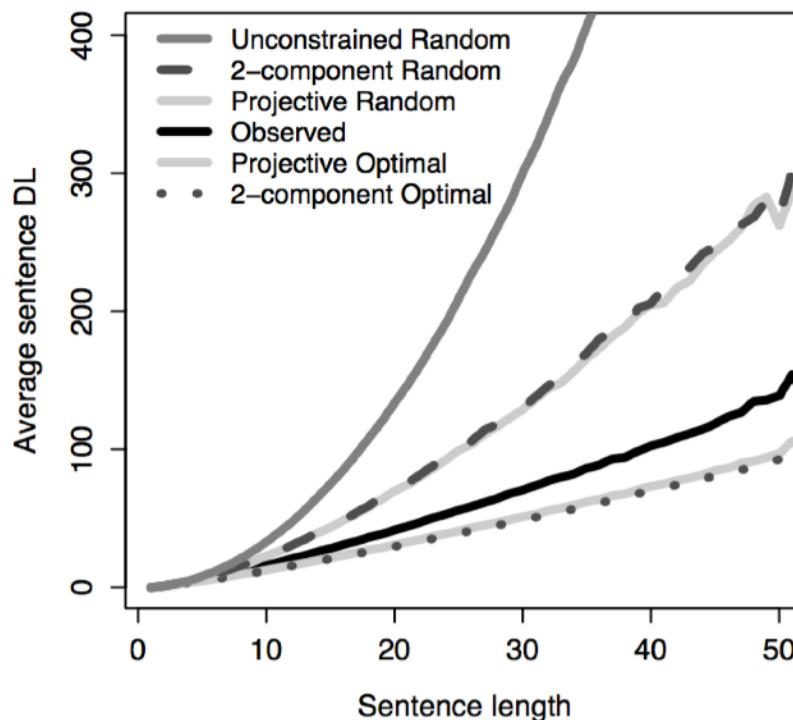
- Idea with long history: short dependencies preferred

# Dependency length and noisy-channel surprisal

- Syntactic dependencies vary in linear distance

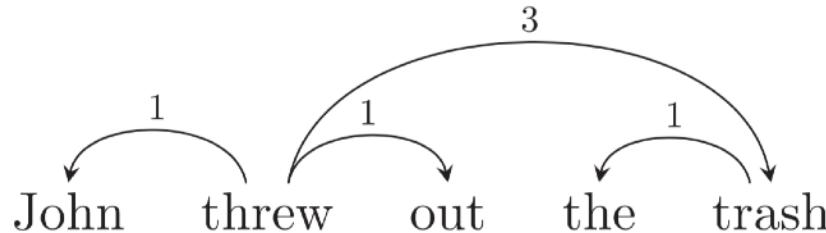


- Idea with long history: short dependencies preferred

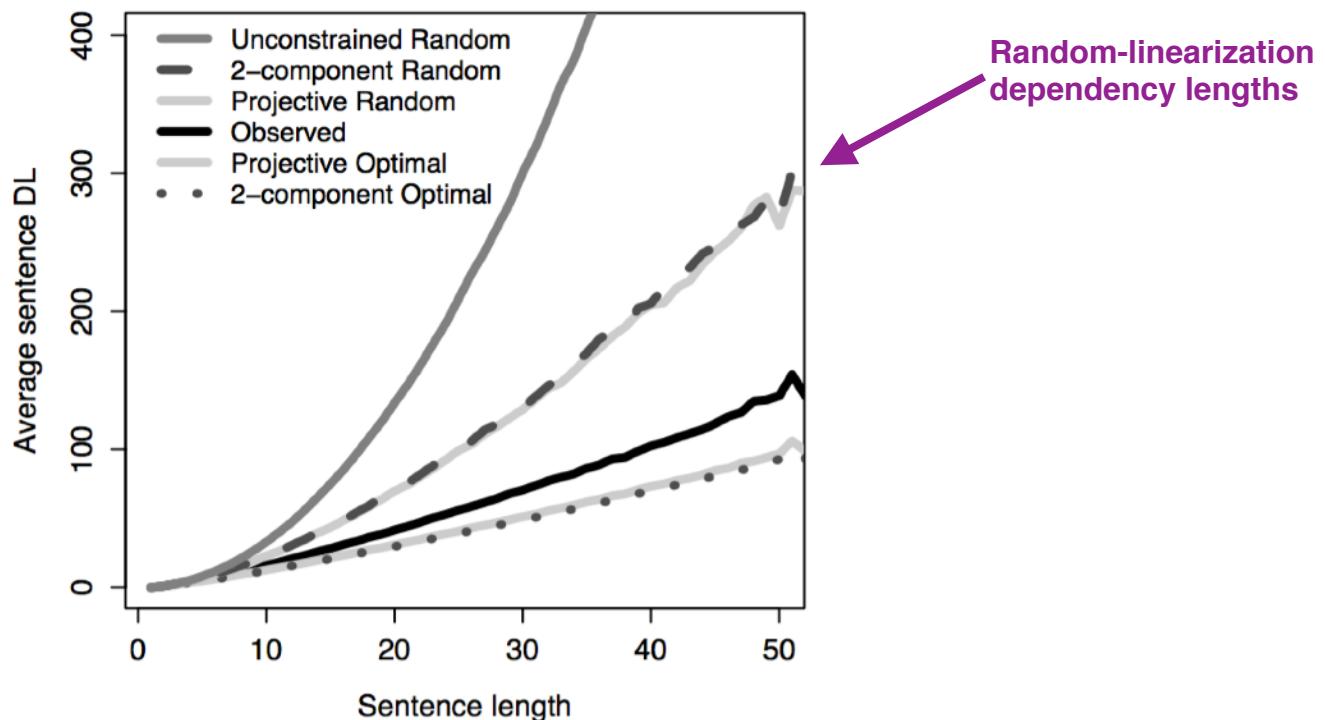


# Dependency length and noisy-channel surprisal

- Syntactic dependencies vary in linear distance

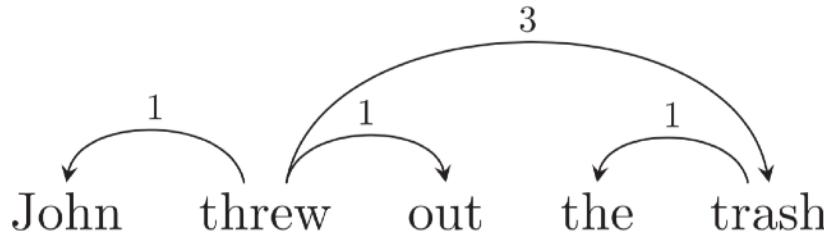


- Idea with long history: short dependencies preferred

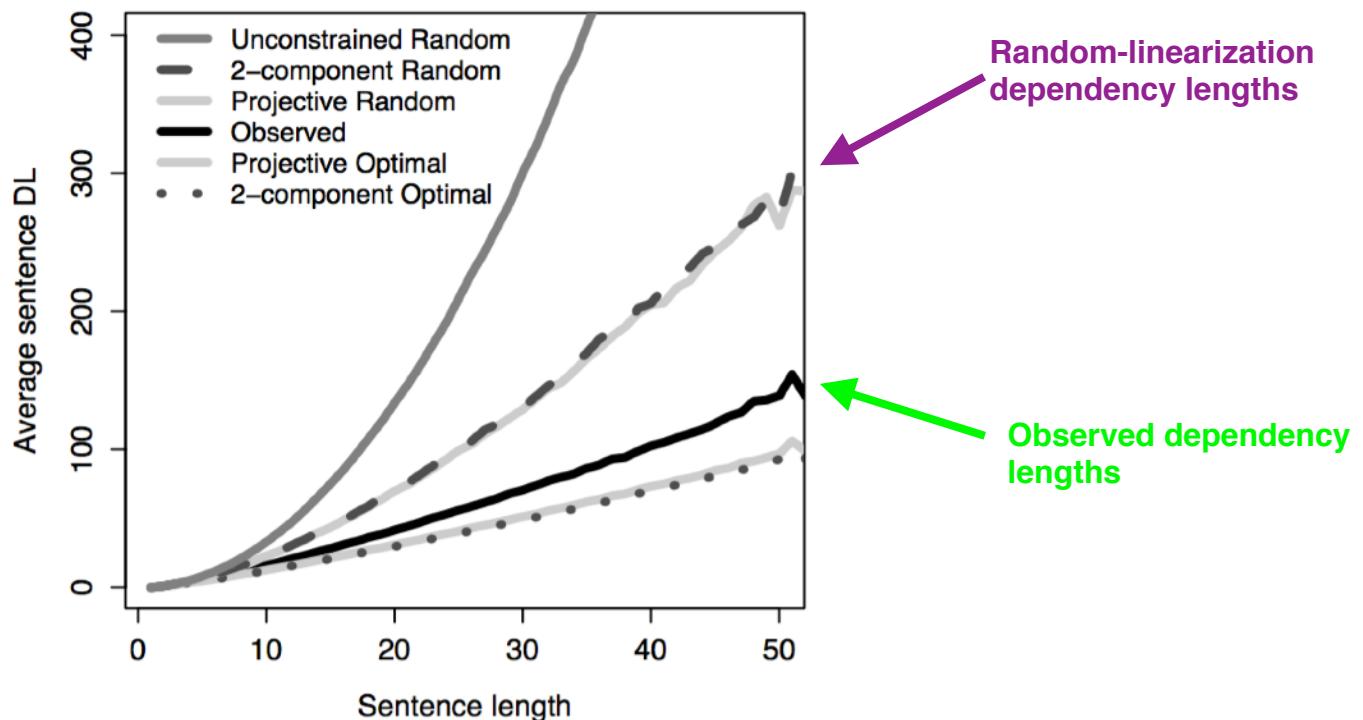


# Dependency length and noisy-channel surprisal

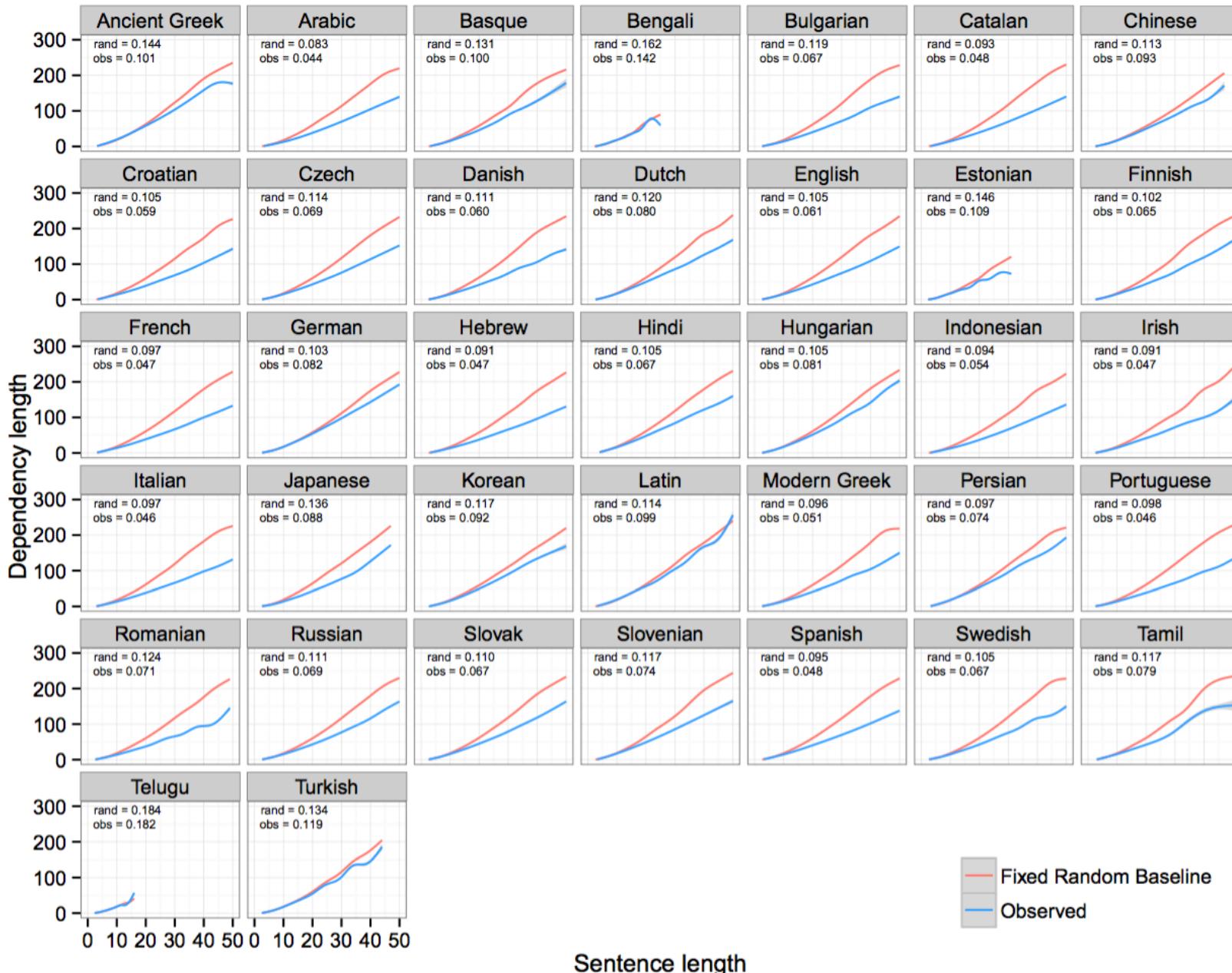
- Syntactic dependencies vary in linear distance



- Idea with long history: short dependencies preferred



# Dependency lengths are short across languages!



# Dependency lengths and the noisy channel

---

- Here: dependency length minimization can be derived from a combination of surprisal & noisy-channel theory



*Richard Futrell*

# From noisy-channel & surprisal to dependency length minimization

context

John threw the old trash sitting in the kitchen

out

# From noisy-channel & surprisal to dependency length minimization

context

John threw the old trash sitting in the kitchen

—

out

# From noisy-channel & surprisal to dependency length minimization

context

John threw the old trash sitting in the kitchen

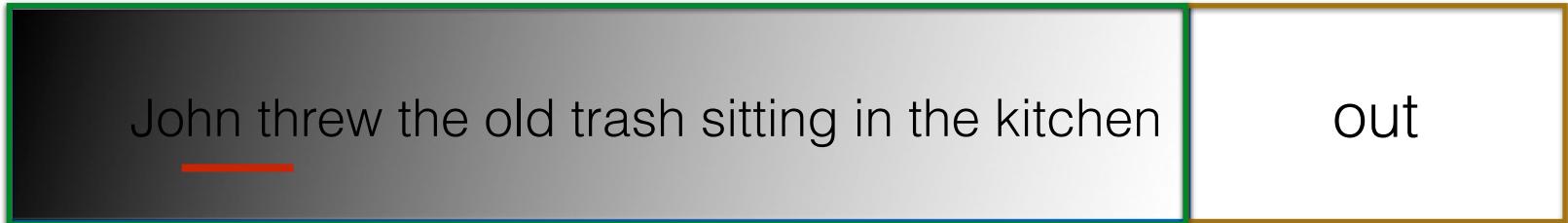
—

out

- Suppose we have an **increasing noise rate** the longer a word has been in memory.

# From noisy-channel & surprisal to dependency length minimization

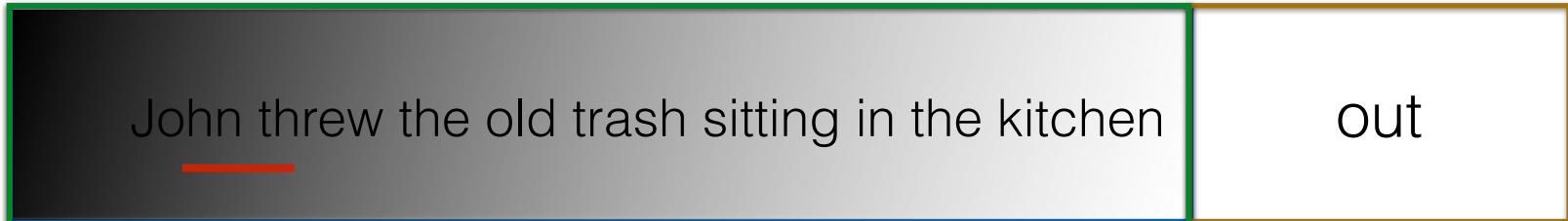
## noisy context



- Suppose we have an **increasing noise rate** the longer a word has been in memory.

# From noisy-channel & surprisal to dependency length minimization

## noisy context



- Suppose we have an **increasing noise rate** the longer a word has been in memory.
- When "threw" is far from "out", then it is less likely to reduce the surprisal of "out": more likely to be affected by noise.

# From noisy-channel & surprisal to dependency length minimization

noisy context



- Suppose we have an **increasing noise rate** the longer a word has been in memory.
- When "threw" is far from "out", then it is less likely to reduce the surprisal of "out": more likely to be affected by noise.

# From noisy-channel & surprisal to dependency length minimization

noisy context



- Suppose we have an **increasing noise rate** the longer a word has been in memory.
- When "threw" is far from "out", then it is less likely to reduce the surprisal of "out": more likely to be affected by noise.
- Noisy-context surprisal increases when **words that predict each other are far apart**.

# From noisy-channel & surprisal to dependency length minimization

noisy context



- Suppose we have an **increasing noise rate** the longer a word has been in memory.
- When "threw" is far from "out", then it is less likely to reduce the surprisal of "out": more likely to be affected by noise.
- Noisy-context surprisal increases when **words that predict each other are far apart**.
- We call this **information locality** (following Gildea & Jaeger, 2015).

# Derivation of Information Locality

## Derivation of Information Locality

- Erasure noise decreases the influence of context:

## Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

## Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John threw the trash out

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(\text{not erased}) \text{pmi}(w; w')$$

John threw the trash out

$h(\text{out})$

The image shows a large block of binary code (0s and 1s) representing the probability of the word "out" given its context. The code is enclosed in a yellow box. A brace on the right side of the box is labeled with the letter "C", indicating that the entire block represents the context of the word "out".

000011010101100101101000011100111011100  
01010101110101100101000000101011100100111  
001001101010110011000101010010011010100110  
100100100001001010011011110010010010001000  
01100011110011110001001111001001011010010  
11000010000110011000010101010011111111100  
110100110011011100000001011000111001111010  
01010001001111110110111100101001011000001  
100111100100001011110001000110000111010001  
001111010100111101110010100011100100100101  
10101100100010111010100001110011101101101  
11010100001100010011000101000100100101000  
0011011001000010010010100010100110000011  
001001101001111011101001101110001101011100  
100001010101000101101001111010110101100111  
011010100001100000110001000001111111111001  
110101000011101001101110000111000111001011  
001110111100011101011100110111111000011100  
011110011001100111010101100101111001100000  
0111100101011110011010001100000000000111100  
110000100100111110110101001101011110001100  
00111101010101101111110111100110010010111  
101010011110110110010000111100100100000000  
111110111101001010000100101000101000001101  
01010110010110100001110011110111001010101  
1110101100101000000010101110010011001001100  
10101100110001010100100110101010011010010010

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John threw the trash out

$h(\text{out}) - P(\text{John not erased}) \text{pmi}(\text{John}; \text{out})$

000011010101100101101000011100111011100  
01010101110101100101000000101011100100111  
001001101010110011000101010010011010100110  
100100100001001010011011110010010010001000  
0110001111001111000100111100100101110100100  
110000100001100110000101010100111111111100  
11010011001101110000001011000111001111010  
01010001001111110110111100101001011000001  
100111100100001011110001000110000111010001  
001111010100111101110010100011100100100101  
10101100100010111010100001110011101101101  
11010100001100010011000101000100100101000  
0011011001000010010010100010100110000011  
001001101001111011101001101110100011011100  
100000101010001011010001111010110101100111  
011010100001100000110001000001111111111001  
110101000011101001101110000111000111001011  
001110111100011101011110011011111100001110  
011110011001100111010101100101111001100000  
0111100101011110011010001100000000000111110  
110000100100111110110101001101011110001100  
00111101010101101111110111100110010010111  
101010011110110110010000111100100100000000  
111110111101001010000100101000101000001101  
010101100101101000011100111101110001010101  
111011110010100000001010111001001100100110  
1010 P(John not erased) pmi(John; out) 0010

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John threw the trash out

$$h(\text{out}) - P(\text{John not erased}) \text{ pmi}(\text{John}; \text{out}) \\ - P(\text{threw not erased}) \text{ pmi}(\text{threw}; \text{out})$$

```

000011010101100101101000011100111011100
01010101110101100101000000101011100100111
001001101010110011000101010010011010100110
100100100001001010011011110010010010001000
01100011110011110001001111001001011010010
110000100001100110000001010101001111111100
110100110011011100000001011000111001111010
0101000100111110110111100101001011000001
100111100100001011110001000110000111010001
001111010100111101110010100011100100100101
10101100100010111010100001110011101101101
1101010000110001001100010100010010010101000
0011011001000010010010100010100110000011
00100110100111101110100110100011010111000
10000101010001011010001111010110101100111
01101010000110000100010001111111111111111100
11010100001110100110001111000111001100111011
001110111100011101011100110111110000111011
0111100100110011101010110010111001100000000
0111100101011110011010001100000000000111100
110 P(threw not erased) pm(i(threw; out)) 100
0011110101010111110111100110010010111111111
1010100111101101100100001111001001000000000
111110111101001010000100101000101000001101
01010110010110100001110011110011110001010101
1110101100101000000010101110010011001001110
1010 P(John not erased) pm(i(John; out)) 0010

```

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John    threw    the    trash    out

$h(\text{out}) - P(\text{John not erased}) \text{pmi}(\text{John}; \text{out})$   
   -  $P(\text{threw not erased}) \text{pmi}(\text{threw}; \text{out})$   
   -  $P(\text{the not erased}) \text{pmi}(\text{the}; \text{out})$

000011010101100101101000011100111011100  
 01010101110101100101000000101011100100111  
 001001101010110011000101010010011010100110  
 1001001000100101001101110010010010001000  
 011000111100111100010011110010010010001000  
 1100001000110011000010101010011111111100  
 11010011001101110000001011000111001111010  
 0101000100111110110111100101001011000001  
 100111100100001011110001000110000111010001  
 0011110101001111011100101000111001001000101  
 101011001000101110101000011110011101101101  
 11010100001100010011000101000100100101000  
 001101100100001001001010100010100110000011  
 00100110100111101111010011011000110111000  
 100001101010111101111010011011000110111000  
 P(the not erased) pmi(the; out) 100111  
 01  
 11010100001110100110110000111000111001011  
 001110111100011101011110011011111000011100  
 0111100100110011101010110010111001100000000  
 0111100101011110011010001100000000000111100  
 110 P(threw not erased) pmi(threw; out) 100  
 0011110101010111101111001100100100100000000  
 1010100111101101100100001111001001000000000  
 111110111101001010000100101000101000001101  
 01010110010110100001110011110011110001010101  
 1110101100101000000010101110010011001001100  
 1010 P(John not erased) pmi(John; out) 0010

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John threw the trash out

$h(\text{out}) - P(\text{John not erased}) \text{pmi}(\text{John}; \text{out})$   
 -  $P(\text{threw not erased}) \text{pmi}(\text{threw}; \text{out})$   
 -  $P(\text{the not erased}) \text{pmi}(\text{the}; \text{out})$   
 -  $P(\text{trash not erased}) \text{pmi}(\text{trash}; \text{out})$

000011010101100101101000011100111011100  
 01010101110101100101000000101011100100111  
 001001101010110011000101010010011010100110  
 10010010001001010011011100100100100010001000  
 01100011110011110001001111001001001000100100  
 1100001000110011000010101010011111111100  
 11010011001101110000001011000111001111010  
 0101000100111110110111100101001011000001  
 100111100100001011110001000110000111010001  
 001111010100111101110010100011100100100101  
 101011001000101110101000011110011101101101  
 110101000011000100110000101000100100100101000  
 0011 P(trash not erased) pmi(trash; out) 0011  
 001001010111101101101010111001001001001000  
 1000011 P(the not erased) pmi(the; out) 100111  
 011010100011101001101100001110001110010111  
 001110111100011101011110011011111000011100  
 0111100100110011101010110010111001100000000  
 011110010101111001101000110000000000111100  
 110 P(threw not erased) pmi(threw; out) 100  
 0011110101010111110111100110010010010010111  
 1010100111101101100100001111001001000000000  
 111110111101001010000100101000101000001101  
 0101011001011101000011100111100101100010101  
 1110101100101000000010101110010011001001100  
 1010 P(John not erased) pmi(John; out) 0010

# Derivation of Information Locality

- Erasure noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

John threw the trash out

$$h(\text{out}) - P(\text{John not erased}) \text{pmi}(\text{John}; \text{out})$$

- $P(\text{threw not erased}) \text{pmi}(\text{threw}; \text{out})$
- $P(\text{the not erased}) \text{pmi}(\text{the}; \text{out})$
- $P(\text{trash not erased}) \text{pmi}(\text{trash}; \text{out})$

000011010101100101101000011100111011100  
 01010101110101100101000000101011100100111  
 001001101010110011000101010010011010100110  
 1001001000100101001101110010010010001000  
 01100011110011110001001111001001011010010  
 1100001000110011000010101010011111111100  
 11010011001101110000001011000111001111010  
 0101000100111110110111100101001011000001  
 100111100100001011110001000110000111010001  
 001111010100111101110010100011100100100101  
 101011001000101110101000011110011101101101  
 110101000011000100110000101000100100100100  
 0011 P(trash not erased) pmi(trash; out) 0011  
 0010010101111011011010101101001001001000  
 1000011 P(the not erased) pmi(the; out) 100111  
 01101010001110100110111000111000111001011  
 00111011110001110101111001101111000011100  
 011110010011001101010110010111001100000000  
 011110010101111001101000110000000000111100  
 110 P(threw not erased) pmi(threw; out) 100  
 001111010101011111011110011001001001001001  
 101010011110110110010000111100100100000000  
 111110111101001010000100101000101000001101  
 010101100101101000011100111100101100010101  
 111010110010100000001010111001001100100110  
 1010 P(John not erased) pmi(John; out) 0010

# Derivation of Information Locality

- Noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$


  
threw      out

$h(\text{out}) - P(\text{threw not erased}) \text{pmi}(\text{threw}; \text{out})$

```

000011010101100101101000011100111011100
01010101110101100101000000101011100100111
001001101010110011000101010010011010100110
100100100001001010011011110010010010001000
01100011110011110001001111001001011010010
1100001000011001100010101010011111111100
110100110011011100000001011000111001111010
01010001001111110110111100101001011000001
100111100100001011110001000110000111010001
001111010100111101110010100011100100100101
101011001000101110101000011110011101101101
11010100001100010011000101000100100100101000
0011011001000100111101001010001010001000111
001001101001111011101001100011010111000111
100001010100111101011001100011101011000111
0110101000011100000110001000001111111111001
110101000011101001110000111000111000111001011
001110111100011101011111001101111100001110
011110011001100111010101100101110001100000000
0111010001001111101101000111100100100000000
110000100100111110110100011010111100011001000
00111101010101011111101111000110010010010111
1010100111101101100100001111001001000000000
111110111101001010000100101000101000001101
01010110010111010000111001111001110001010101
1110101100101000000010101110010011001001000110
101011001100010101001001101010100110100100010

```

}
C

# Derivation of Information Locality

- Noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

threw

out

$h(\text{out}) - P(\text{threw not erased}) \text{ pmi}(\text{threw}; \text{out})$

0000110101011001011010000111001111011100  
010101011110101100101000000010101110010011  
0010011010101100110001010100100110100110  
100100100001001010011011110010010001000  
0110001110011110001001111001001011010010  
1100001000011001100001010101001111111100  
110100110011011110000001011000111001111010  
01010001001111101101111100101001011000001  
100111100100001011110001000110000111010001  
001111010100111101110010100011100100100101  
101011001000101110101000011110011101101101  
110101000001100010011000101000100100101000  
001101100100001001001010100010100110000011  
001001101001111011110100110100011010111000  
1000001010101000101101001111010110101100111  
01101010000110000011000100000111111111001  
110101000011101001101110000111000111001011  
00111011110001110101111001101111100001110  
011110011001100111010101100101111001100000  
01111001010111100110100011000000000111110  
11000001001001111011010100110101110001100  
00111101010101101111110111100110010010111  
101010011111011010010001111001001000000000  
111110111101001010000100101000101000001101  
01  
1110101010100100000001011001011001010101010  
0101011010001010100100110101001101010010010

C } C

# Derivation of Information Locality

- Noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

threw out

$h(\text{out}) - P(\text{threw not erased}) \text{ pmi}(\text{threw}; \text{out})$

- When context items are far, their cost-reducing influence decreases.

```

0000011010101011001011010000111001111011100
010101011110101100101000000010101110010011
001001101010110011000101010010011010100110
1001001000010010101111001001000100010001000
011000111100111100010011110010010111010010
11000010000110011000010101010011111111100
110100110011011100000001011000111001111010
01010001001111101101111100101001011000001
1001111100100001011110001000110000111010001
001111010100111101110010100011100100100101
101011001000101110101000011110011101101101
110101000001100010011000101000100100101000
001101100100001001001010100010100110000011
001001101001111011110100110100011010111000
1000001010101000101101001111010110101100111
011010100001100000110001000001111111111001
1101010000111010011011100001110001110010111
001110111100011101011110011011111100001110
0111100110011001110101011000101111001100000
0111100101011110011001000110000000000111110
11000001001001111101101001001101011110001100
001111010101011011111110111100110010010111
1010100111110110110100001111001001000000000
1111101111101001010000100101000101000001101
0101010111110110110100001010111010100101101
P(threw not erased) pmi(threw, out)
111010110100101010000101011001001000101001
1010110011000101010001010100100110100110100

```

# Derivation of Information Locality

- Noise decreases the influence of context:

$$C(w|\text{context}) \approx h(w) - \sum_{w' \in \text{context}} P(w' \text{ not erased}) \text{pmi}(w; w')$$

**threw**      **out**

$h(\text{out}) - P(\text{threw} \text{ not erased}) \text{ pmi}(\text{threw}; \text{out})$

- When context items are far, their cost-reducing influence decreases.
    - Similar to the concept of decay in cue effectiveness  
(Qian & Jaeger, 2012)

```

000001101010101100010110100000111001111011100
0101010111101011001010000000010101110010011
001001101010110011000101010010011010100110
1001001000010010100110111100100100010001000
011000111001111000100111100100101101001010
11000010000110011000010101010011111111100
110100110011011100000001011000111001111010
010100010011111101101111100101001011000001
100111100100001011110001000110000111010001
001111010100111101110010100011100100100101
101011001000101110101000011110011101101101
110101000001100010011000101000100100101000
001101100100001001001010100010100110000011
001001101001111011110100110100011010111000
1000001010101000101101001111010110101100111
011010100001100000110001000001111111111001
110101000011101001101110000111000111001011
001110111100011101011110011011111100001110
011110001100110011010101100101111001100000
011110010101011110010100011000000000111110
1100000100100111110101101001101011110001100
001111010101011011111101111001100010010111
101010011111010110010000111100100100000000
111110111101001010000100101000101000001101
010101011110100010010010010100010100010101
111010110010010000000010101100101010010101
010101100110001010100100110101010011010010010

```

C }

*P(threw not erased) pmi(threw, out)*

# Information Locality

## Information Locality

- **Information locality:** prediction of processing difficulty when words that predict each other (have high mutual information) are far apart.

## Information Locality

- **Information locality:** prediction of processing difficulty when words that predict each other (have high mutual information) are far apart.
- How does this relate to **dependency locality?**

## Information Locality

- **Information locality:** prediction of processing difficulty when words that predict each other (have high mutual information) are far apart.
- How does this relate to **dependency locality?**
- Hypothesis: **Words in syntactic dependencies have high mutual information.**

# Information Locality

- **Information locality:** prediction of processing difficulty when words that predict each other (have high mutual information) are far apart.
- How does this relate to **dependency locality?**
- Hypothesis: **Words in syntactic dependencies have high mutual information.**
  - If this is true, then we can see dependency locality effects as a subset of information locality effects.

# Information Locality

- **Information locality:** prediction of processing difficulty when words that predict each other (have high mutual information) are far apart.
- How does this relate to **dependency locality?**
- Hypothesis: **Words in syntactic dependencies have high mutual information.**
  - If this is true, then we can see dependency locality effects as a subset of information locality effects.
- We will show that the hypothesis is true in dependency corpora.

# Do Dependencies Have High Mutual Information?

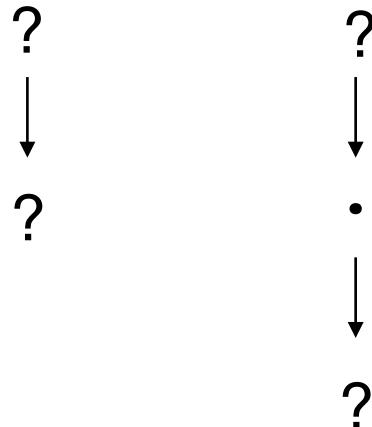
# Do Dependencies Have High Mutual Information?

?

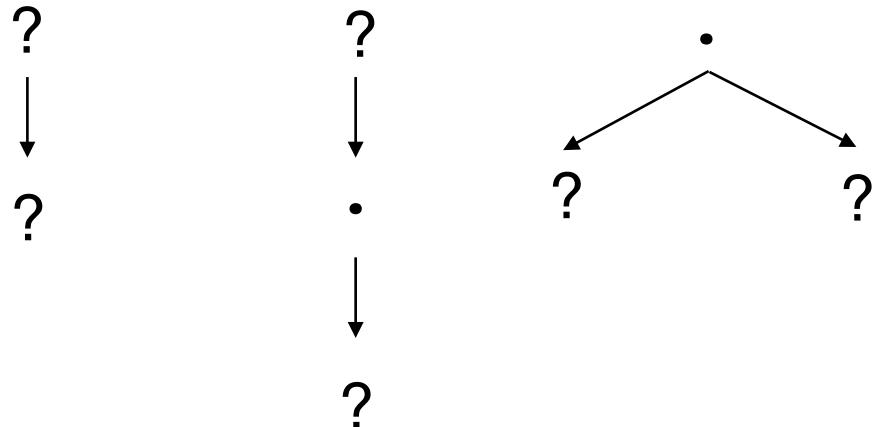


?

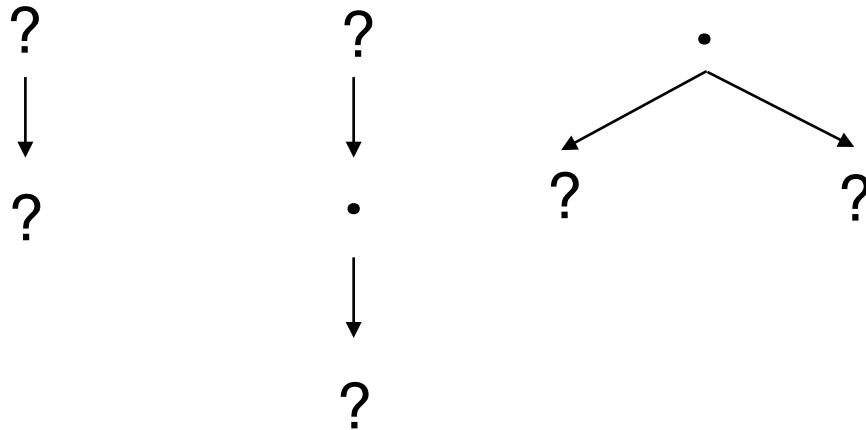
# Do Dependencies Have High Mutual Information?



# Do Dependencies Have High Mutual Information?

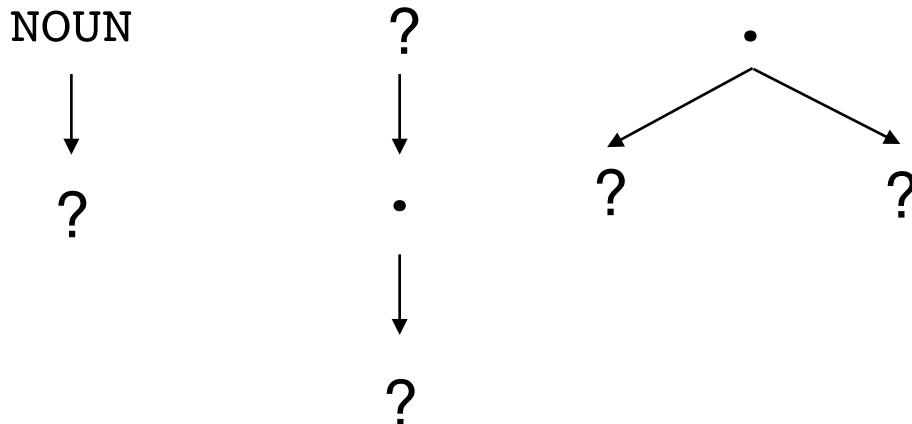


# Do Dependencies Have High Mutual Information?



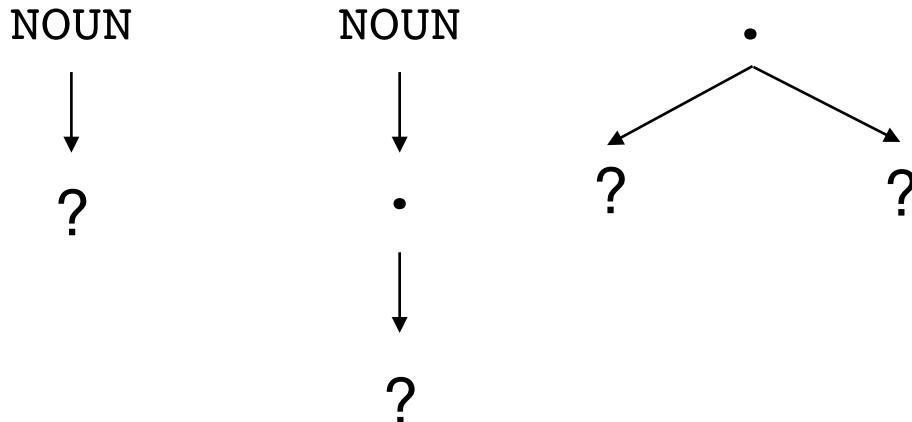
- We calculated mutual information values over part-of-speech tags for pairs of words in the UD corpora.

# Do Dependencies Have High Mutual Information?



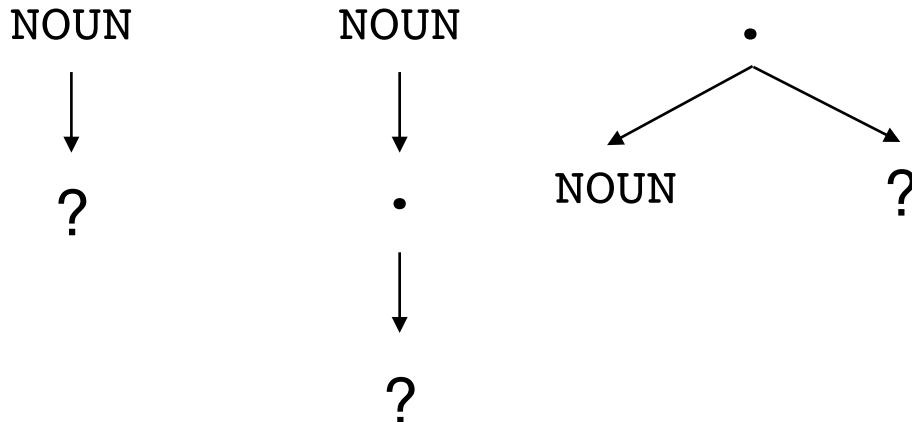
- We calculated mutual information values over part-of-speech tags for pairs of words in the UD corpora.

# Do Dependencies Have High Mutual Information?



- We calculated mutual information values over part-of-speech tags for pairs of words in the UD corpora.

# Do Dependencies Have High Mutual Information?

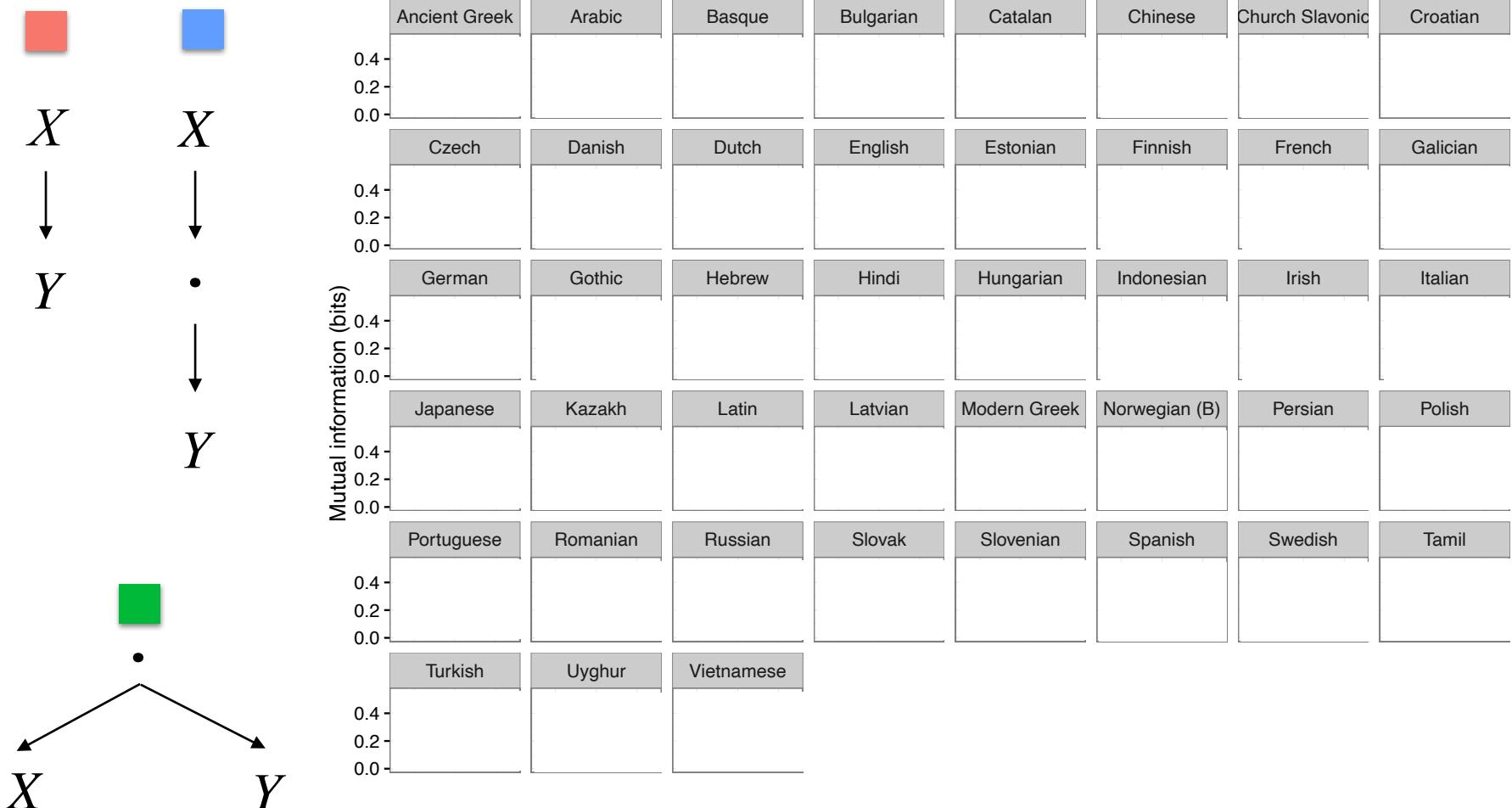


- We calculated mutual information values over part-of-speech tags for pairs of words in the UD corpora.

# Do Dependencies Have High Mutual Information?



# Do Dependencies Have High Mutual Information?



# Do Dependencies Have High Mutual Information?



# Comprehension as exploration of input

---

- Broader ongoing goal: develop eye-movement control model integrating the insights discussed thus far:
  - Probabilistic linguistic knowledge
  - Uncertain input representations
  - Principles of adaptive, rational action
- *Reinforcement learning* is an attractive tool for this

# A rational reader

---

- Very simple framework:
  - Start w/ prior expectations for text (linguistic knowledge)
  - Move eyes to get perceptual input
  - Update beliefs about text as visual arrives (Bayes' Rule)
- Add to that:
  - Set of *actions* the reader can take in discrete time
  - A *behavior policy*: how the model decides between actions

# A first-cut behavior policy

---

# A first-cut behavior policy

---

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters:  $\alpha$  and  $\beta$
- Define *confidence* in a character position as the probability of the most likely character

# A first-cut behavior policy

---

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters:  $\alpha$  and  $\beta$
- Define *confidence* in a character position as the probability of the most likely character

*From the closet, she pulled out a \*acket for the upcoming game*

# A first-cut behavior policy

---

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters:  $\alpha$  and  $\beta$
- Define *confidence* in a character position as the probability of the most likely character

*From the closet, she pulled out a \*acket for the upcoming game*

$$P(\text{jacket})=0.38$$

$$P(\text{racket})=0.59$$

$$P(\text{packet})=0.02$$

...

# A first-cut behavior policy

---

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters:  $\alpha$  and  $\beta$
- Define *confidence* in a character position as the probability of the most likely character

*From the closet, she pulled out a \*acket for the upcoming game*

Confidence=0.59

$$P(\text{jacket}) = 0.38$$

$$P(\text{racket}) = 0.59$$

$$P(\text{packet}) = 0.02$$

...

# A first-cut behavior policy

---

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters:  $\alpha$  and  $\beta$
- Define *confidence* in a character position as the probability of the most likely character

*From the closet, she pulled out a \*acket for the upcoming game*

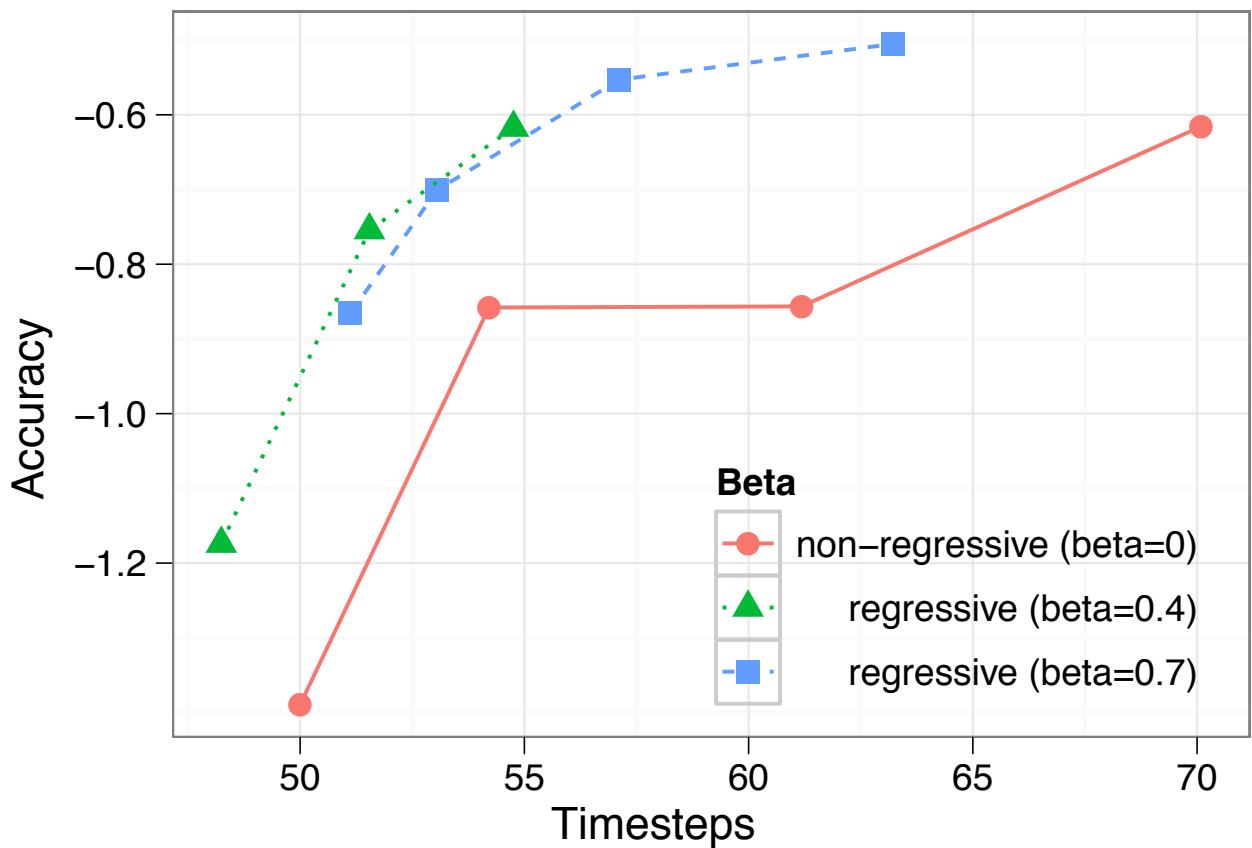
Confidence=0.59

$P(\text{jacket})=0.38$   
 $P(\text{racket})=0.59$   
 $P(\text{packet})=0.02$   
...

- Move left to right, bringing up confidence in each character position until it reaches  $\alpha$
- If confidence in a previous character position drops below  $\beta$ , regress to it
- Finish reading when you're confident in everything

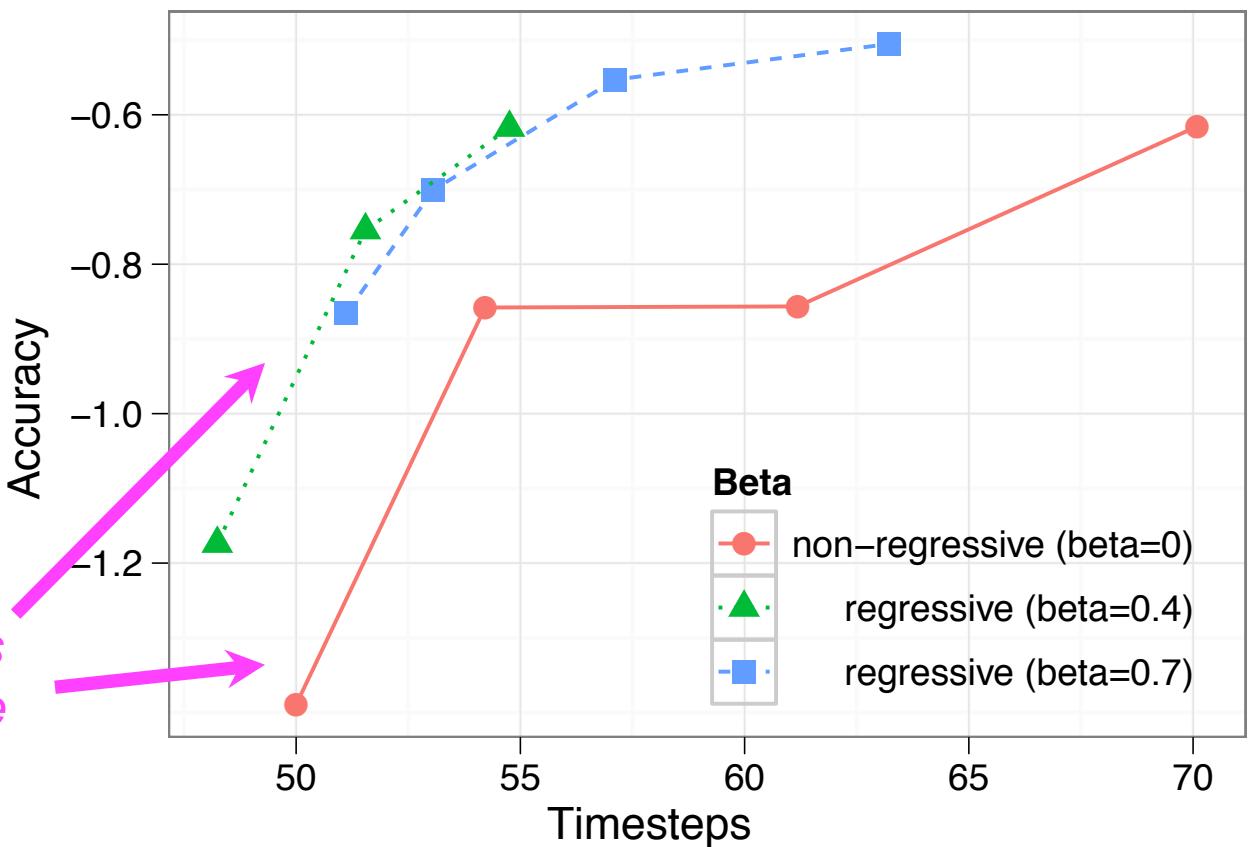
# (Non)-regressive policies

- Non-regressive policies have  $\beta=0$
- Hypothesis: non-regressive policies strictly dominated
- Test: estimate speed and accuracy of various policies on reading the the Schilling et al. (1998) corpus



# (Non)-regressive policies

- Non-regressive policies have  $\beta=0$
- Hypothesis: non-regressive policies strictly dominated
- Test: estimate speed and accuracy of various policies on reading the the Schilling et al. (1998) corpus



# Goal-based adaptation

---

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

$\gamma$	$\alpha$	$\beta$
0.025		
0.1		
0.4		

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

$\gamma$	$\alpha$	$\beta$
0.025	0.90	0.99
0.1	0.36	0.80
0.4	0.18	0.38

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

$\gamma$	$\alpha$	$\beta$	Timesteps	Accuracy
0.025	0.90	0.99		
0.1	0.36	0.80		
0.4	0.18	0.38		

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

$\gamma$	$\alpha$	$\beta$	Timesteps	Accuracy
0.025	0.90	0.99	41.2	P(correct)=0.98
0.1	0.36	0.80	25.8	P(correct)=0.41
0.4	0.18	0.38	16.4	P(correct)=0.01

# Goal-based adaptation

---

- Open frontier: modeling the adaptation of eye movements to specific reader goals
- We set a *reward function*: relative value  $\gamma$  of speed (finish reading in  $T$  timesteps) versus accuracy (guess correct sentence with probability  $L$ )
- PEGASUS simplex-based optimization (Ng & Jordan, 2000)

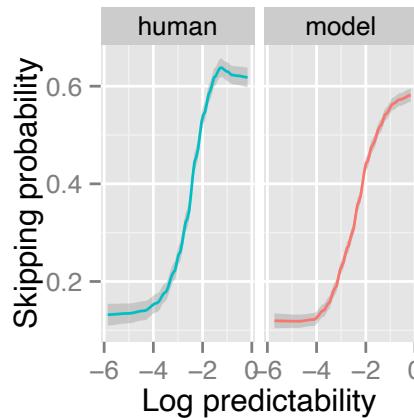
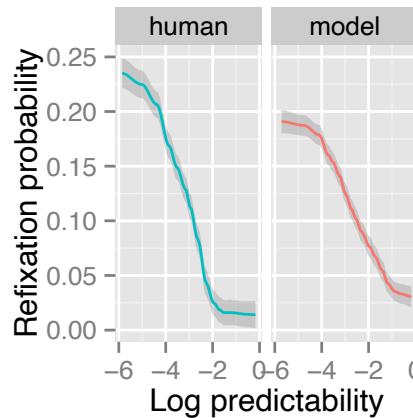
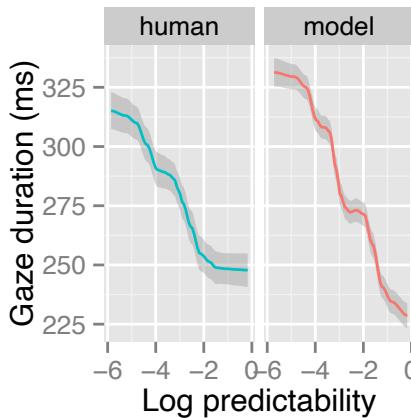
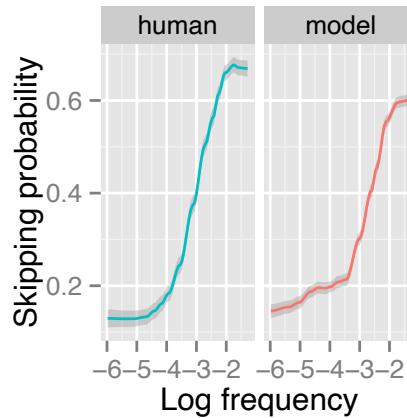
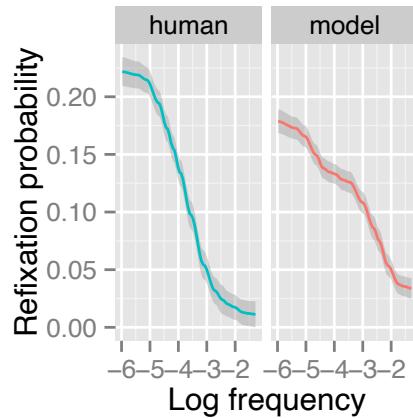
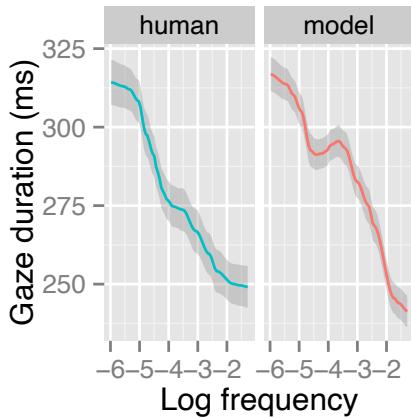
$\gamma$	$\alpha$	$\beta$	Timesteps	Accuracy
0.025	0.90	0.99	41.2	$P(\text{correct})=0.98$
0.1	0.36	0.80	25.8	$P(\text{correct})=0.41$
0.4	0.18	0.38	16.4	$P(\text{correct})=0.01$

- The method works, and gives intuitive results

# Empirical match with human reading

- Benchmark measures in eye-movement modeling:

frequency



predicts size and  
shape of all effects

predictability

Bicknell & Levy (2012)

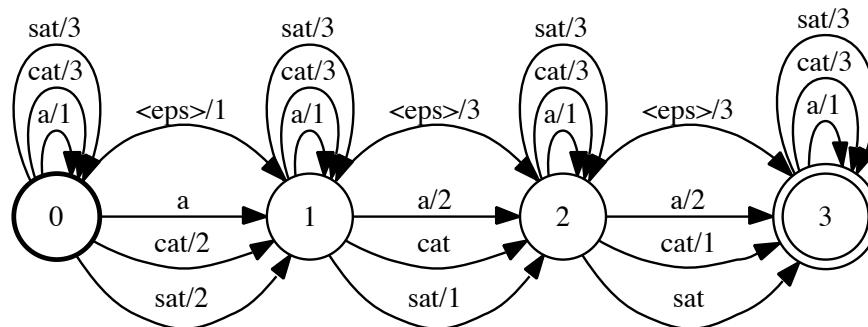
# Success at empirical benchmarks

---

- Other models (E-Z Reader, SWIFT) get these too, but *stipulate* rel'nship between word properties & “processing rate”
- We *derive* these relationships from simple principles of noisy-channel perception and rational action

# Noisy-channel processing: summary

- Noisy-channel models help us understand
  - Basic capabilities of human language comprehension
  - Outstanding puzzles in syntactic processing
- These models open up a rich typology of new sentence processing effects
- There is growing evidence for these effects
- These models pose new theoretical opportunities and architectural challenges for the study of human linguistic cognition



# References I

-  Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the 34th annual meeting of the Cognitive Science Society* (pp. 1320–1325).
-  Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: John Wiley & Sons.
-  Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1168–1178). Uppsala, Sweden.
-  Bicknell, K., & Levy, R. (2012a). Why long words take longer to read: The role of uncertainty about word length. In *Proceedings of the 3rd annual workshop on Cognitive Modeling and Computational Linguistics* (pp. 21–30).

## References II

-  Bicknell, K., & Levy, R. (2012b). Word predictability and frequency effects in a rational model of reading. In *Proceedings of the 34th annual meeting of the Cognitive Science Society* (pp. 126–131). Sapporo, Japan.
-  Booth, T. L. (1969). Probabilistic representation of formal languages. In *IEEE conference record of the 1969 tenth annual symposium on switching and automata theory* (pp. 74–81).
-  Crocker, M., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6), 647–669.
-  Fodor, J. D. (2002). Psycholinguistics cannot escape prosody. In *Proceedings of the speech prosody conference*.

# References IV

-  Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 688–698).
-  Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33), 10336–10341.
-  Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1–76.
-  Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O’Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

# References V

-  Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056.
-  Gibson, E., & Thomas, J. (1999). The perception of complex ungrammatical sentences as grammatical. *Language & Cognitive Processes*, 14(3), 225–248.
-  Gildea, D., & Jaeger, T. F. (2015). Human languages order information efficiently. *CoRR*, abs/1510.02823. arXiv: 1510.02823
-  Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the annual meeting of the association for computational linguistics*, Prague, Czech Republic.
-  Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34, 286–310.

# References VI

-  Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, Pennsylvania.
-  Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4), 609–642.
-  Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.
-  Itti, L., & Baldi, P. (2005). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*.
-  Jelinek, F., & Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context free grammars. *Computational Linguistics*, 17(3), 315–323.

## References VII

-  Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4), 613–632.
-  Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of acl*.
-  Levy, R. (2008a). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Waikiki, Honolulu.
-  Levy, R. (2008b). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

# References VIII

-  Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1055–1065).
-  Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). Hove: Psychology Press.
-  Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086–21090.
-  MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692–715.

# References IX

-  Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2), 269–311.
-  Park, Y. A., & Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the 10th annual meeting of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) conference* (pp. 335–343). Boulder, Colorado, USA.
-  Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proceedings of the 38th annual meeting of the Cognitive Science Society* (pp. 378–383).
-  Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57, 348–379.

# References X

-  Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.
-  Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 33(3), 550–569.
-  Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2), 165–201.
-  Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4), 355–370.
-  Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.

# References XI

- 
- Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010).  
Short-term forgetting in sentence comprehension:  
Crosslinguistic evidence from verb-final structures.  
*Language & Cognitive Processes*, 25(4), 533–567.