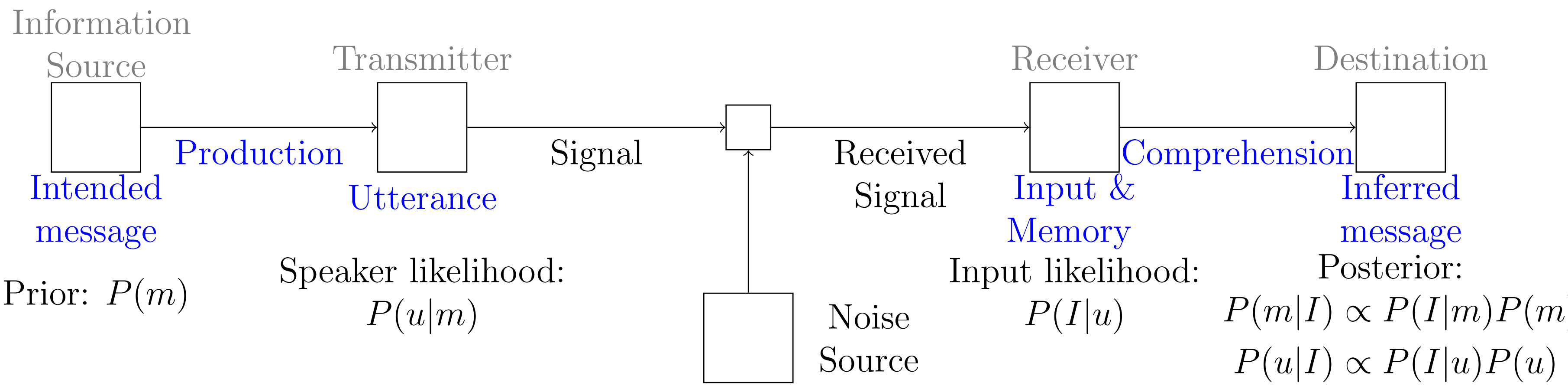
Noisy-channel sentence comprehension theory



Roger Levy

9.19: Computational Psycholinguistics

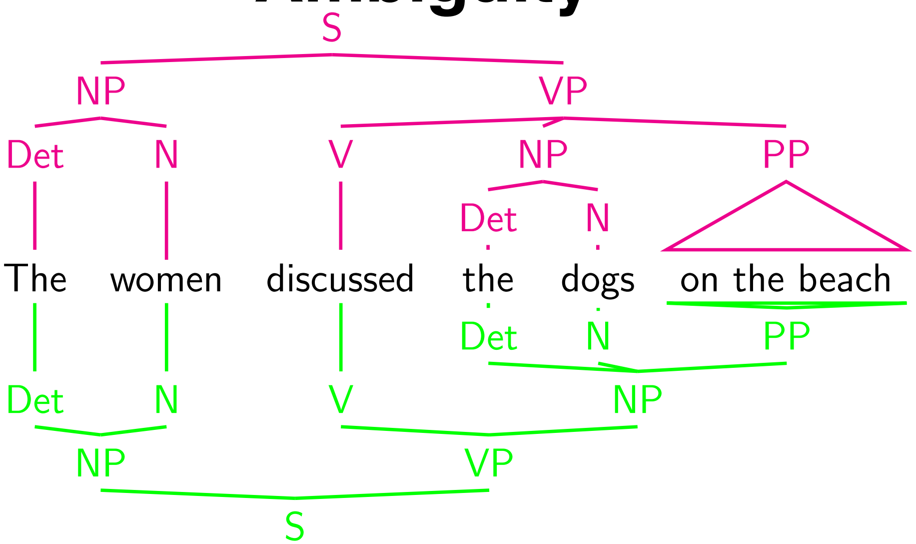
13 November 2023

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

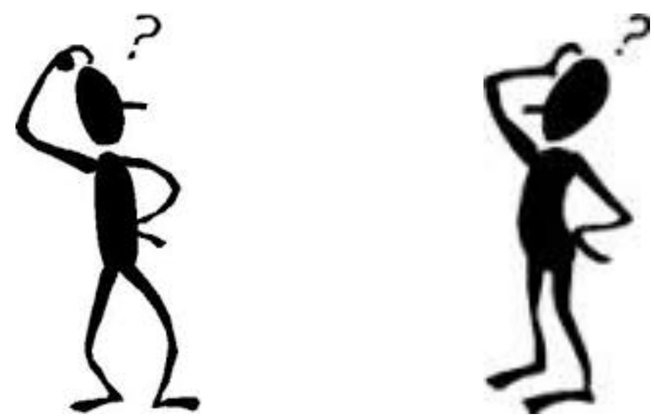
Ambiguity



Environmental noise

Memory Limitations

Incomplete knowledge of one’s interlocutors



Rational analysis

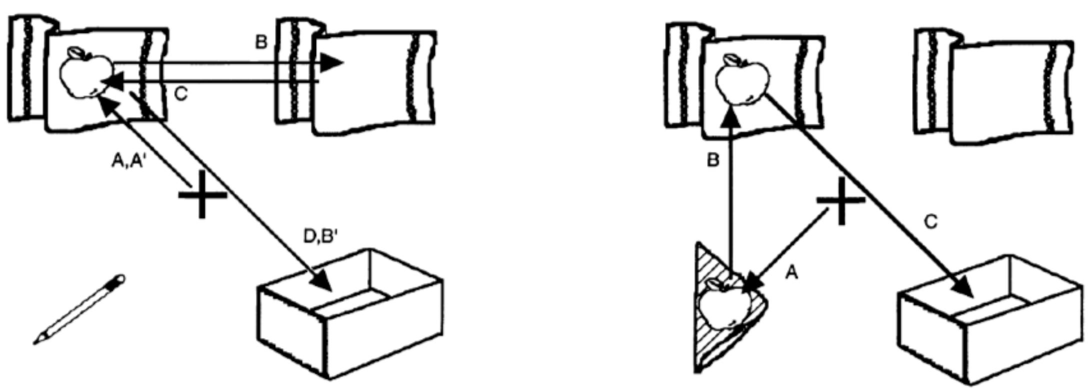
- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively

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

(Anderson, 1990, 1991)

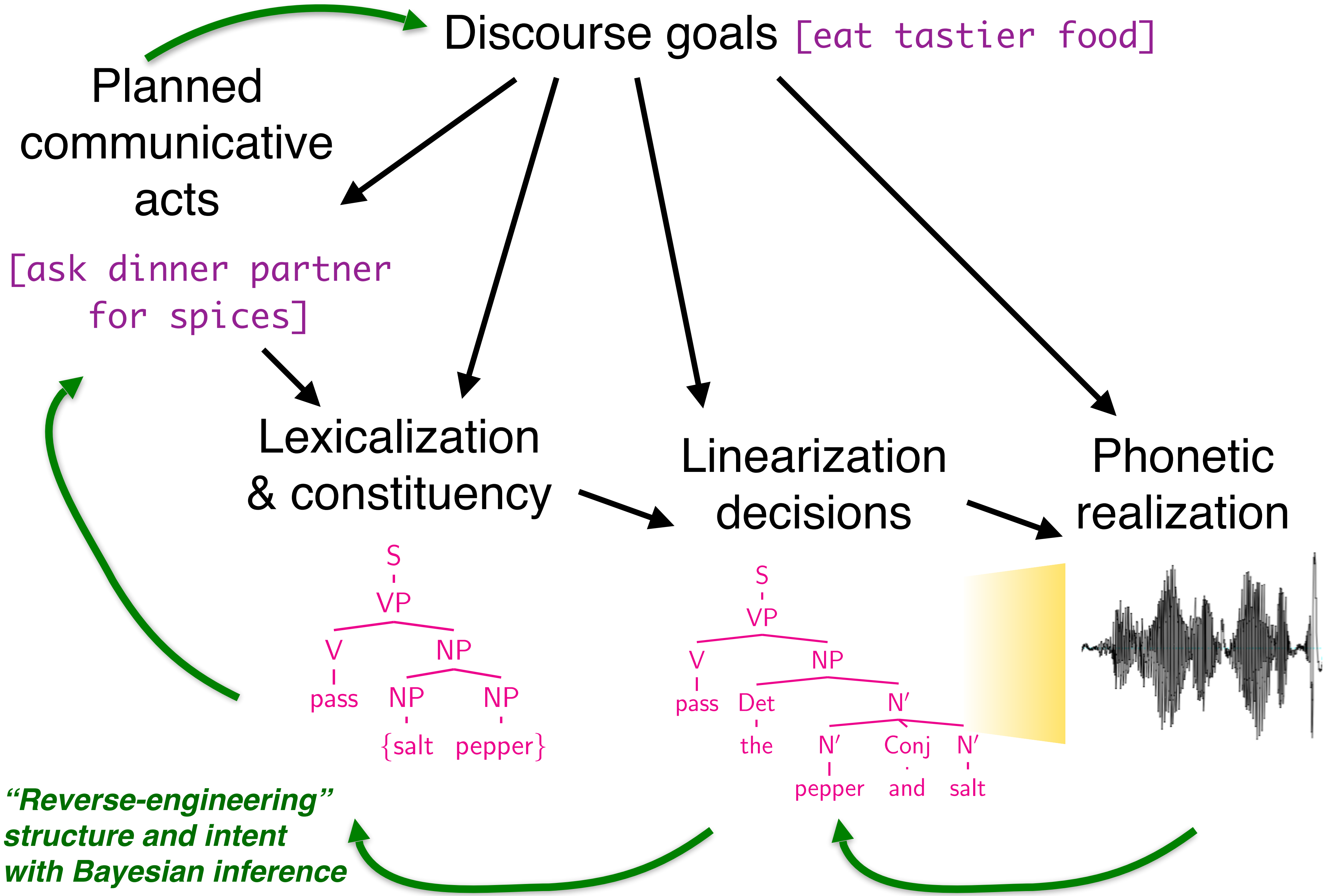
Efficient comprehension as rational, goal-driven

- 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



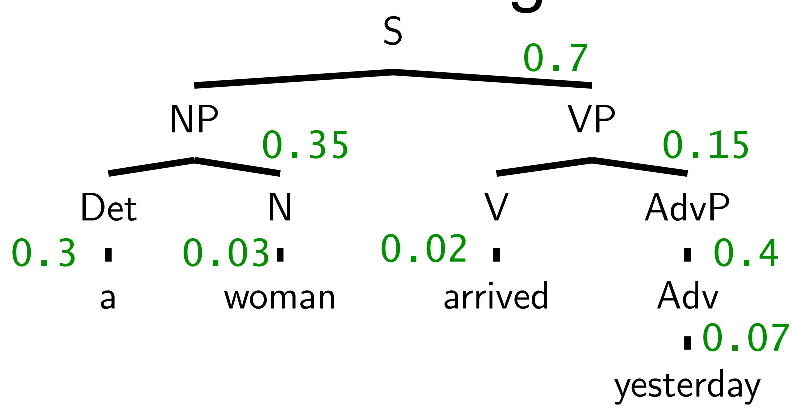
Surprisal summary: psycholinguistic evidence

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



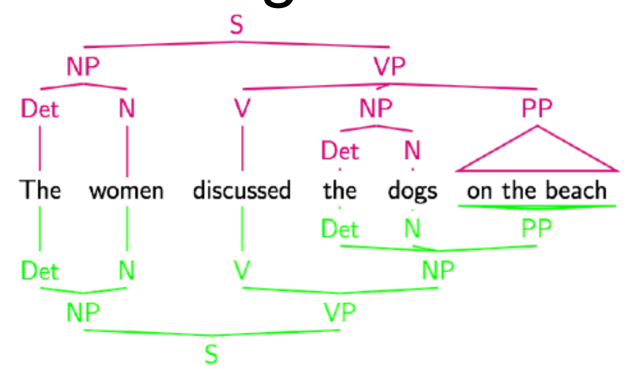
P(T) = 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07
= $1.85 \cdot 10^{-7}$

- Surprisal

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

$$\left[\approx \log \frac{1}{P(w_i|w_{1\ldots i-1})}\right]$$

- Global disambiguation

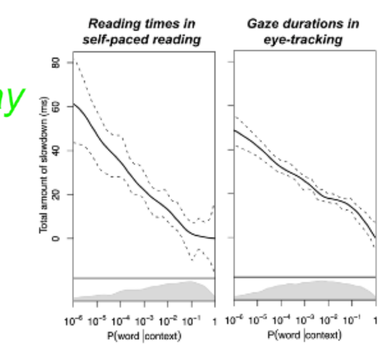


- Garden-pathing

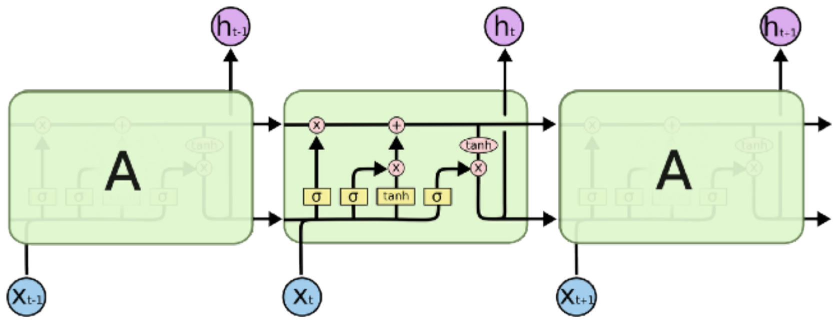
When the dog scratched the vet $\boxed{\textit{removed}}$ the muzzle.

- Prediction & reading times

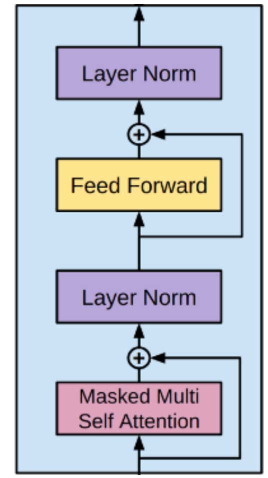
my brother came inside to…
the children went outside to…
play



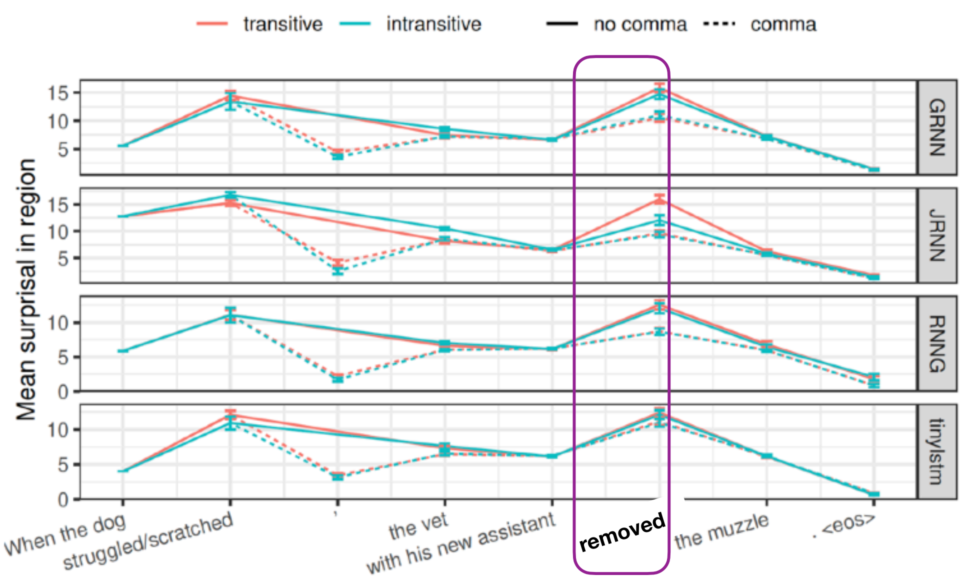
Syntax-like surprisal from deep-learning models



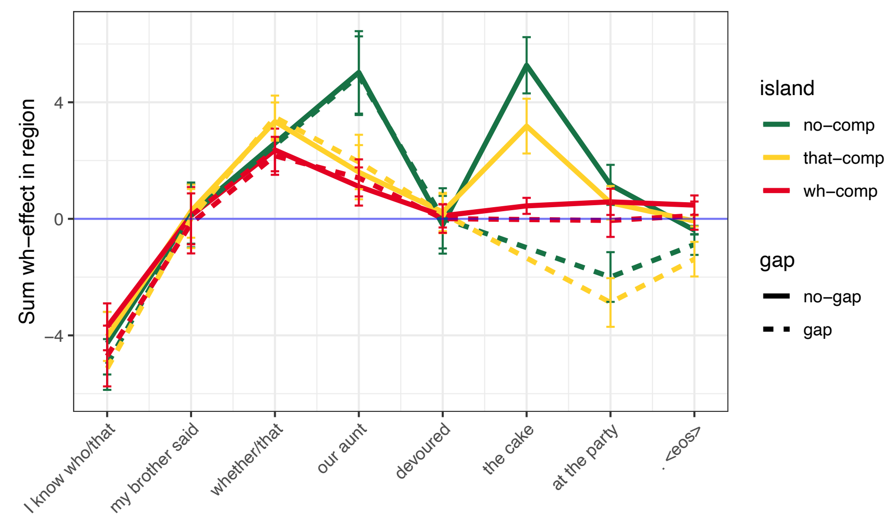
(Elman, 1990; Hochreiter & Schmidhuber, 1997)



(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

- Try to understand this sentence:

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

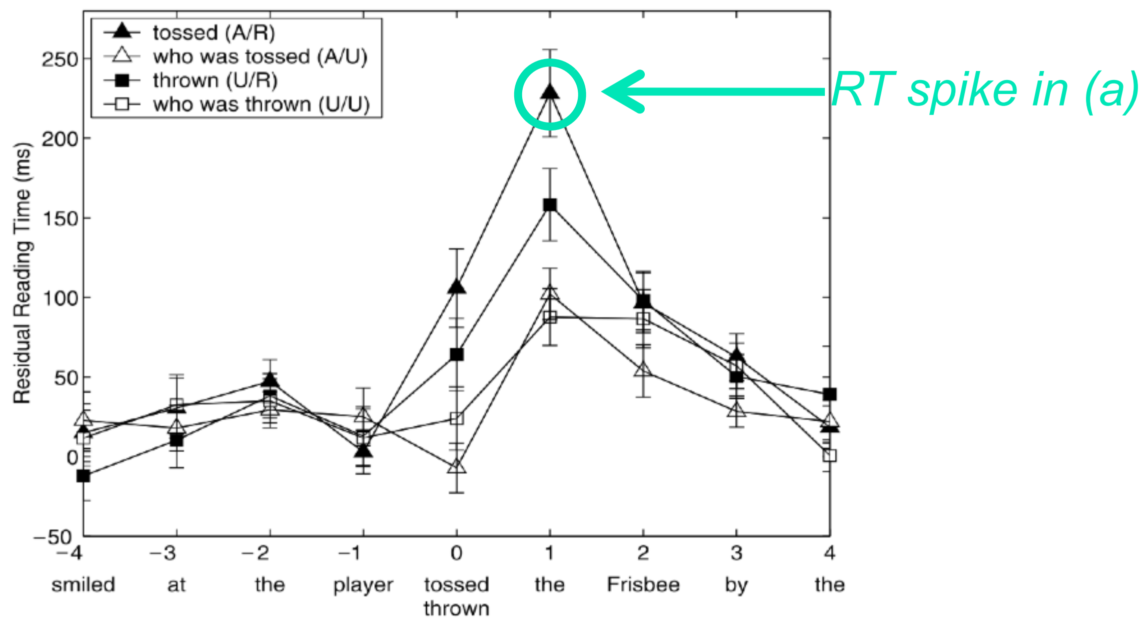
…and contrast this with:

  (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)

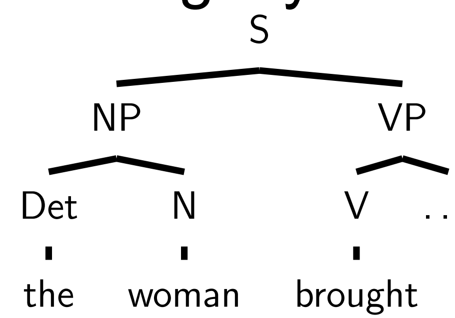


Tabor et al. (2004, JML)

Why is *tossed/thrown* interesting?

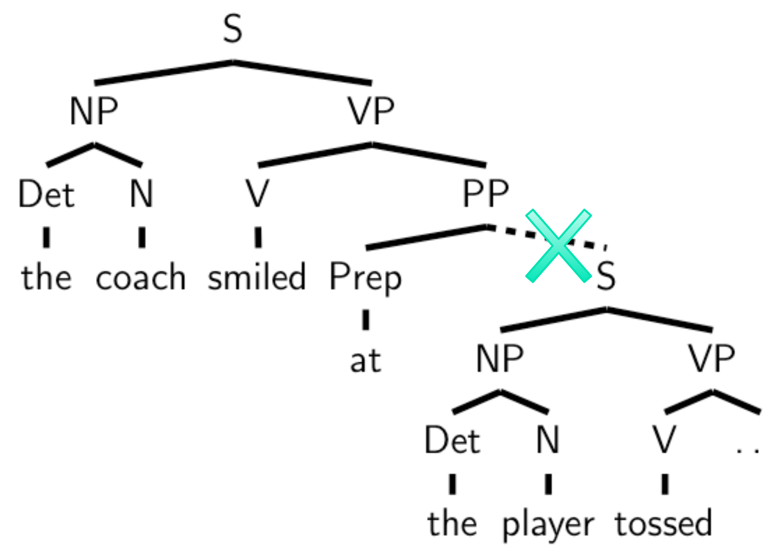
- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - *The woman brought the sandwich…tripped*

verb?
participle?



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

verb?
participle?



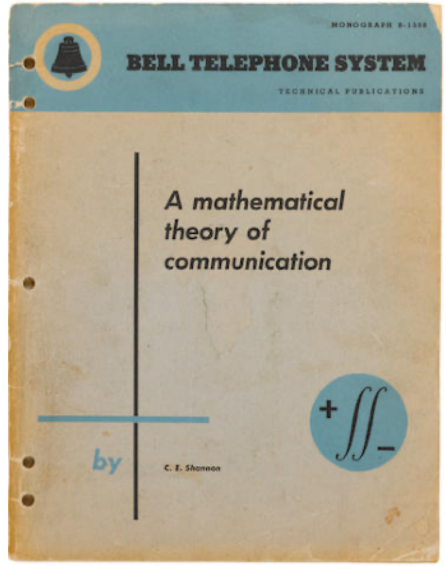
- *A challenge for rational models:* ***failure to condition on relevant context***

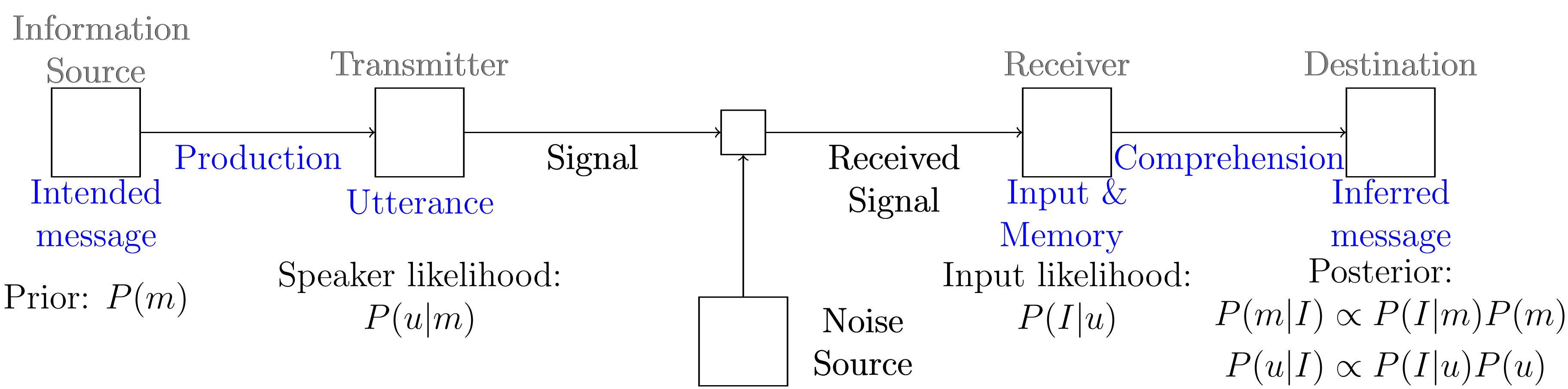
Uncertain input in language comprehension

- 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)





(Levy, 2008; Gibson et al., 2013)

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 *I*:

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

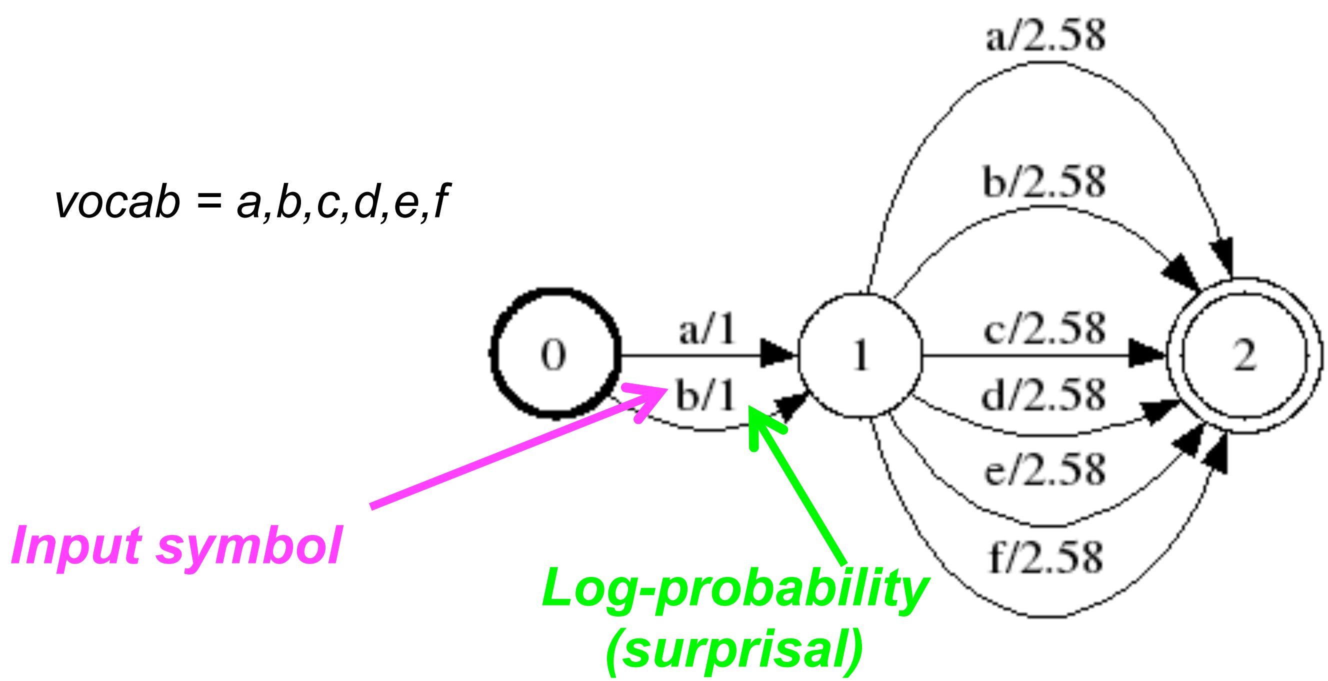
$$\propto Q(\mathbf{w},\mathbf{w}^*)$$

Levy (2008, EMNLP)

Representing noisy input

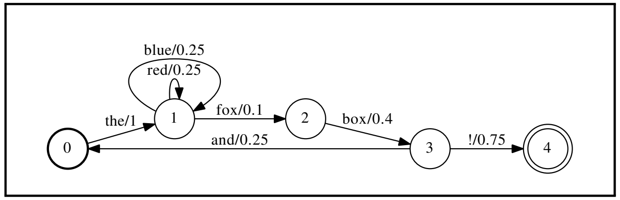
- How can we represent the type of noisy input generated by a word sequence?
- *Probabilistic finite-state automata* (pFSAs; Mohri, 1997) *are a good model*

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;
- Σ is a finite set of terminal symbols;
- $F \subseteq Q$ is the set of FINAL STATES;
- Δ is a finite set of TRANSITIONS each of the form $q \overset{i}{\rightsquigarrow} q'$, meaning that "if you are in state q and see symbol i you can consume it and move to state q'";
- λ is a function mapping transitions to real numbers (weights);
- ρ is a function mapping final states to real numbers (weights).

Weighted finite-state automata (2)

- Q is a finite set of STATES $q_0 q_1 \ldots q_N$, with q_0 the designated START STATE;
- Σ is a finite set of terminal symbols;
- $F \subseteq Q$ is the set of FINAL STATES;
- Δ is a finite set of TRANSITIONS each of the form $q \overset{i}{\leadsto} q'$, meaning that "if you are in state q and see symbol i you can consume it and move to state q'";
- λ is a function mapping transitions to real numbers (weights);
- ρ is a function mapping final states to real numbers (weights).

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

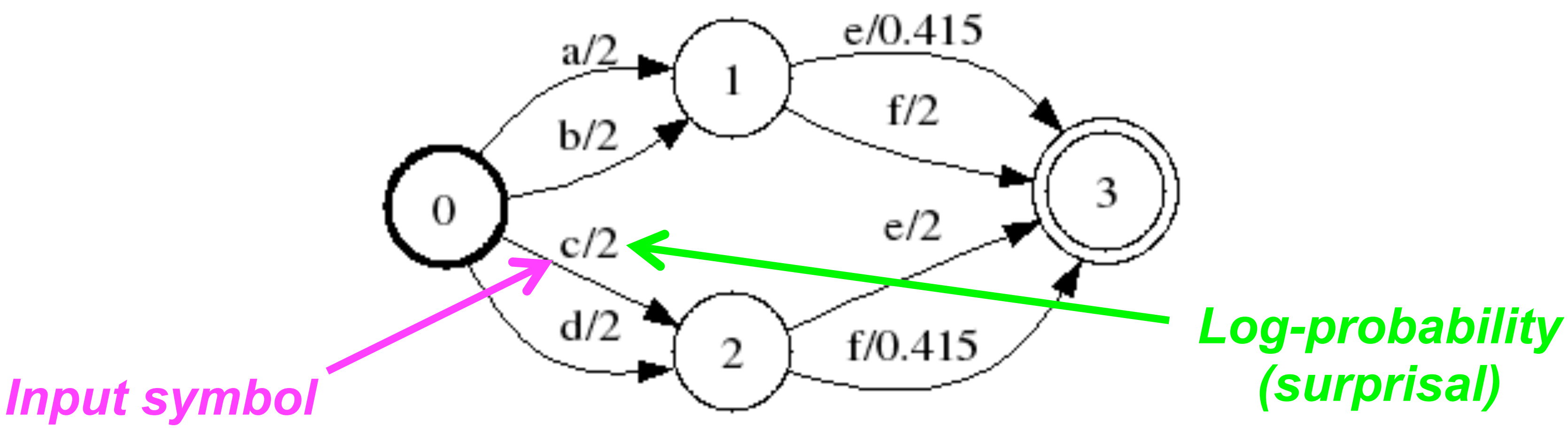
$$q_0 \underset{1}{\overset{w_1}{\leadsto}} \underset{2}{\overset{w_2}{\leadsto}} \ldots \underset{N-1}{\overset{w_{N-1}}{\leadsto}} \underset{N}{\overset{w_N}{\leadsto}} q^*$$

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

$$P(q_0 \underset{1}{\overset{w_1}{\leadsto}} \underset{2}{\overset{w_2}{\leadsto}} \ldots \underset{N-1}{\overset{w_{N-1}}{\leadsto}} \underset{N}{\overset{w_N}{\leadsto}} q^*) = \rho(q^*) \prod_{i=1}^{N} \lambda(\underset{i}{\leadsto}) \tag{1}$$

Probabilistic Linguistic Knowledge

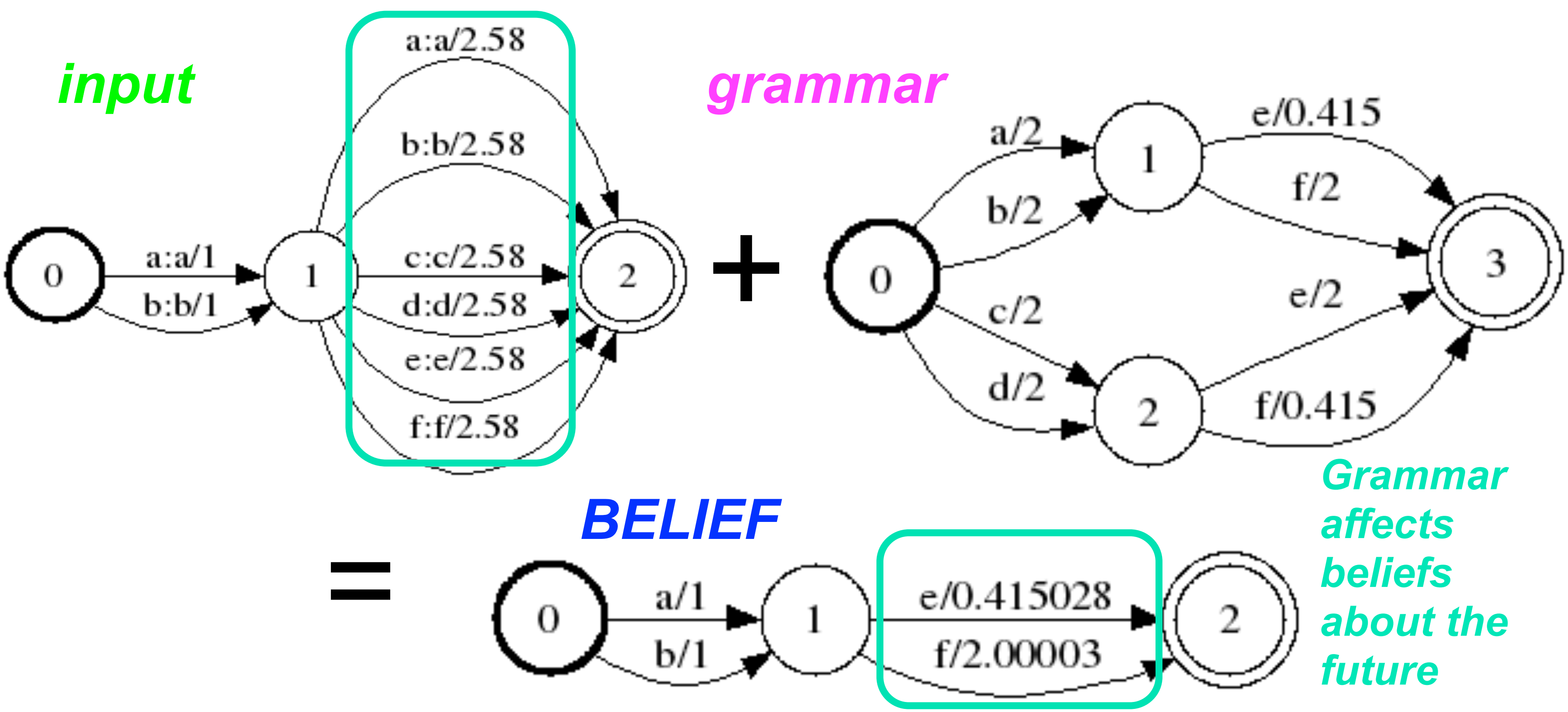
- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



- 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

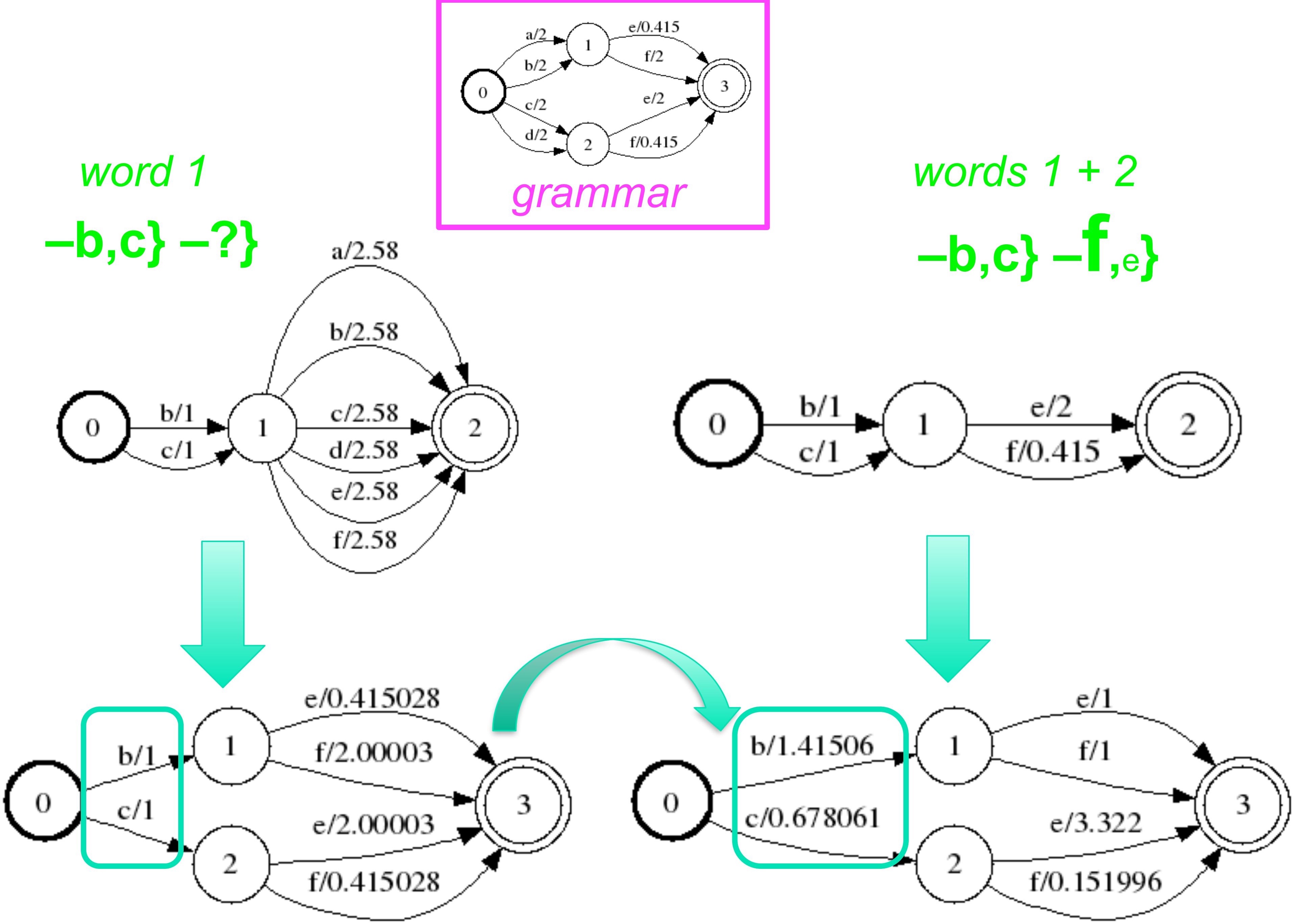
Combining grammar & uncertain input

- 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*



Revising beliefs about the past

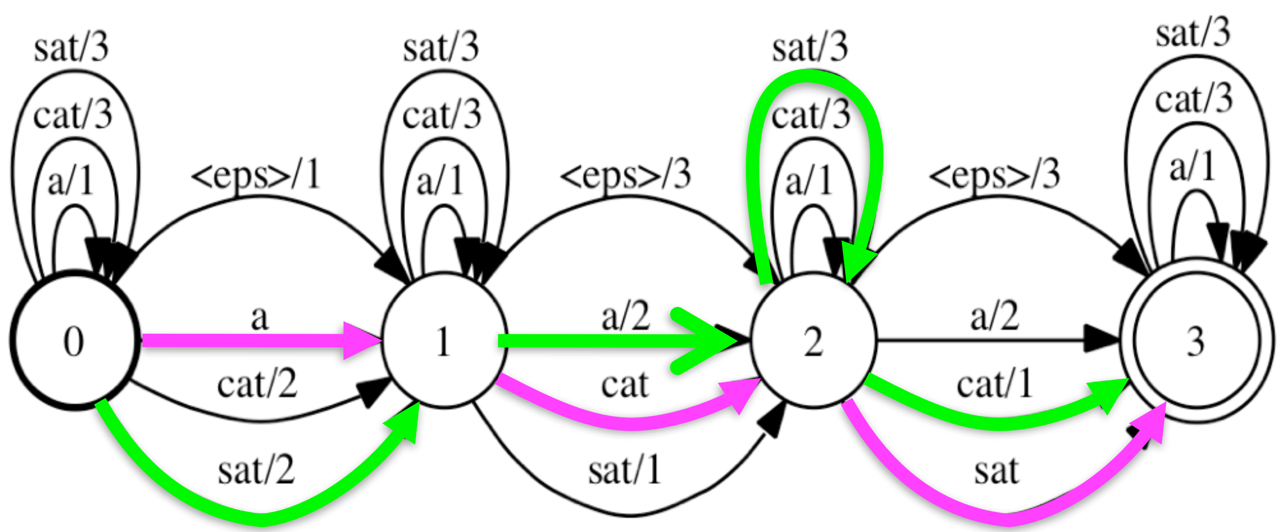
- When we're uncertain about the future, grammar + partial input can affect beliefs about what will happen
- With uncertainty of the past, grammar + future input can affect beliefs about *what has already happened*


grammar
a/2
b/2
c/2
d/2
e/0.415
f/2
e/2
f/0.415
word 1
–b,c} –?}
a/2.58
b/2.58
c/2.58
d/2.58
e/2.58
f/2.58
b/1
c/1
words 1 + 2
–b,c} –f,e}
b/1
c/1
e/2
f/0.415
e/0.415028
f/2.00003
e/2.00003
f/0.415028
b/1
c/1
b/1.41506
c/0.678061
e/1
f/1
e/3.322
f/0.151996

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(**w**,**w***): a WFSA based on Levenshtein distance between words (K_{LD}):



Cost(*a cat sat*)=0

Cost(*sat a sat cat*)=8

Result of K_{LD} applied to **w*** = ***a cat sat***

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

(Anderson, 1990, 1991)

Incremental inference under uncertain input

- Near-neighbors make the "incorrect" analysis "correct":

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

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

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

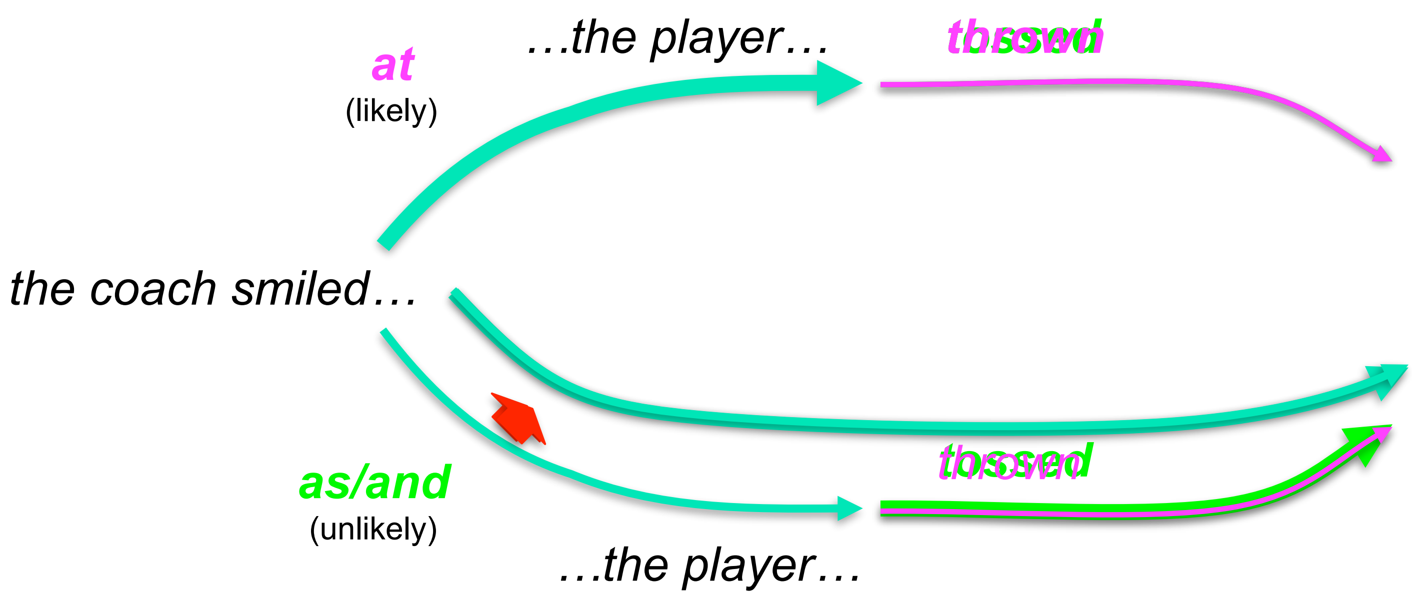
- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively

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

(Anderson, 1990, 1991)

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

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

- Q(**w**,**w***) comes from K_{LD} (with minor changes)
- P_C(**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_i about changes in beliefs 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)

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

- 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

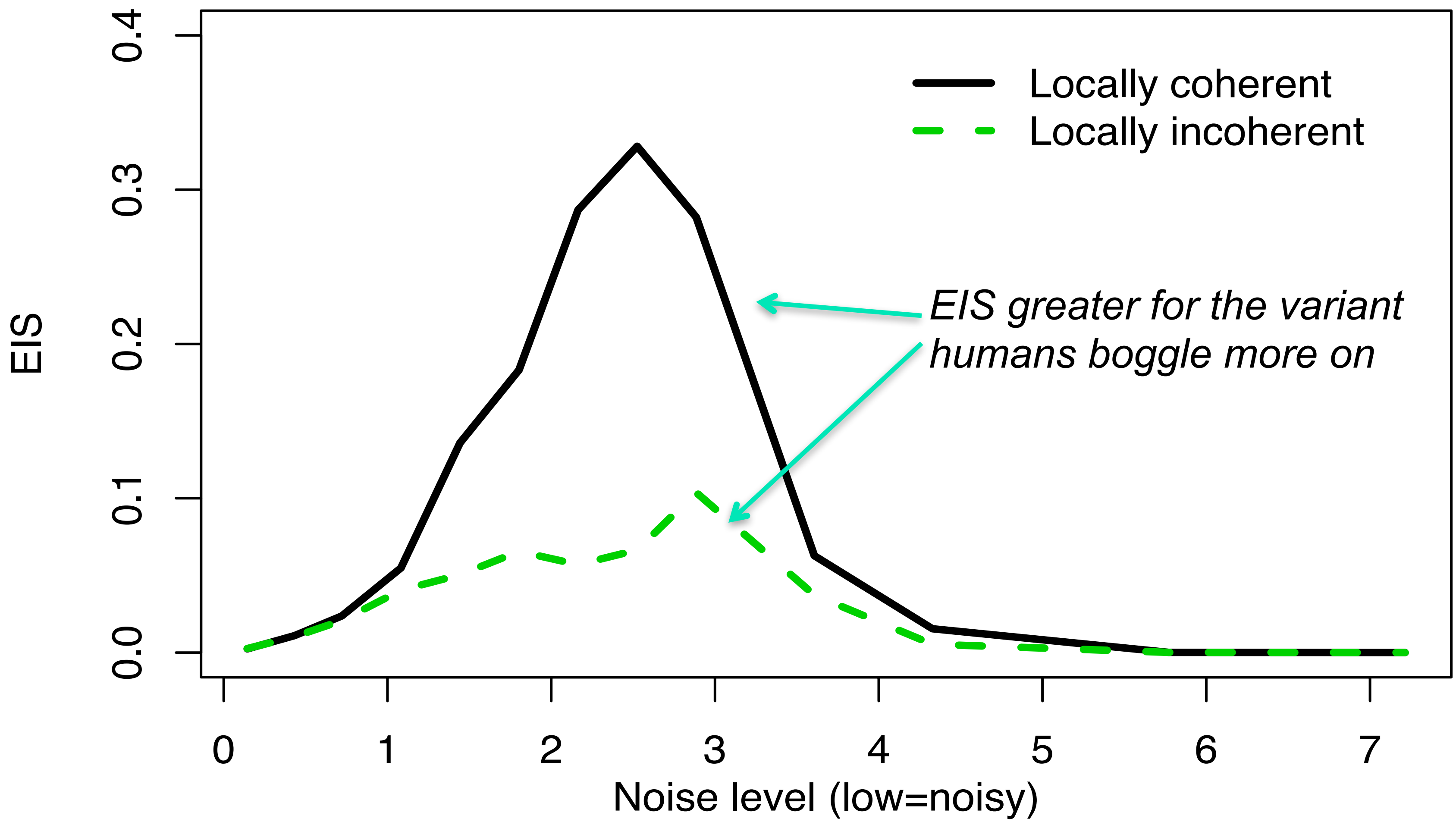
strings up to but excluding word j

- The change induced by w_i is the ***error identification signal*** EIS_i, defined as

$$D\left(\underbrace{P_i\left(w_{[0,i)}\right)}_{\text{new distribution}} \,\middle|\middle|\, \underbrace{P_{i-1}\left(w_{[0,i)}\right)}_{\text{old distribution}}\right)$$

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)

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
-

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

- 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

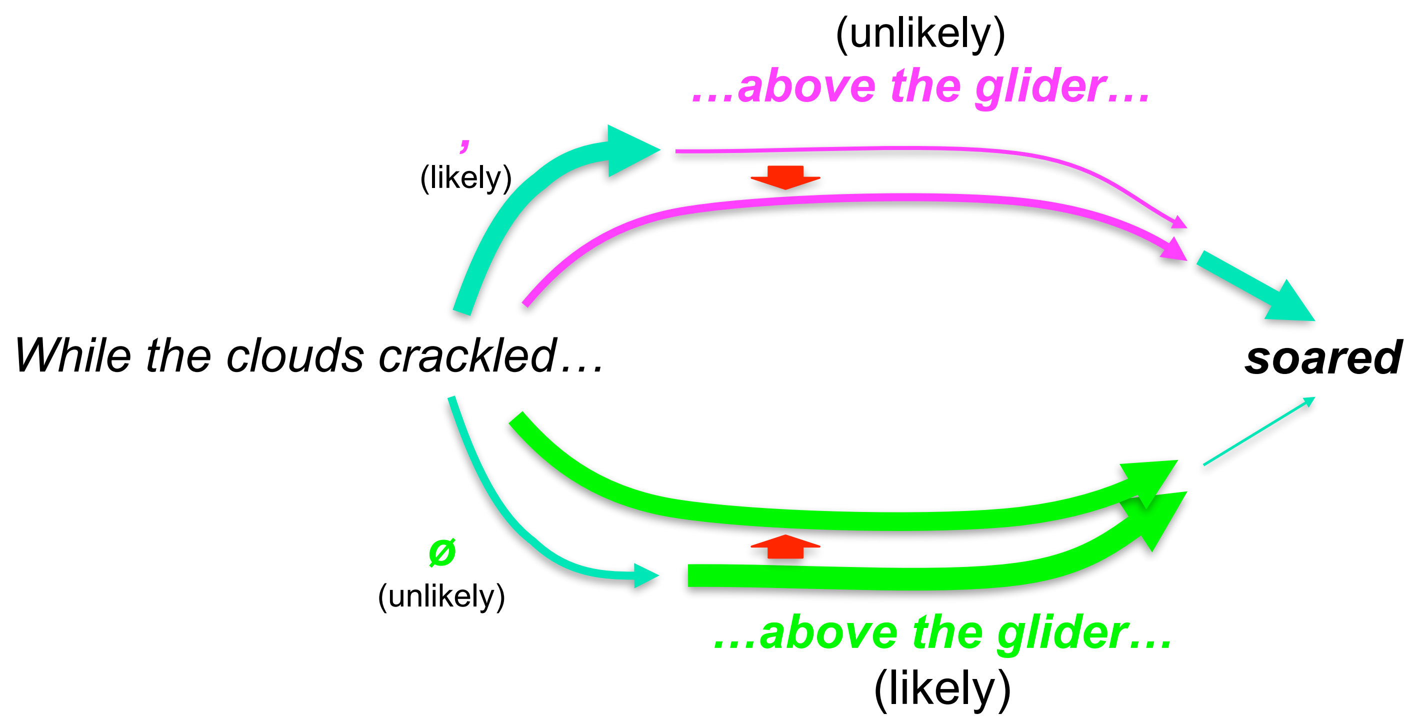
  - "With a comma after *mending* there would be no syntactic garden path left to be studied." (Fodor, 2002)
- We'll see that the story is slightly more complicated.

Prediction 2: hallucinated garden paths

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

- 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



- 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!

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

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

```
While-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
- Simple test of whether beliefs about previous input can be revised

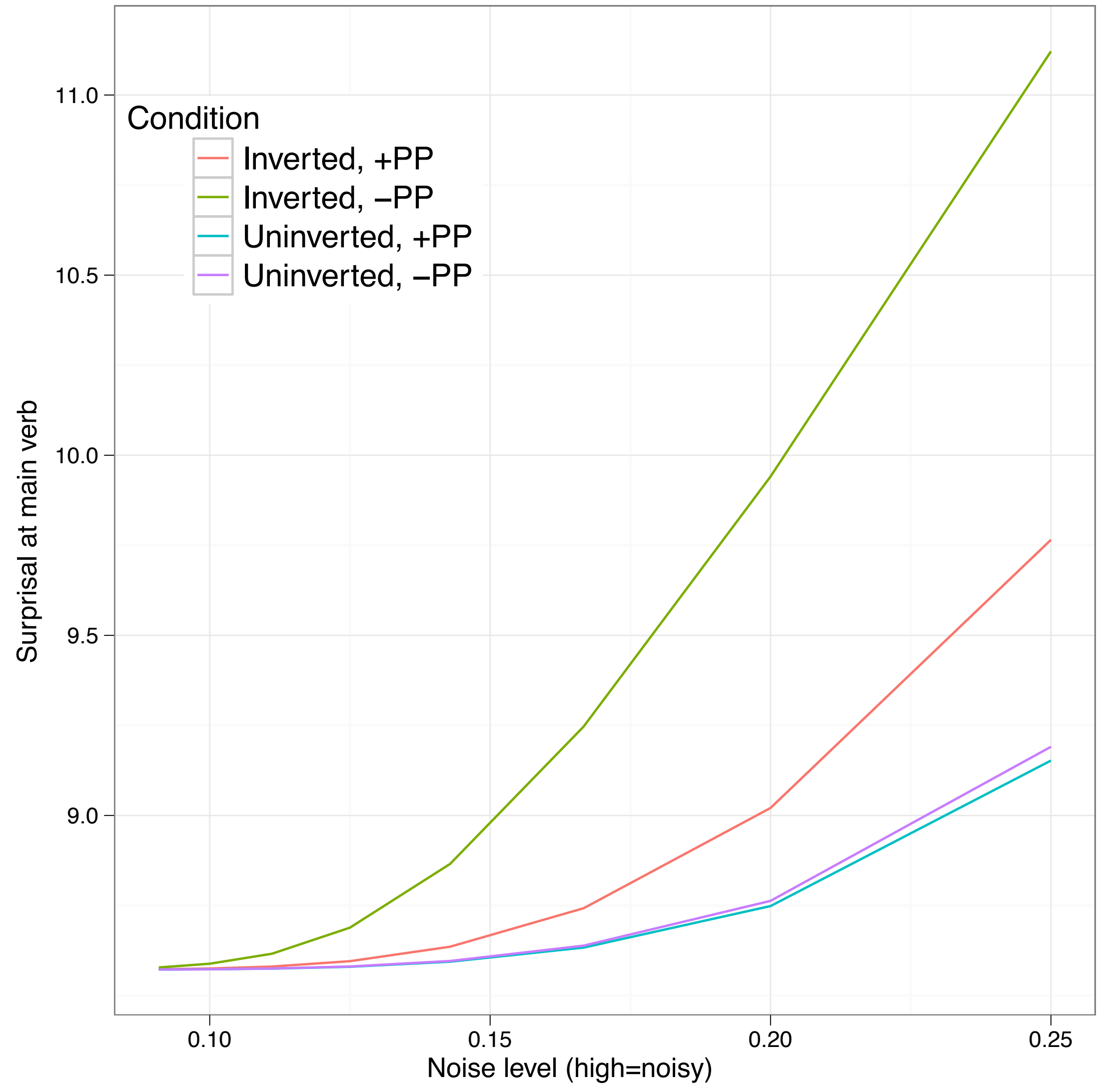
Model predictions

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

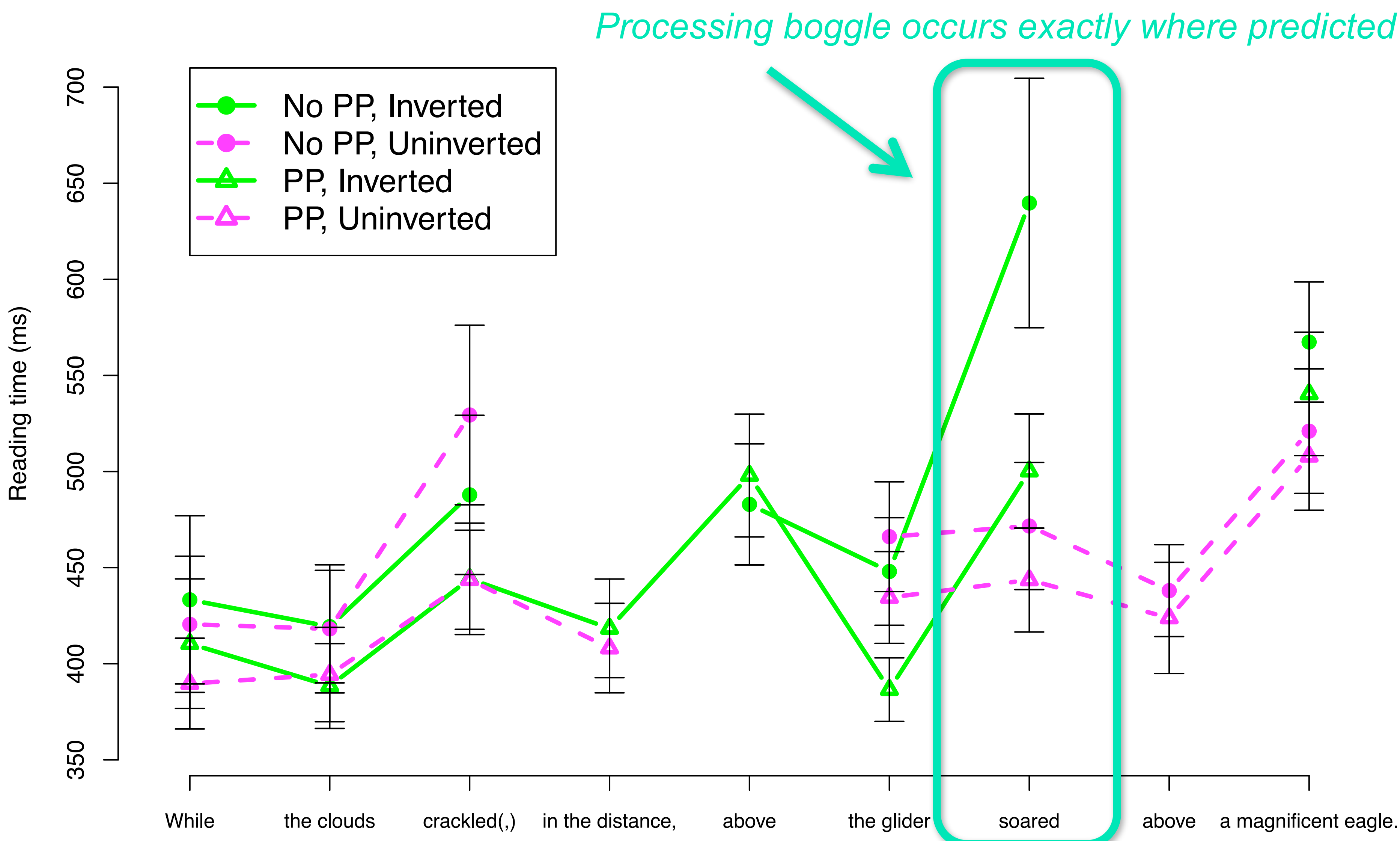
While the clouds crackled in the distance, above the glider soared a magnificent eagle.

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



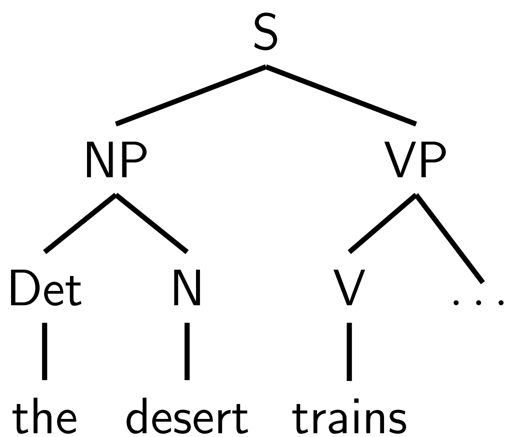
Hallucinated garden-path summary

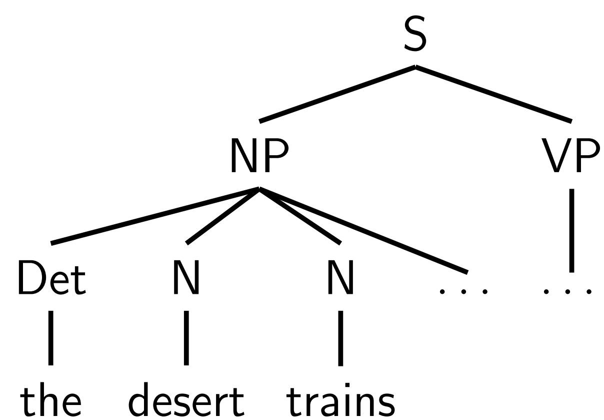
- The *at/toward* study showed that comprehenders *note the possibility of* alternative strings and *act on it*
- This study showed that comprehenders can actually *devote resources to* grammatical analyses inconsistent with the surface string

Hallucinated garden paths cont'd

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

I know that the desert trains could resupply the camp.





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

(Frazier & Rayner, 1987; Macdonald, 1993)

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"

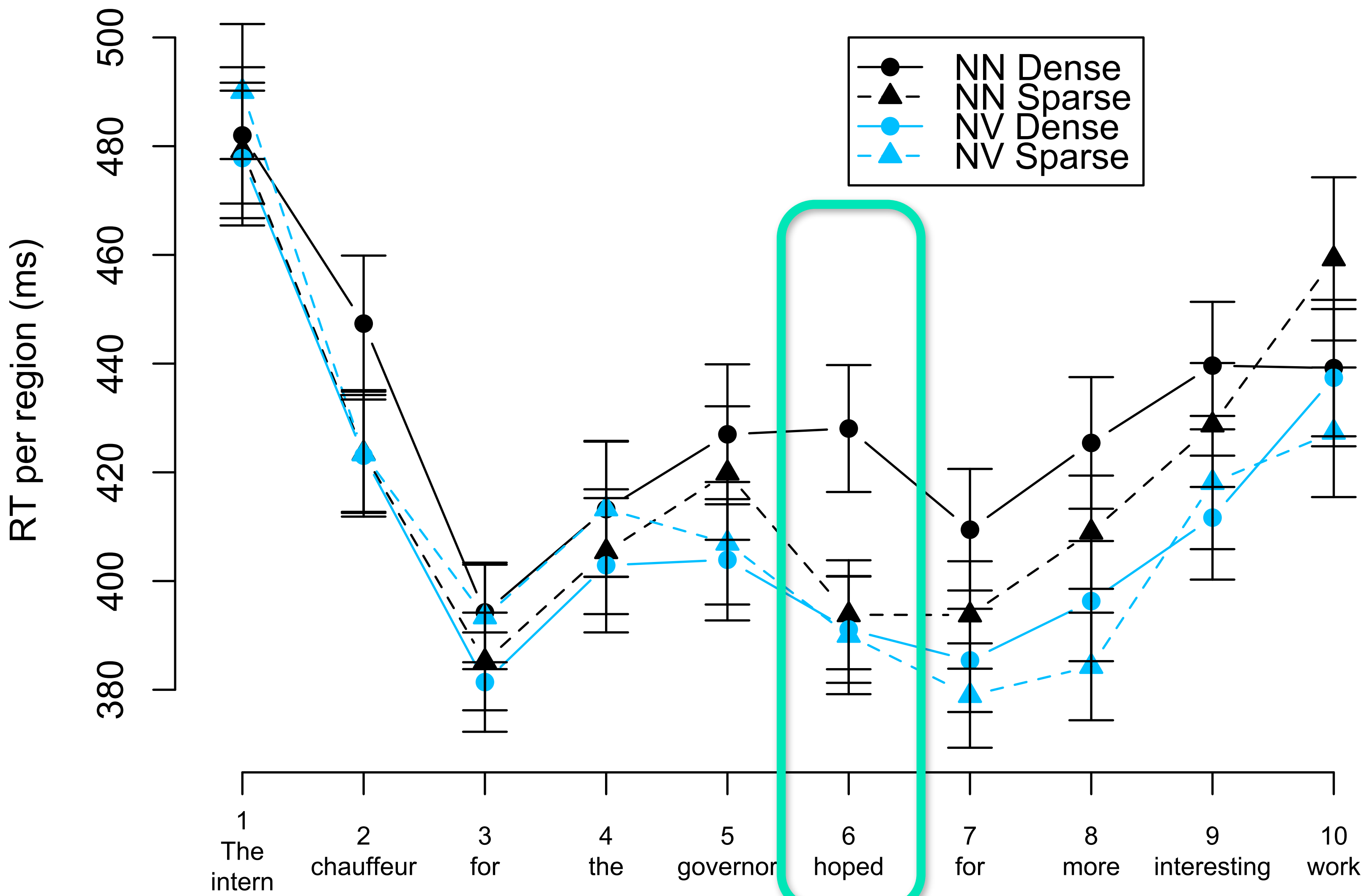
The inexperienced chauffeur for the governor hoped for more interesting work.
[NN, "sparse" neighborhood]

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

(Bergen, Levy, & Gibson, 2012)

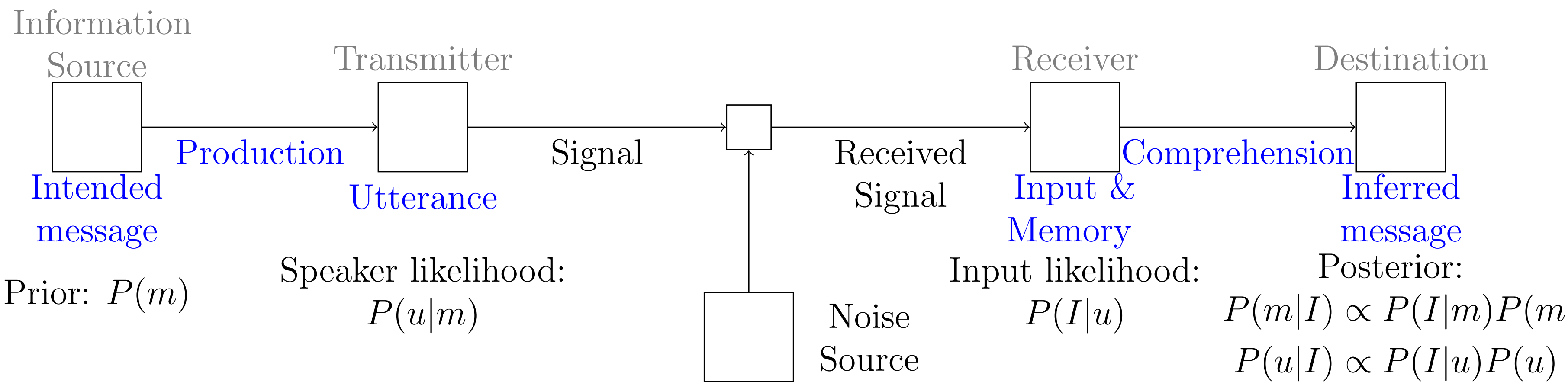
Results

- RT spike at disambiguating region for NN Dense



(Bergen, Levy, & Gibson, 2012)

Noisy-channel theory of language processing



(Shannon, 1948; Levy, 2008; Gibson et al., 2013)

Simple question-answering

`The woman lost the diamond.`

Did the woman lose something? **Yes**

`The ball kicked the girl.`

Did the girl kick something? **No**

`The businessman benefited from the tax law.`

Did the tax law benefit from anything? **No**

`The cook baked a cake Lucy.`

Was something baked for Lucy? **No** *(Yes?)*

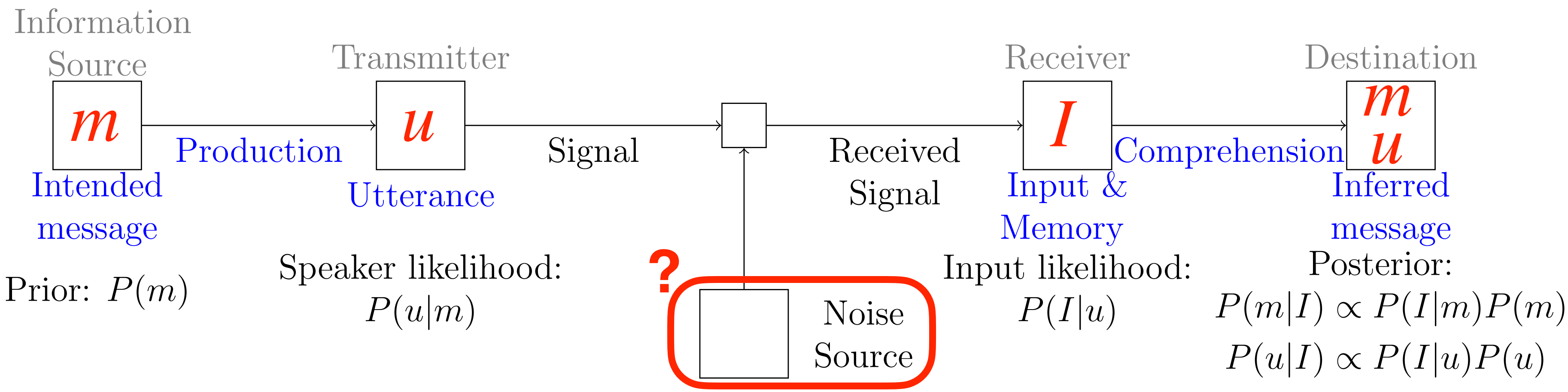
Over 2/3 of answers!

(Ferreira, 2003; Gibson et al., 2013)

Noisy-channel semantic interpretation?

$I \leftarrow$ `The cook baked a cake Lucy.`

m? *Was something baked for Lucy?*



In two semantically plausible "neighbor" sentences, the answer is "yes":

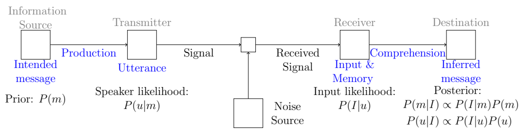
`The cook baked a cake Lucy.`

Hypothesized noise operation: **deletion** `for`

`The cook baked a cake Lucy.`

Hypothesized noise operation: **exchange** `Lucy a cake`

Predictions for implausible sentences



$$P(m\,|\,I) \propto \underbrace{P(I\,|\,m)}_{\text{Noise operation}}\underbrace{P(m)}_{\text{Plausibility}}$$

Non-literal interpretation?

Double Object/Benefactive-*for* alternation

		Deletion/ insertion	***Exchange***
Implausible	**`The cook baked a cake Lucy.`**	Yes	Yes
	`The cook baked Lucy for a cake.`	Yes	Yes
Plausible	**`The cook baked Lucy a cake.`**	No	No
	`The cook baked a cake for Lucy.`	No	No

Active/Passive alternation

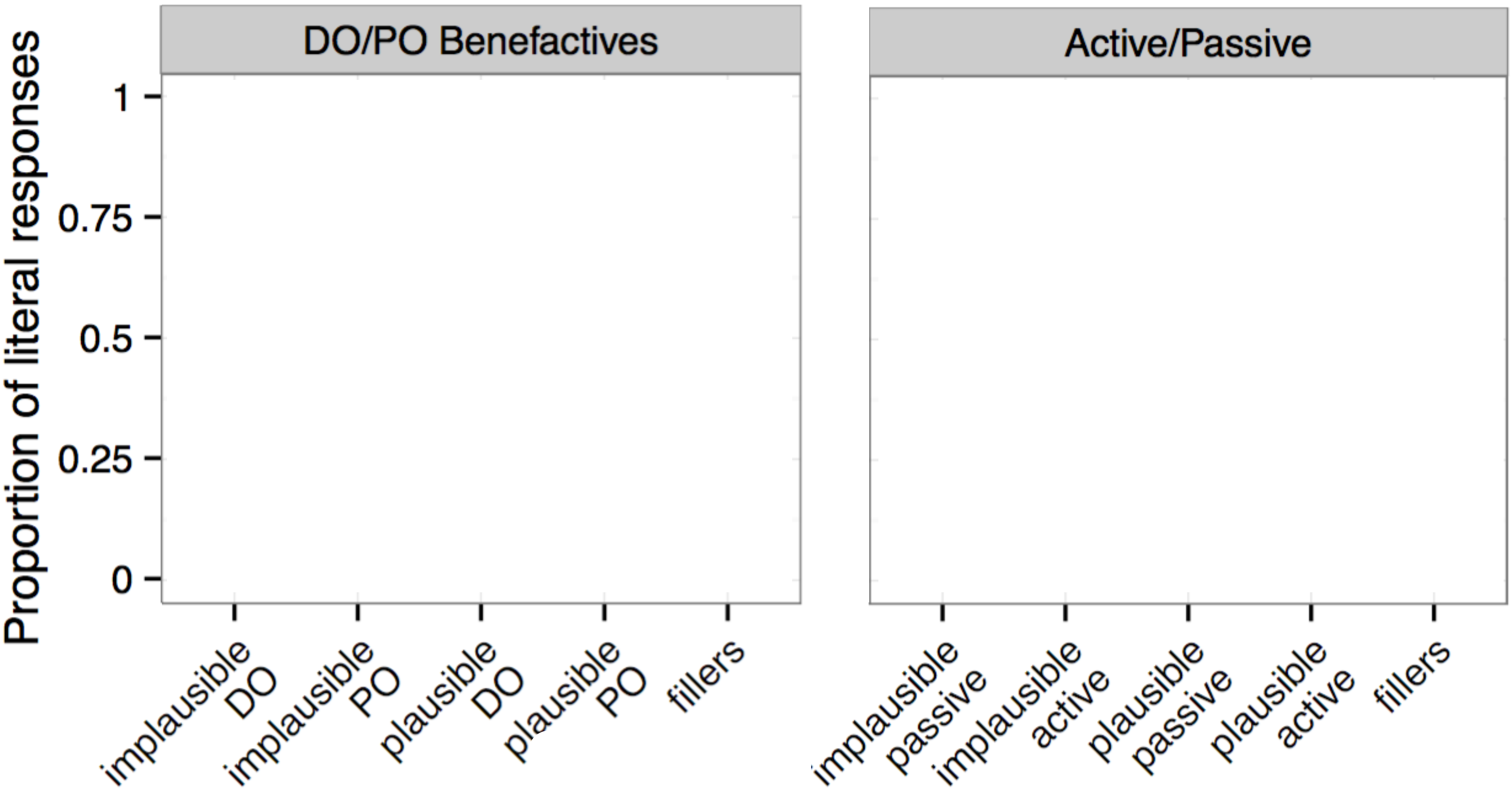
		Deletion/ insertion	***Exchange***
Implausible	**`The ball kicked the girl.`**	No	Yes
	`The girl was kicked by the ball.`	No	Yes
Plausible	**`The girl kicked the ball.`**	No	Yes
	`The ball was kicked by the girl.`	No	Yes

(Gibson et al., 2013)

Literal vs. non-literal interpretation rates

Non-literal interpretations for implausible sentences?

Insertion/Deletion	***Exchange***	***Insertion/Deletion***	***Exchange***
Yes	**Yes**	**No**	**Yes**



(Gibson et al., 2013; data from replication by Poppels & Levy, 2016)

Five alternations in an insertion/deletion model

English constructions	Change	Implausible version
1. Active/passive	Two insertions	c. The girl was kicked by the ball. (passive)
	Two deletions	d. The ball kicked the girl. (active)
2. Subject-locative/ object-locative	One deletion, one insertion	c. The table jumped onto a cat. (object-locative)
	One insertion, one deletion	d. Onto the cat jumped a table. (subject-locative)
3. Transitive/intransitive	One insertion	c. The tax law benefited from the businessman. (intransitive)
	One deletion	d. The businessman benefited the tax law. (transitive)
4. DO/PO goal	One insertion	c. The mother gave the daughter to the candle. (PO-goal)
	One deletion	d. The mother gave the candle the daughter. (DO-goal)
5. DO/PO benefactive	One insertion	c. The cook baked Lucy for a cake. (PO-benef)
	One deletion	d. The cook baked a cake Lucy. (DO-benef)

c=inferred insertion   d=inferred deletion

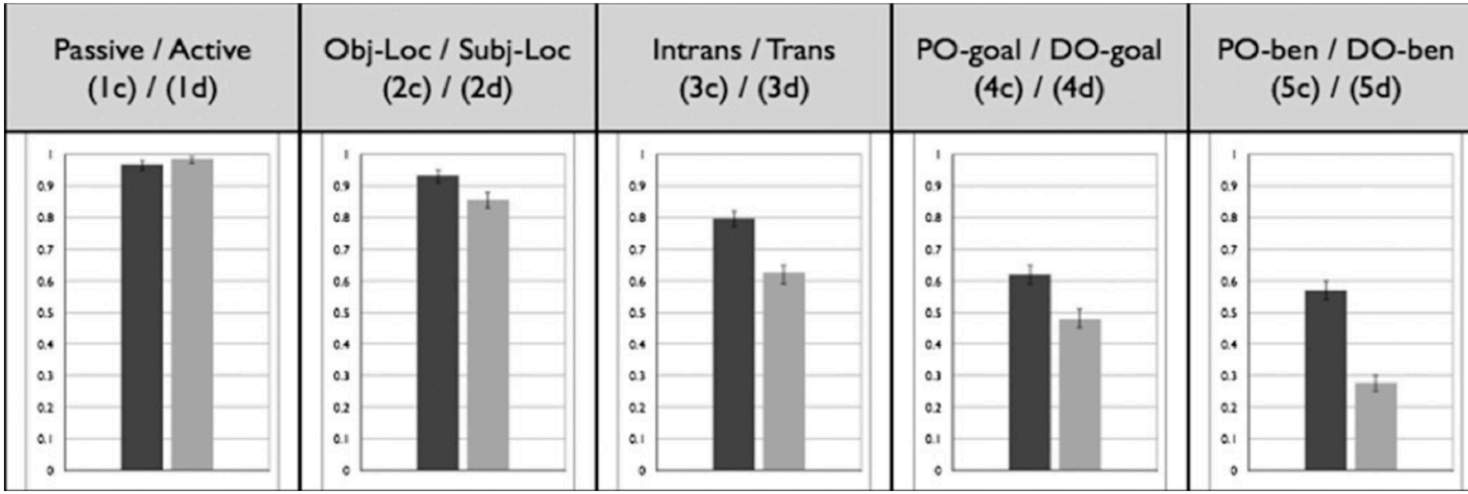
(Gibson et al., 2013; plausible versions not shown here)

Five alternations in an insertion/deletion model

$$P(m \mid I) \propto \underline{P(I \mid m)}\,\underline{P(m)}$$

Noise operation

Plausibility



Base experiment

20 experimental items, 60 plausible & grammatically normal fillers → 10/80 implausible trials

Fillers with syntactic errors

"A legislator lied to the consultant a new bill"

"A bystander was the fireman by rescued in the nick of time"

Many implausible trials

100 experimental items, 60 plausible & grammatically normal fillers → 50/160 implausible trials

(Gibson et al., 2013)

Inferring deletions versus insertions

$$P(m \mid I) \propto \underbrace{P(I \mid m)}_{\text{Noise operation}} \underbrace{P(m)}_{\text{Plausibility}}$$

1 Delete

2 Choose deletion location

`The cook baked a cake ~~for~~ Lucy.`

↓

`The cook baked a cake Lucy.`

1 Insert

2 Choose insertion location

`The cook baked Lucy a cake.`

for

3 Choose what to insert

↓

`The cook baked Lucy for a cake.`

Noisy-channel prediction: inferring deletions should be intrinsically easier than inferring insertions!

Five alternations in an insertion/deletion model

$$P(m|I) \propto \underbrace{P(I|m)}_{\text{Noise operation}}\underbrace{P(m)}_{\text{Plausibility}}$$

Base experiment

20 experimental items, 60 plausible & grammatically normal fillers → 10/80 implausible trials

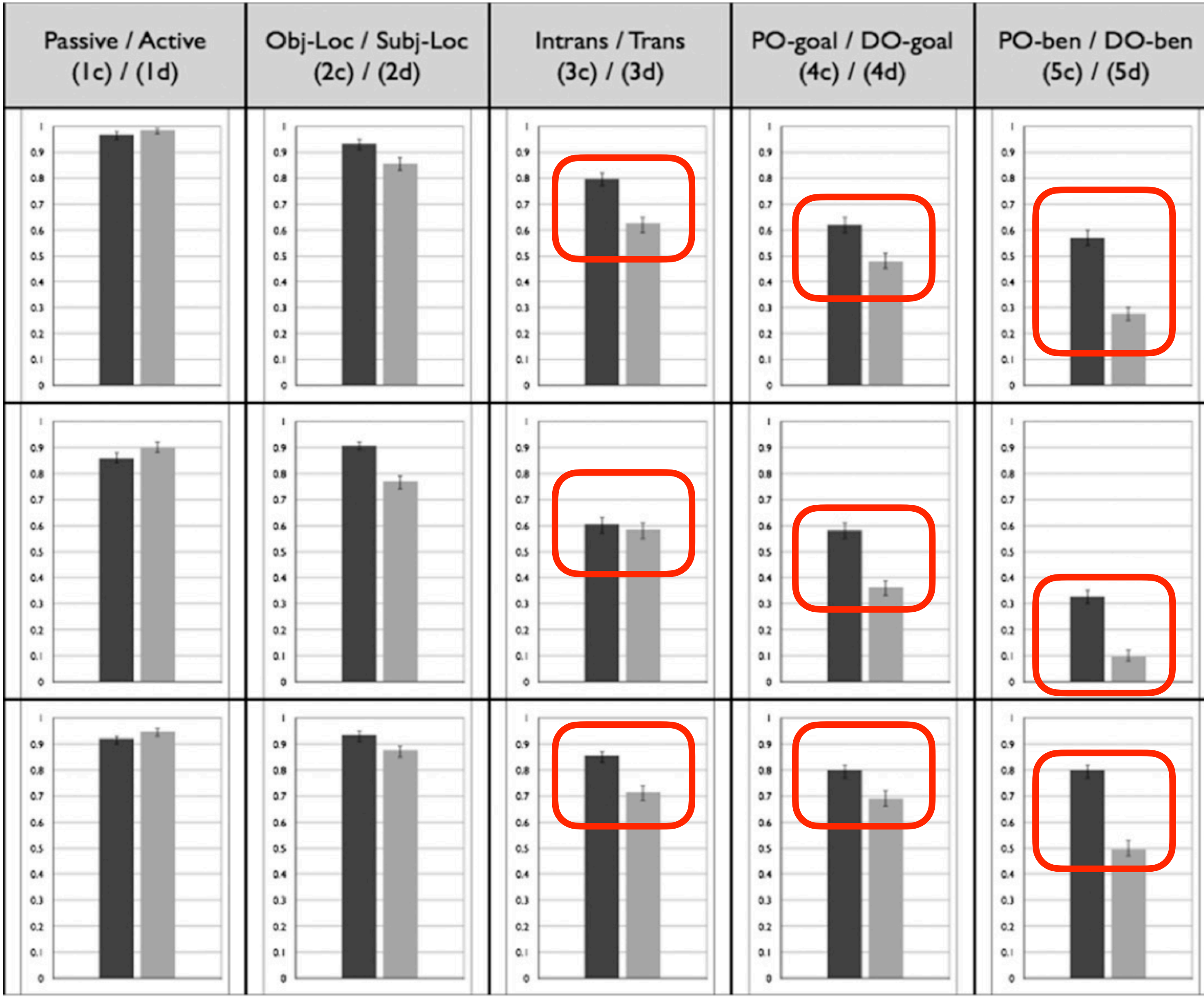
Fillers with syntactic errors

"A legislator lied to the consultant a new bill"

"A bystander was the fireman by rescued in the nick of time"

Many implausible trials

100 experimental items, 60 plausible & grammatically normal fillers → 50/160 implausible trials



(Gibson et al., 2013)

In the real world (2008)

I'm not going to solely blame all of man's activities on changes in climate.

Sarah Palin (images credit Gage Skidmore)

I'm not one to attribute every activity of man to climate change.

(Credit to Colin Phillips for bringing these examples to light)

Corpora of speech errors

Anticipations	John dropped his cuff of coffee
	reek long race
Perseverations	John gave the goy (=gave the boy)
	Spanish speaping people
Exchanges	teep a cape (=keep a tape)
	the nipper is zarrow
	Fancy getting your model renosed (=nose remodeled)

(Fromkin, 1971; Garrett, 1975, inter alia)

Revisiting the possibility of exchanges

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

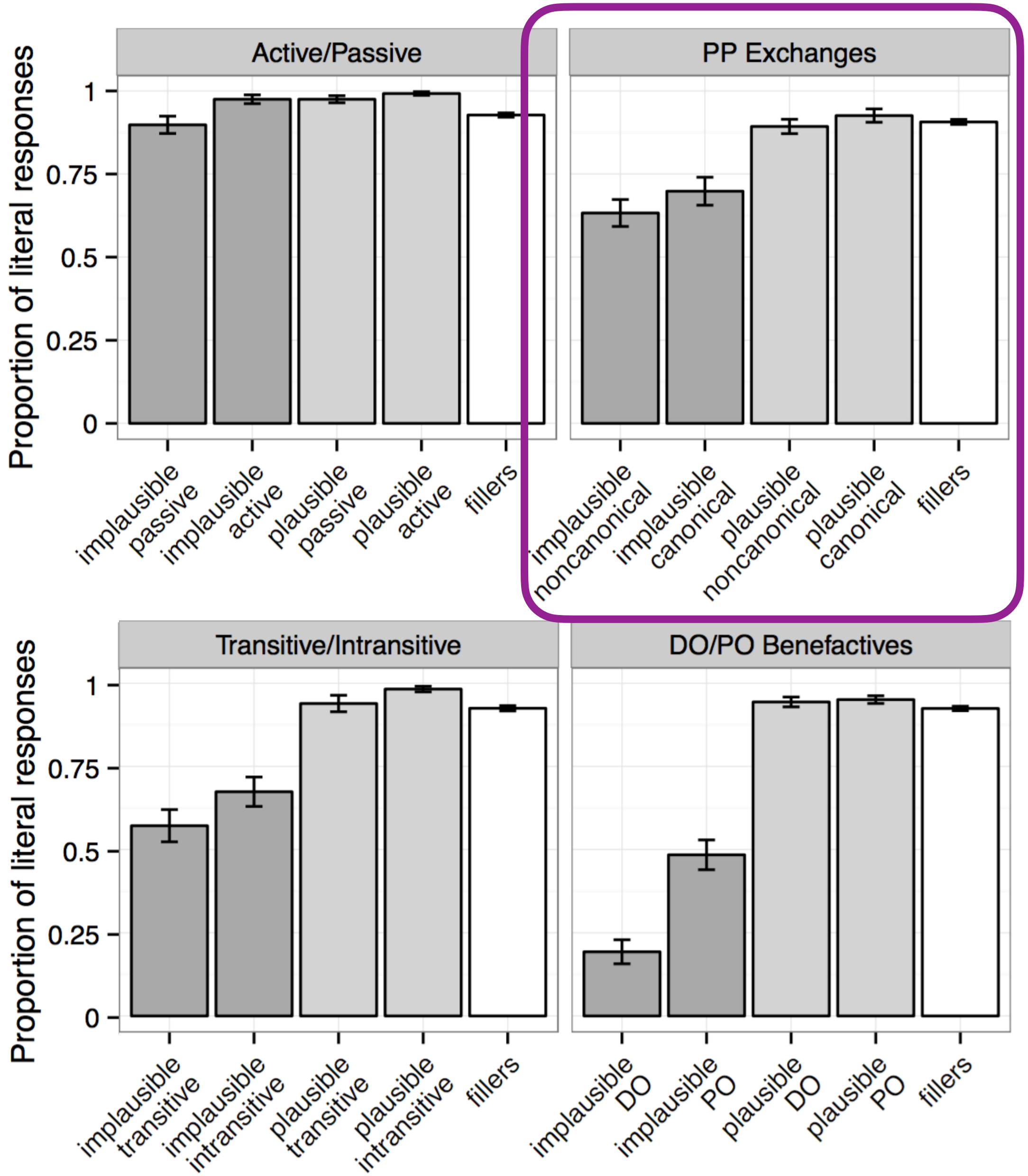
- An occasional speech error of mine that I've noticed for years, but that no one ever notices me make
- Extraordinarily unlikely under an insertions/deletions noise model
- But reasonably likely if word ***exchanges*** are admitted

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]

Did something fall to the floor?

(Poppels & Levy 2016)

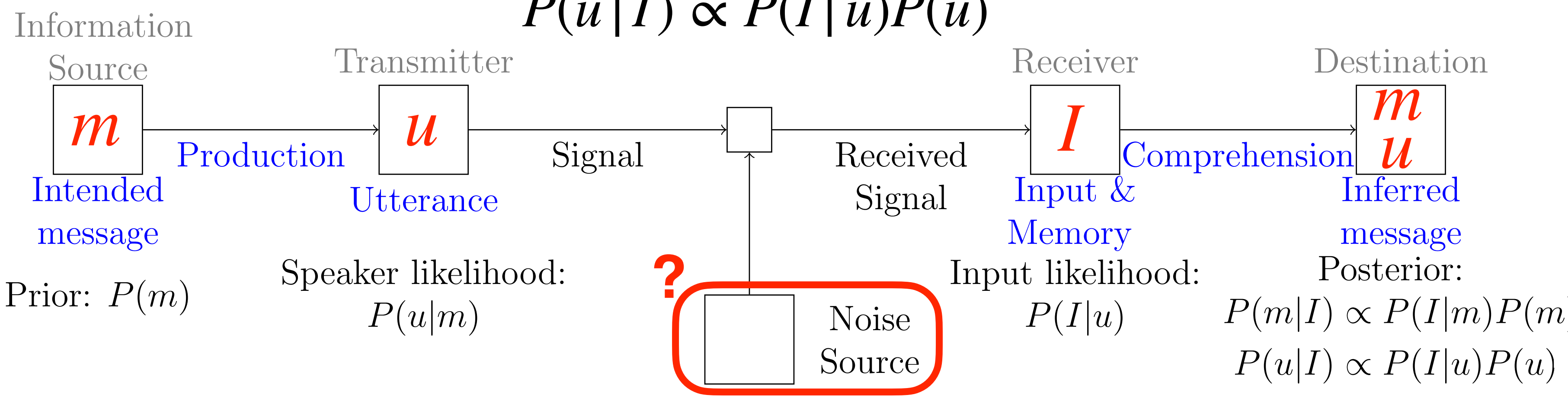
Exchanges in the noise model



(Poppels & Levy 2016)

Probing inferred intended utterances

$$P(u \mid I) \propto P(I \mid u)P(u)$$



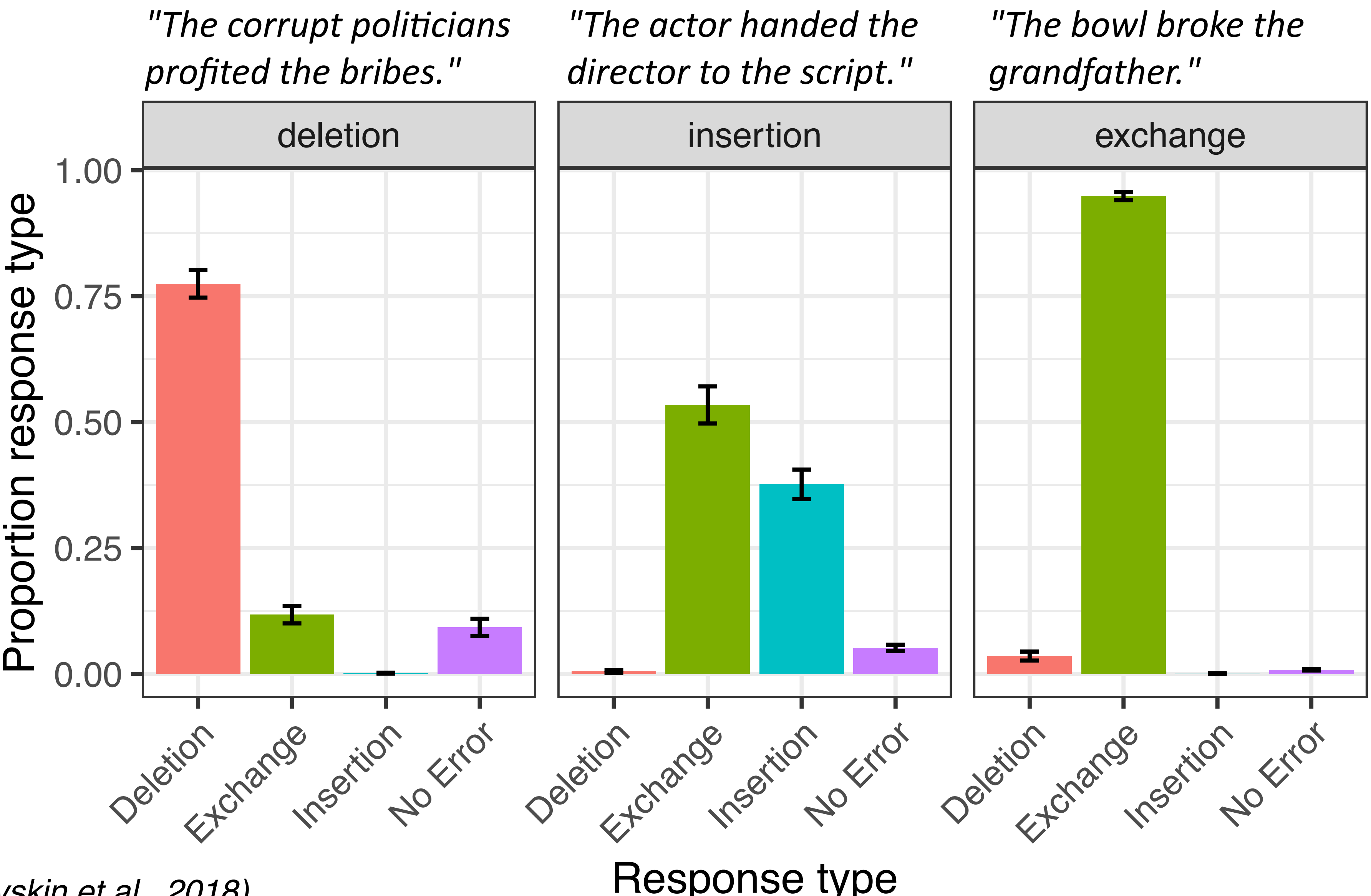
Experiment "cover story": read transcriptions of speech that might have errors, retype with edits if they think the speaker might have meant something else

The ball kicked the girl. **The judge gave the athlete to the prize.**

The ball kicked the girl. No error	The judge gave the athlete the prize. Insertion
The girl kicked the ball. Exchange	The judge gave the athlete a prize. Insertion
The ball was kicked by the girl. Deletion	The judge gave the prize to the athlete. Exchange

(Ryskin et al., 2018)

Probing inferred intended utterances



(Ryskin et al., 2018)

Noisy-channel interpretation summary

- The noisy-channel framework suggests investigating global interpretations as well as incremental processing
- "Non-literal" interpretations can be very frequent for the right stimuli
- Interpretations broadly follow Bayesian principle of trade-off between prior and likelihood
  - Deletions easier to infer than insertions
  - Higher grammatical error rate in environment→more non-literal inference
  - More implausible sentences in environment→less non-literal inference
- *However*, status of exchange errors in the noise model remains a mystery

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.

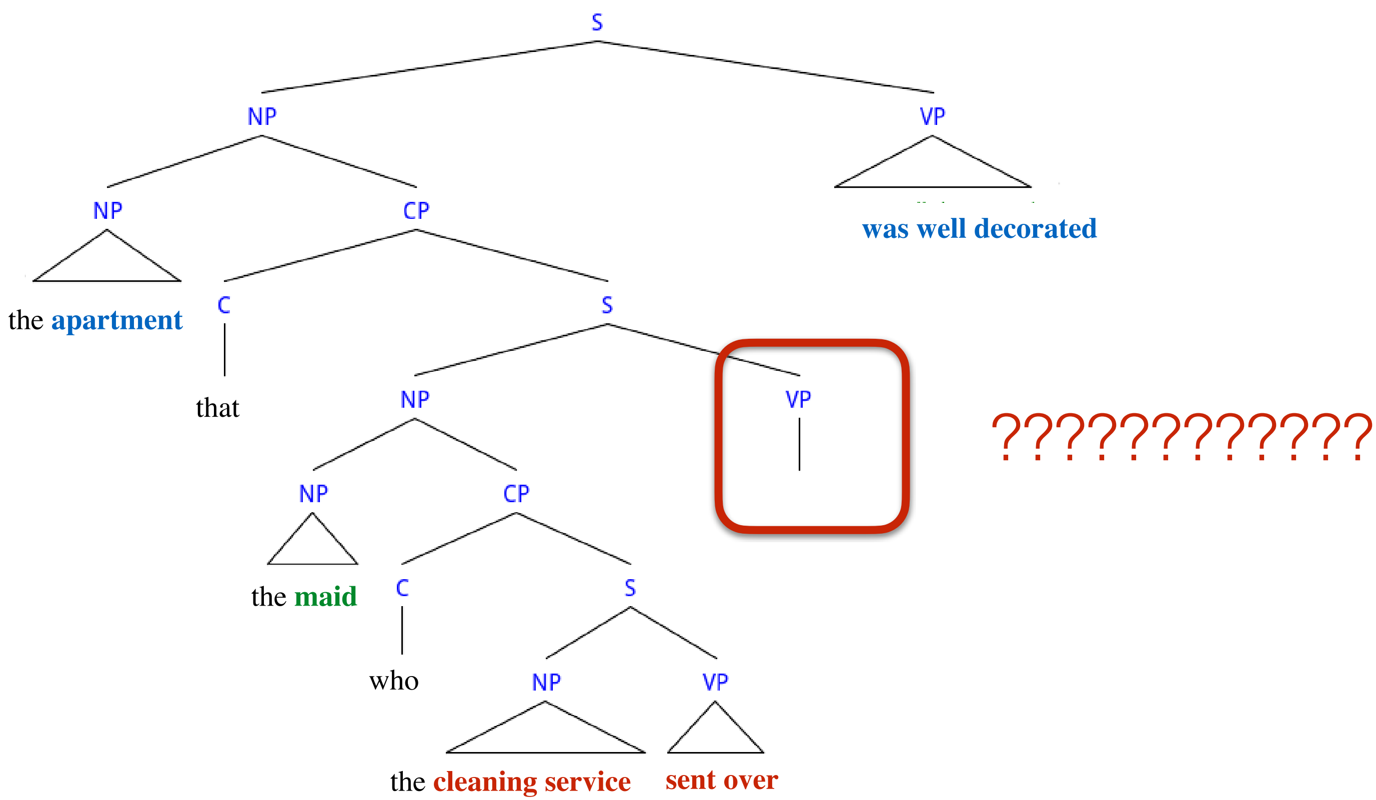
(Futrell & Levy, 2017; Futrell et al., 2020)

(Slide courtesy Richard Futrell)

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**. 👎



(Slide courtesy Richard Futrell)

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?

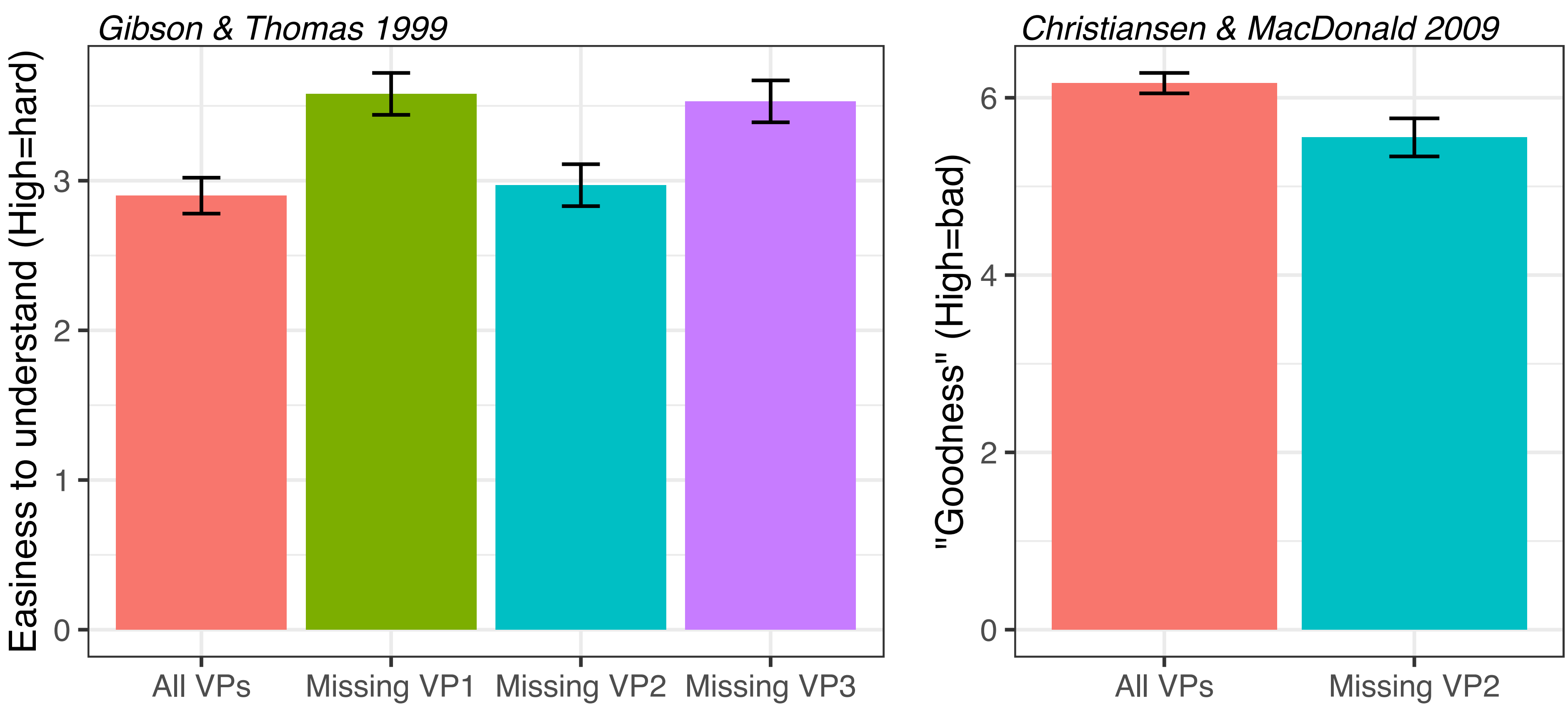
(Slide courtesy Richard Futrell)

Gibson & Thomas 1999: whole-sentence reading

The ancient manuscript that the graduate student who the new card catalog had confused a great deal was studying in the library was missing a page.

Christiansen & MacDonald 2009: word-by-word self-paced reading, follows by rating

The chef who the waiter who the busboy offended appreciated admired the musicians.



Structural Forgetting

1. *The apartment that the maid who the cleaning service sent was well-decorated. 👍
   Die Wohnung, die das Zimmermädchen, das der Reinigungsdienst übersandte, war gut eingerichtet. 👎

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

(Slide courtesy Richard Futrell)

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]

(Slide courtesy Richard Futrell)

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**.)

C(NOUN THAT NOUN THAT NOUN VERB VERB) < *C*(NOUN THAT NOUN THAT NOUN VERB VERB VERB)

C(2 VERBS) < *C*(3 VERBS)

(Slide courtesy Richard Futrell)

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.

(*Slide courtesy Richard Futrell*)

Noisy-Context Surprisal Account of Structural Forgetting

- Futrell & Levy (2017) demonstrate that this works for toy grammars of English and German.

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

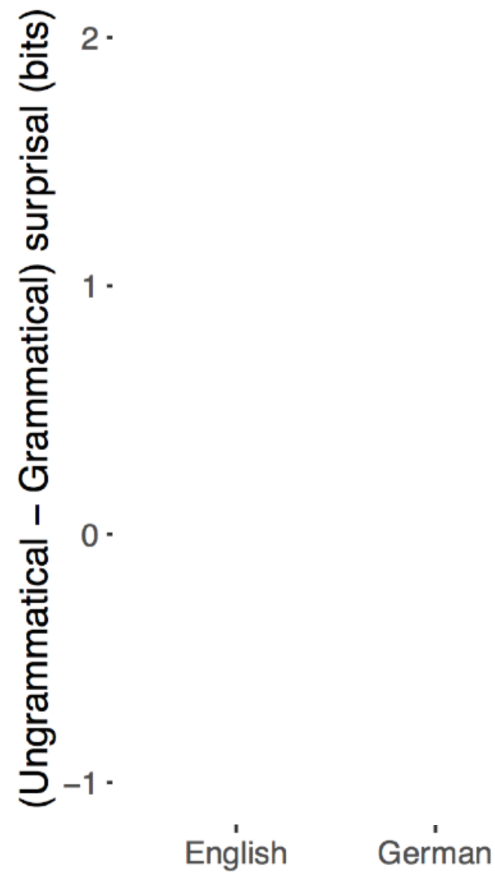
```
NOUN VERB
NOUN PREP NOUN VERB
NOUN THAT VERB NOUN VERB
NOUN THAT NOUN VERB VERB
NOUN THAT NOUN THAT NOUN...
```

Plus **deletion noise**: every token in the context is forgotten (deleted) with probability d

(Slide courtesy Richard Futrell)

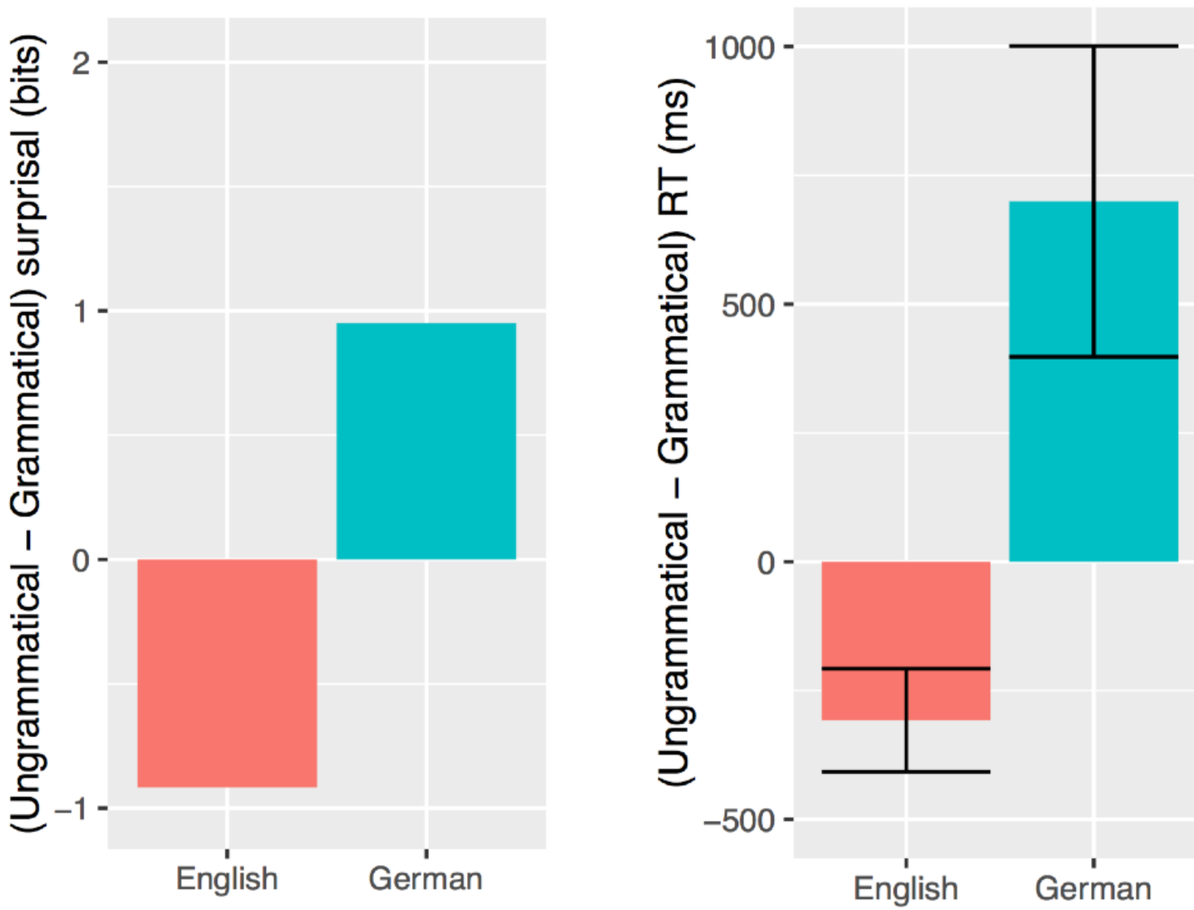
Noisy-Context Surprisal Account of Structural Forgetting

- 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:



(Slide courtesy Richard Futrell)

Noisy-Context Surprisal Account of Structural Forgetting



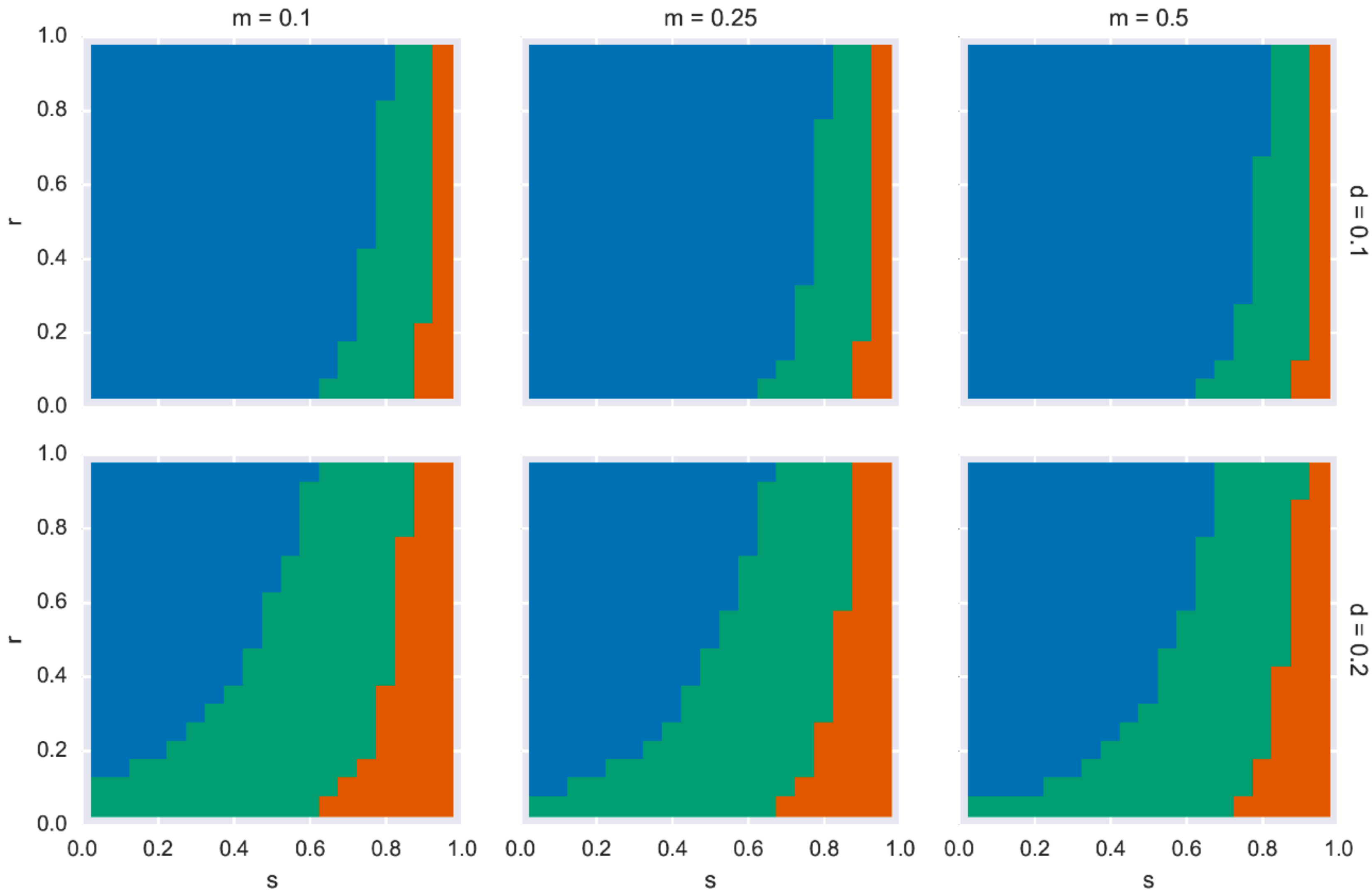
Vasishth et al. (2010)

(Slide courtesy Richard Futrell)

Robustness to choice of model parameters

m Modifier probability
s Probability of English RC being verb-final
d Probability of context token deletion

= English+German-like pattern



(Futrell & Levy, 2017)

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).

(Slide courtesy Richard Futrell)

Structural Forgetting and the Noisy Channel

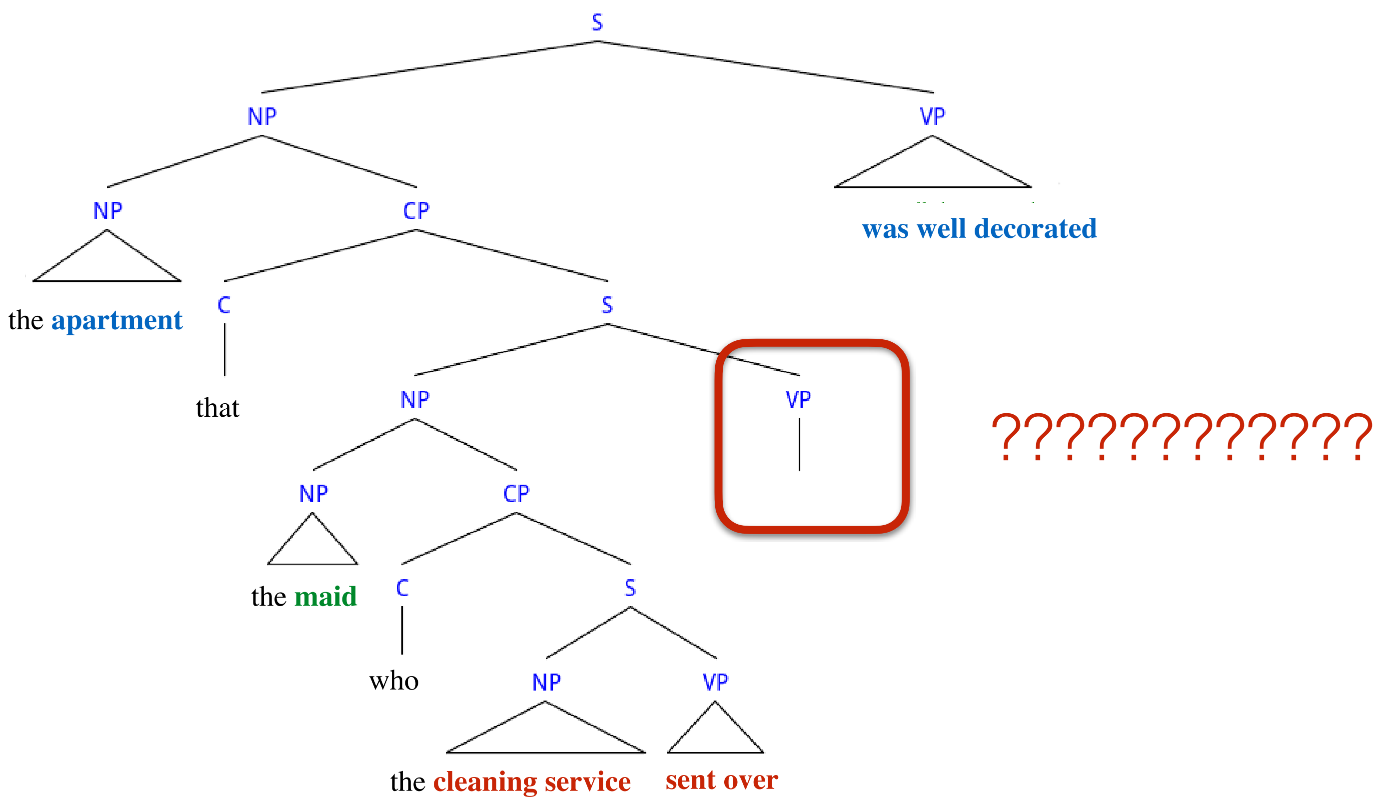
1. The apartment that the maid who the cleaning service sent over was well-decorated.
2. The apartment that the maid who the cleaning service sent over cleaned was well-decorated.

(Futrell & Levy, 2017)

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.)
- 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 was well-decorated. 👍

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

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

   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**. 👎

- These contexts are more common in German than English (Roland et al., 2007).
  - English: the maid [that <u>cleaned</u> the apartment] **80%**
    the apartment [that the maid <u>cleaned</u>] **20%**
  - German: das Dienstmädchen, [das die Wohnung <u>reinigte</u>]
    die Wohnung, [die das Dienstmädchen <u>reinigte</u>]

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**.)

C(NOUN THAT NOUN THAT NOUN VERB VERB) < *C*(NOUN THAT NOUN THAT NOUN VERB VERB VERB)

C(2 VERBS) < *C*(3 VERBS)

Noisy-Context Surprisal Account of Structural Forgetting

$$C(2\ \textsc{verbs}) \ < \ C(3\ \textsc{verbs})$$

- 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

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

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

NOUN VERB

NOUN PREP NOUN VERB

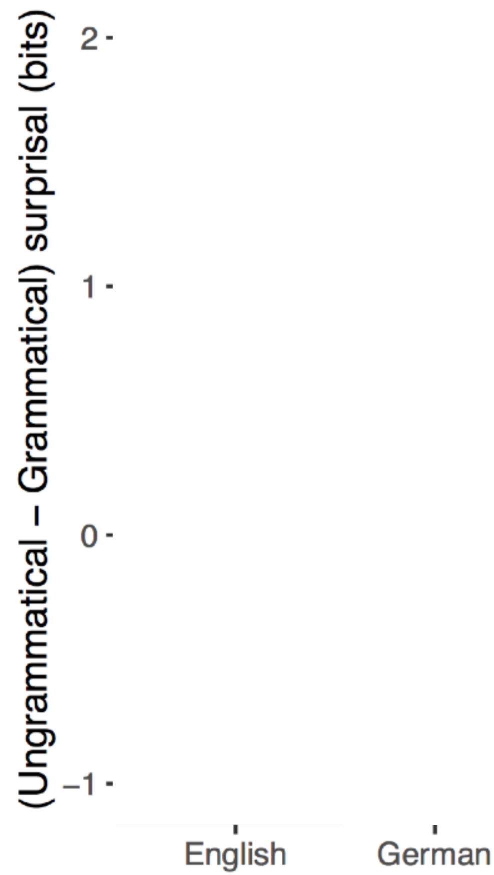
NOUN THAT VERB NOUN VERB

NOUN THAT NOUN VERB VERB

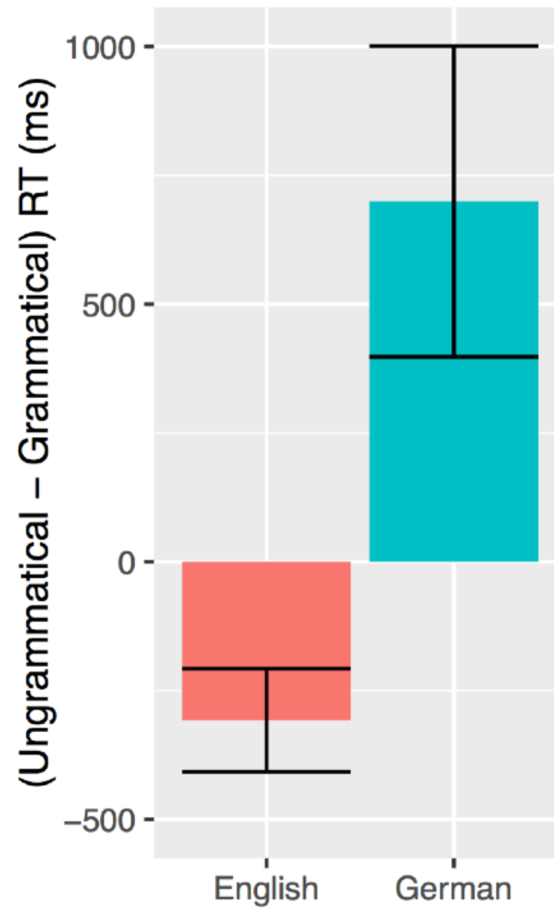
NOUN THAT NOUN THAT NOUN...

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



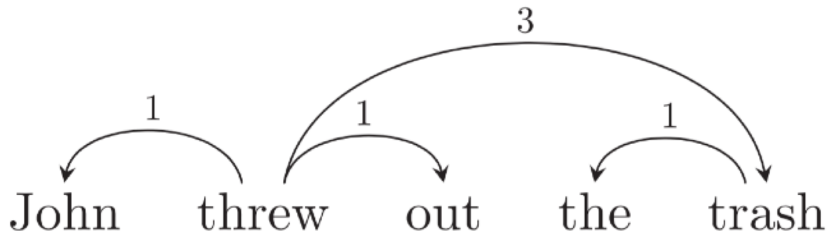
Vasishth et al. (2010)

Noisy-Context Surprisal Account of Structural Forgetting

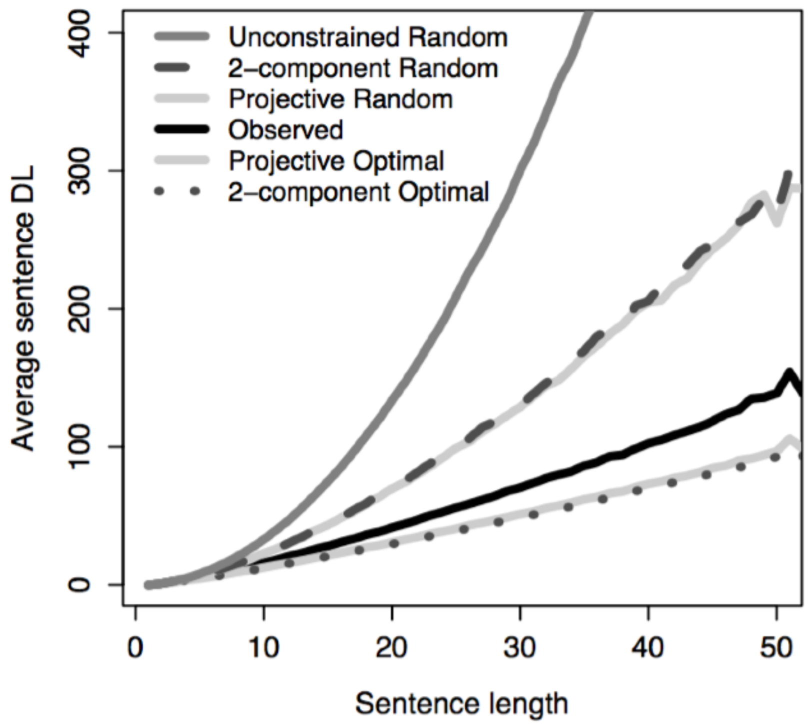
- Probability that a context is remembered depends on its prior probability.
  - Noisy-context surprisal *explains* the behavior of the RNN in Frank et al. (2016): the RNN is using a lossily compressed / noisy representation of context.
- The model has an explicit grammar (competence), but cannot apply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

Dependency length and noisy-channel surprisal

- Syntactic dependencies vary in linear distance



- Idea with long history: short dependencies preferred



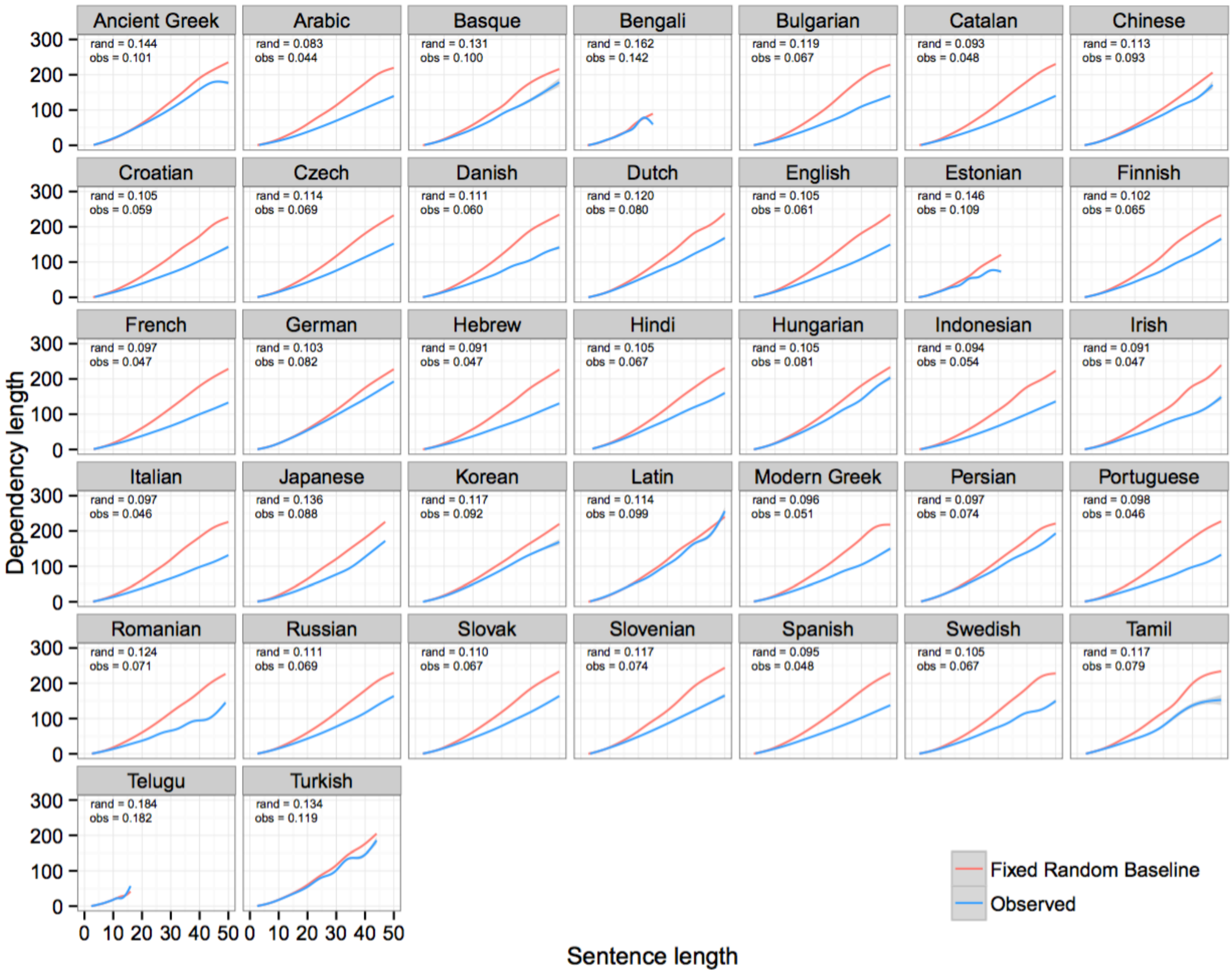
Random-linearization dependency lengths

Observed dependency lengths

(Hawkins, 1994; Gibson, 1998, 2000; Gildea & Temperley, 2007, 2009; Park & Levy, 2009; Futrell et al., 2015)

(Futrell et al., 2015)

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

(Futrell & Levy, 2017)

From noisy-channel & surprisal to dependency length minimization

noisy context

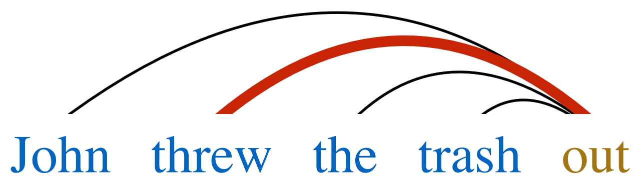
- Suppose we have an **increasing noise rate** the longer a word has been in memory.
- When "threw" is far from "out", then it is less likely to reduce the surprisal of "out": more likely to be affected by noise.
- Noisy-context surprisal increases when **words that predict each other are far apart**.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

(Futrell & Levy, 2017)

Derivation of Information Locality

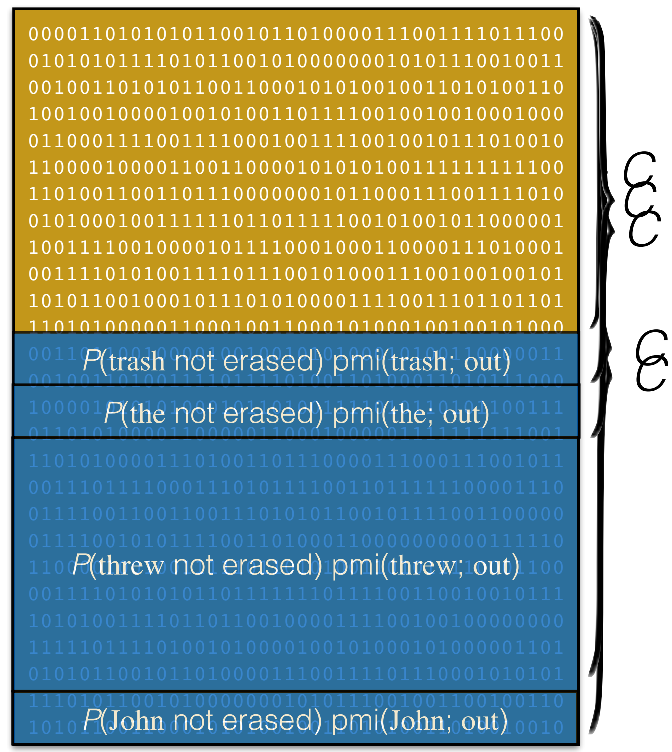
- Erasure noise decreases the influence of context:

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



h(out) - *P*(John not erased) pmi(John; out)

- *P*(threw not erased) pmi(threw; out)
- *P*(the not erased) pmi(the; out)
- *P*(trash not erased) pmi(trash; out)

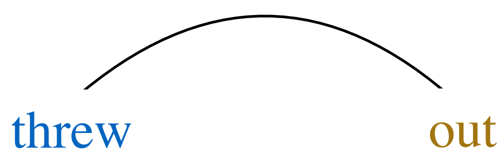


(Futrell & Levy, 2017)

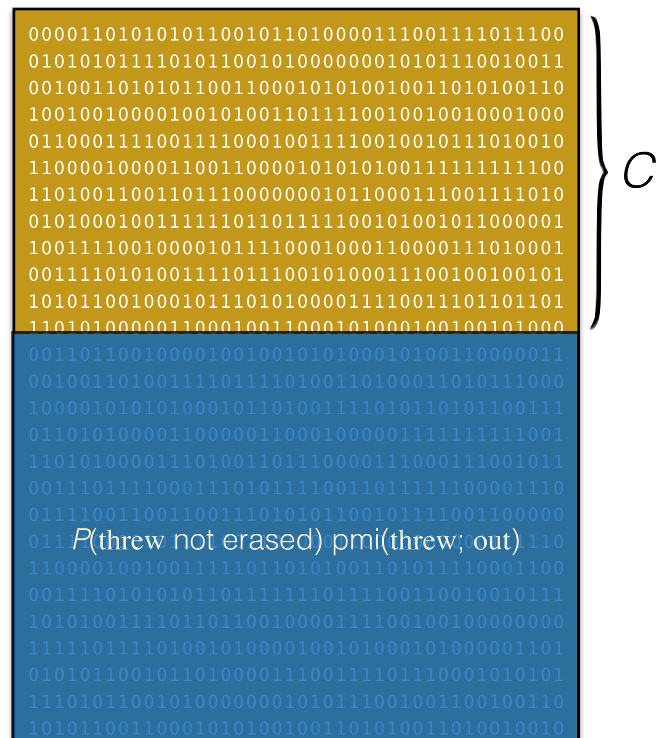
Derivation of Information Locality

- Noise decreases the influence of context:

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



h(out) - P(threw not erased) pmi(threw; out)



(Futrell & Levy, 2017)

Derivation of Information Locality

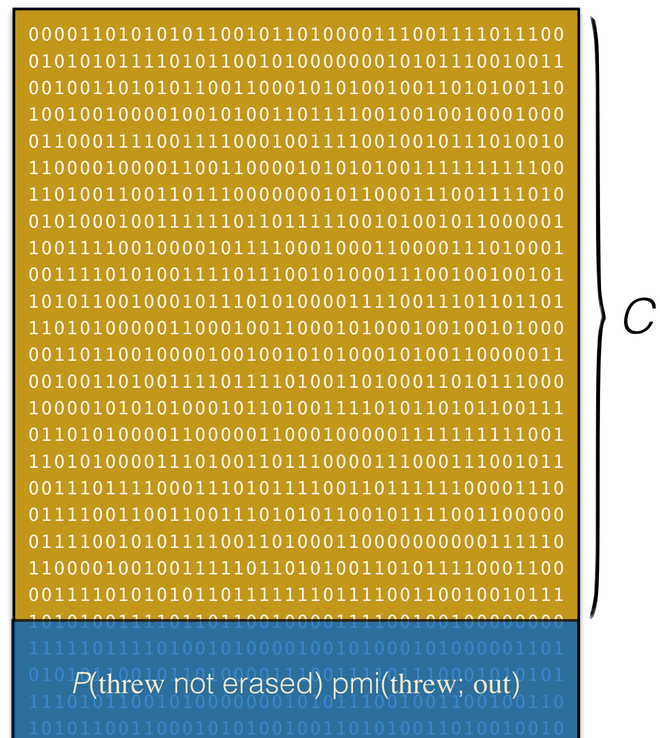
- Noise decreases the influence of context:

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

threw out

h(out) - P(threw not erased) pmi(threw; out)

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



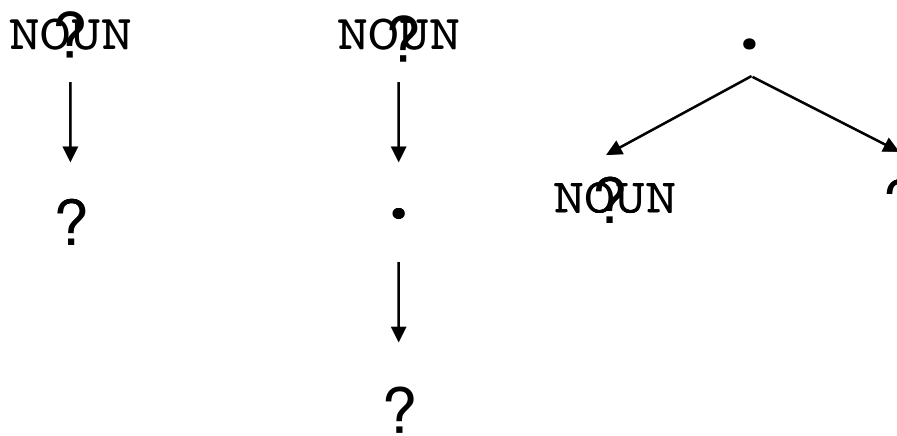
(Futrell & Levy, 2017)

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.

(Futrell & Levy, 2017)

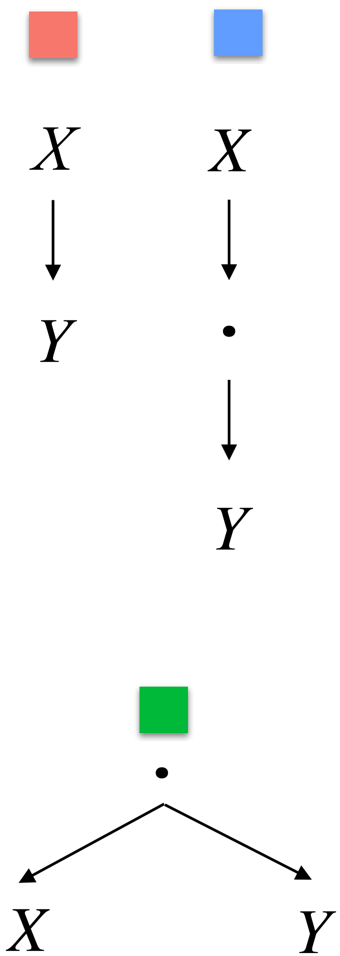
Do Dependencies Have High Mutual Information?



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

(Futrell & Levy, 2017)

Do Dependencies Have High Mutual Information?



(Futrell & Levy, 2017)

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

(Bicknell & Levy, 2010, 2012ab)

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

(Bicknell & Levy, 2010, 2012)

A first-cut behavior policy

- Actions: *keep fixating*; *move the eyes*; or *stop reading*
- Simple behavior policy with two parameters: α and β
- 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*

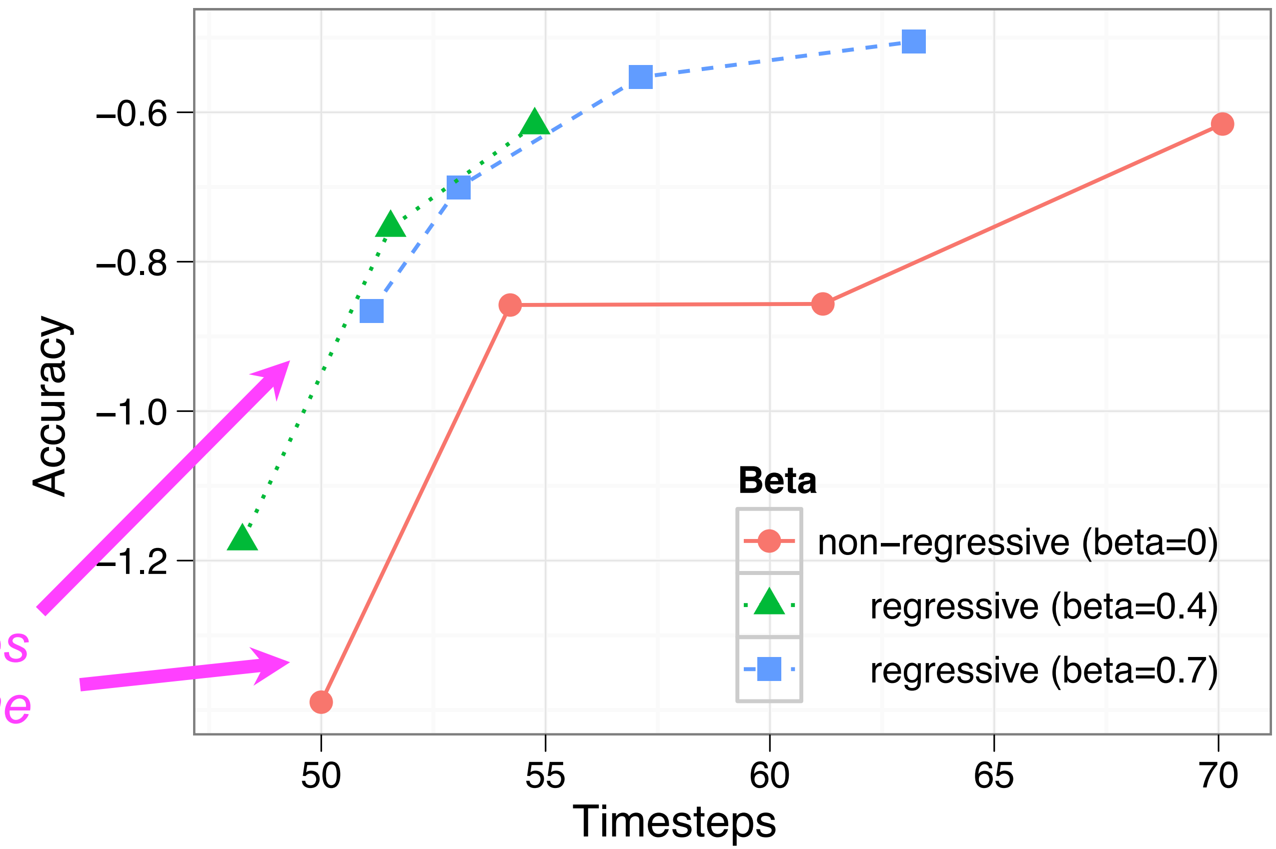
Confidence=0.59

```
P(jacket)=0.38
P(racket)=0.59
P(packet)=0.02
...
```

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

(Non)-regressive policies

- *Non-regressive policies* have *β=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 always beaten by some regressive policy

Goal-based adaptation

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

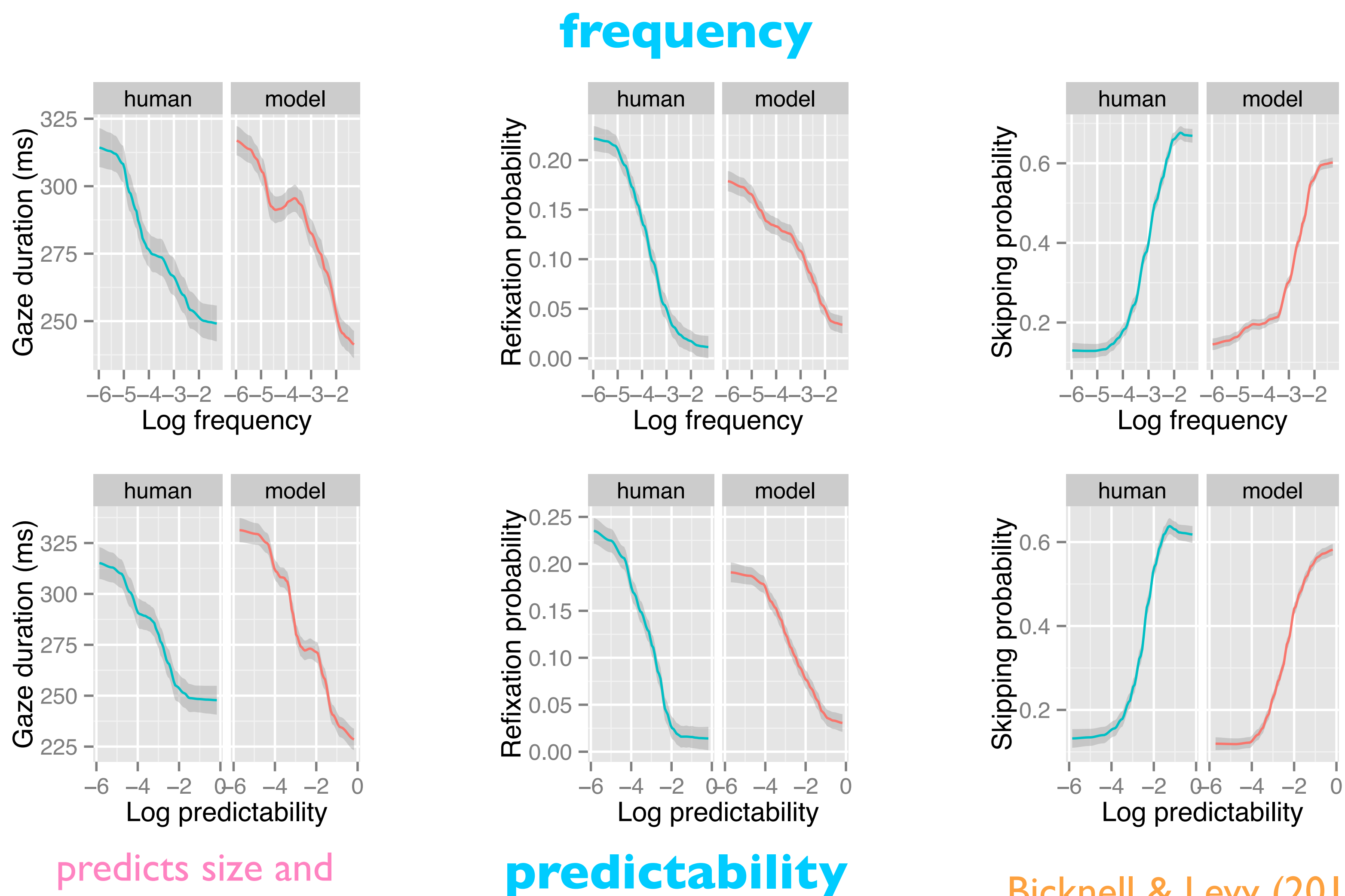
γ	α	β
0.025		
0.1		
0.4		

- The method works, and gives intuitive results

(Bicknell & Levy, 2010)

Empirical match with human reading

- Benchmark measures in eye-movement modeling:



predicts size and shape of all effects

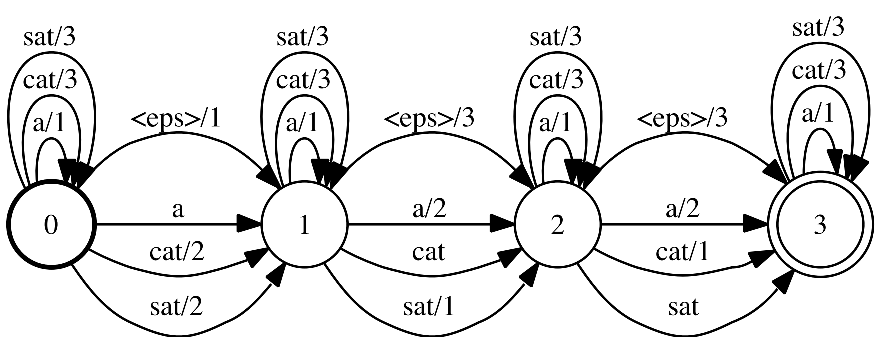
Bicknell & Levy (2012)

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.