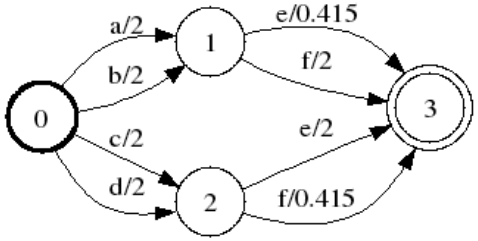
UCSD

A Fully Rational Model of Local-coherence Effects

Modeling Uncertainty about the Linguistic Input in Sentence Comprehension



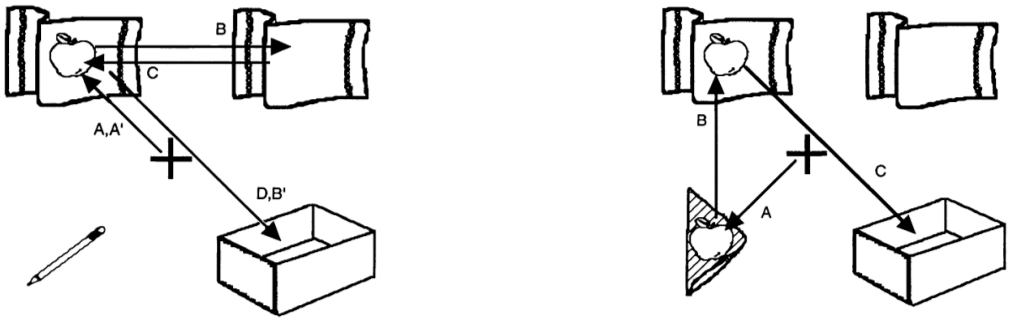
Roger Levy

UC San Diego

CUNY — 14 March 2008

Incrementality and Rationality

- 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 the information available*, whenever possible
- This is what *incrementality* is
- We have lots of evidence that people do this often



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

But…what do you think of this sentence?

The coach smiled at the player tossed the frisbee.

…and contrast this with:

The coach smiled at the player thrown the frisbee.

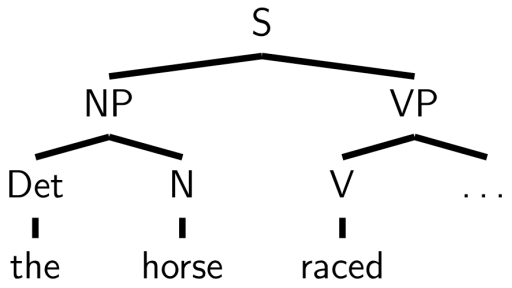
The coach smiled at the player who was thrown the frisbee.

(Tabor et al., 2004)

Why is this sentence so interesting?

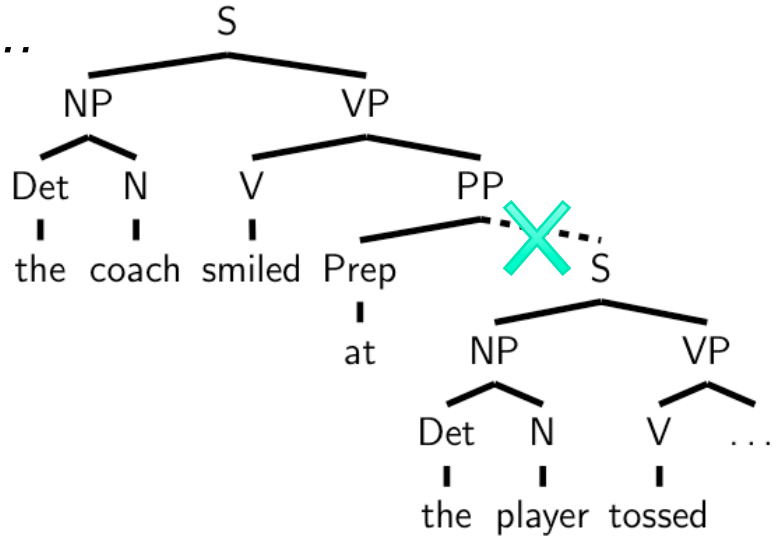
- As with classic garden-path sentences, a part-of-speech ambiguity leads to misinterpretation
  - *The horse raced past the barn…*

verb?
participle?



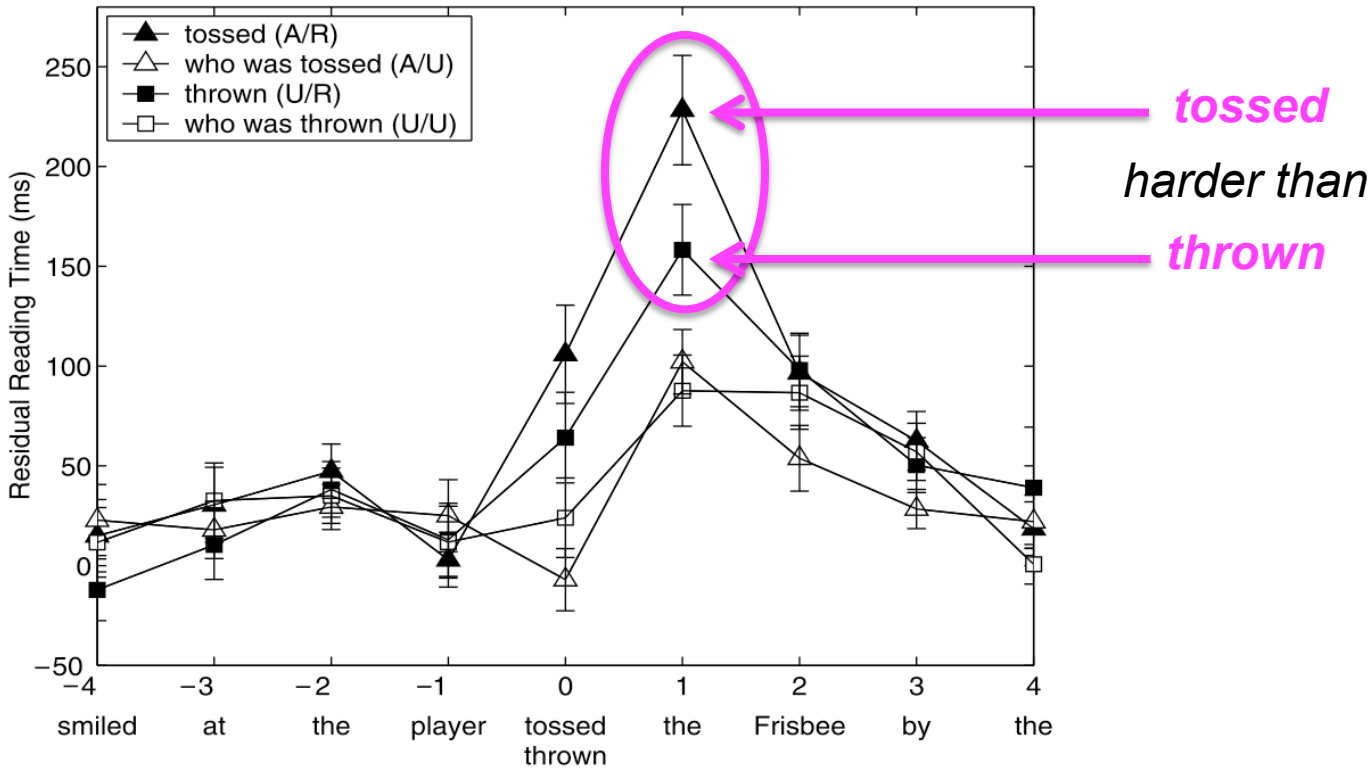
- But here, context "should" be enough to avoid garden-pathing
  - *The coach smiled at the player tossed…*

verb?
participle?



- Yet the main-verb POS "interferes" with processing

Behavioral correlates (Tabor et al., 2004)



- Also, Konieczny (2006, 2007) found compatible results in stops-making-sense and visual-world paradigms
- ***These results are problematic for theories requiring global contextual consistency*** (Frazier, 1987; Gibson, 1991, 1998; Jurafsky, 1996; Hale, 2001, 2006)

Contextual constraint & rationality

- Let's recast the problem in even more general terms
- *Rational* models of comprehension: the comprehender uses *all the information currently available*
- In local-coherence sentences, the comprehender seems to be systematically *ignoring* available information
- Local-coherence effects' challenge: ***to what extent is human sentence comprehension rational?***

Existing proposed theories

- Proposed models posit a context-ignoring, *bottom-up* component of comprehension:
  - Gibson, 2006

$$P(cat_i|w_i, \text{context}) \propto P(cat_i|\text{context})P(cat_i|w_i)$$

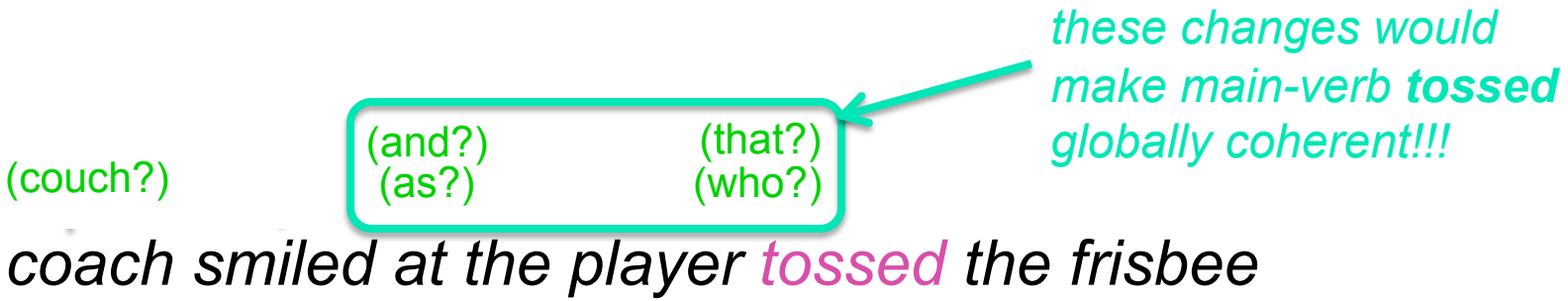
  - Tabor & Hutchins, 2004; Tabor, 2006
  - Hale, 2007
- To the extent that these models are rational, it can only be in terms of "bounded rationality" (Simon 1957)
- ***To what extent do we want to bound the rationality of human sentence comprehension?***

Today's proposal

- I simply want to argue that it is premature to conclude from local-coherence effects that the parser's rationality *must* be bounded in this respect
- There is another possibility that has been overlooked thus far
- Instead of relaxing the assumption about *rational use of context*, we ***may*** instead have misspecified the *input representation*

Relaxing assumptions about input

- Traditionally, the input to a sentence-processing model has been a *sequence of words*
- But really, input to sentence processor should be more like the output of a word-recognition system



- That means that the possibility of *misreading/ mishearing* words must be accounted for
- On this hypothesis, local-coherence effects are about *what the comprehender wonders whether she might have seen*

Inference through a noisy channel

- So how can we model sentence comprehension when the input is still noisy?
- A *generative probabilistic grammatical model* makes inference over uncertain input possible
  - This is the ***noisy channel*** from NLP/speech recognition
  - Inference involves Bayes' Rule

Prior: *Comprehender's knowledge of language*

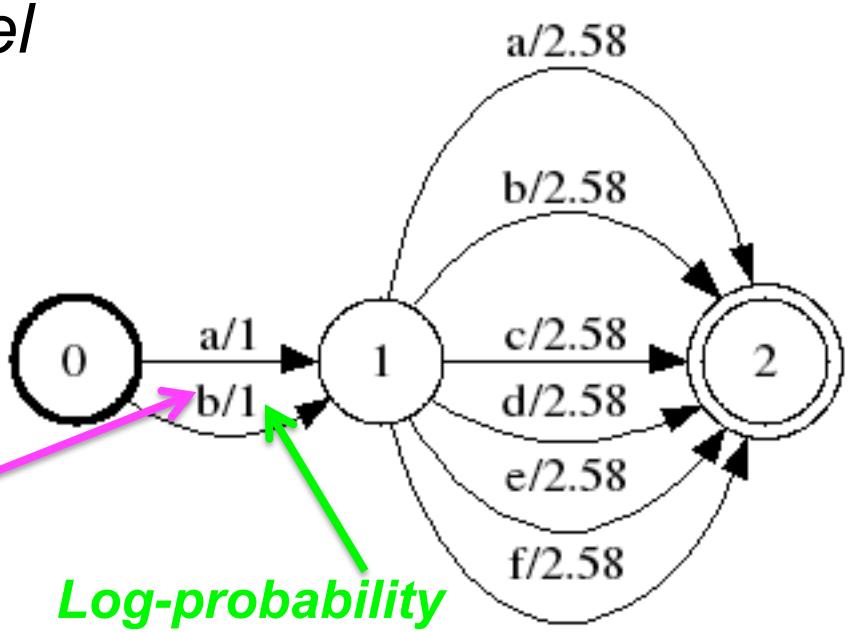
$$P(\text{words}|\text{input}, \text{grammar}) \propto P(\text{input}|\text{words}, \text{grammar})P(\text{words}|\text{grammar})$$
$$\propto P(\text{input}|\text{words})P(\text{words}|\text{grammar}) \quad [\text{by assumption}]$$

Evidence: *Noisy input probability, dependent only on the "words" generating the input*

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



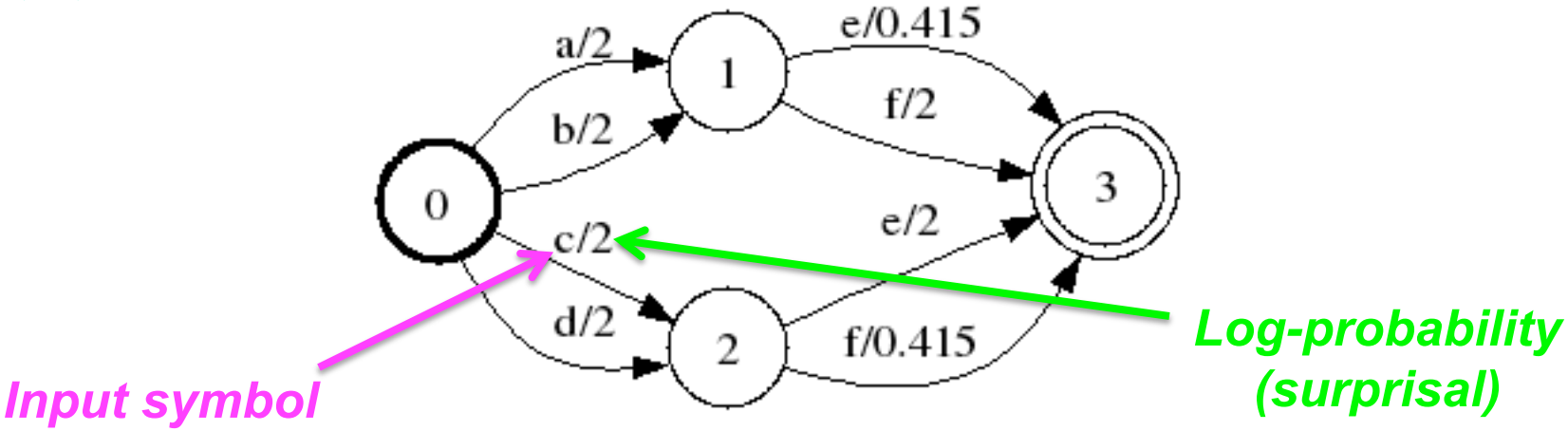
Input symbol

Log-probability (surprisal)

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

Probabilistic Linguistic Knowledge

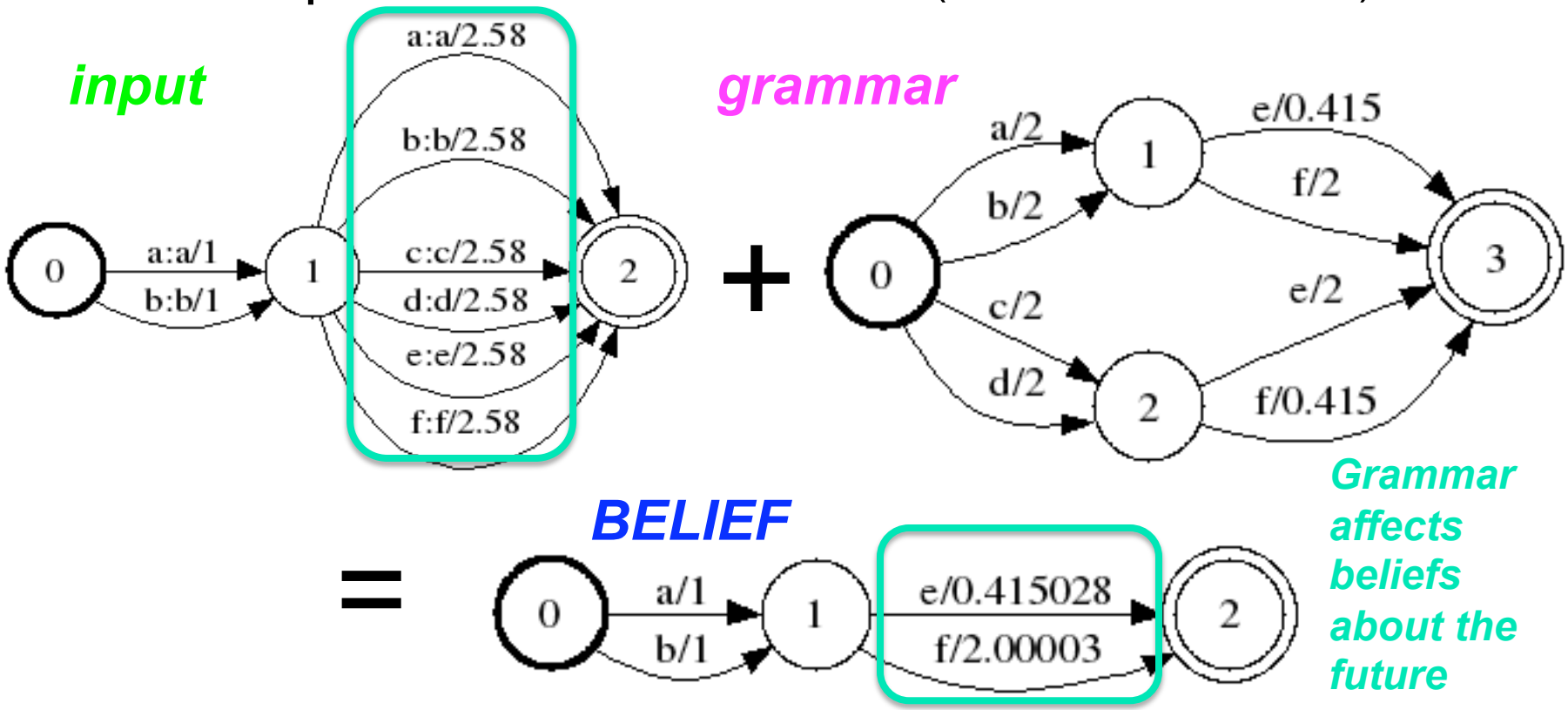
- A generative probabilistic grammar determines beliefs about *which strings are likely to be seen*
  - Probabilistic Context-Free Grammars (PCFGs; Booth, 1969)
  - Probabilistic Minimalist Grammars (Hale, 2006)
  - Probabilistic Finite-State Grammars (Mohri, 1997; Crocker & Brants 2000)



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

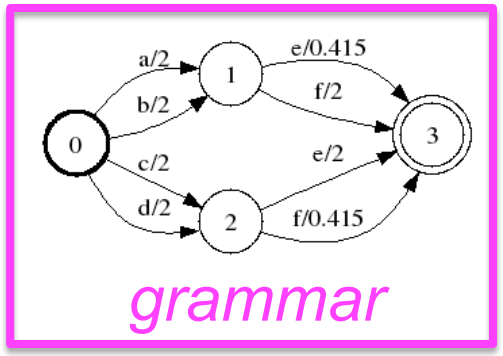
Combining grammar & uncertain input

- Bayes' Rule says that the *evidence* and the *prior* should be combined (multiplied)
- For probabilistic grammars, this combination is the formal operation of *intersection* (see also Hale, 2006)



Revising beliefs about the past

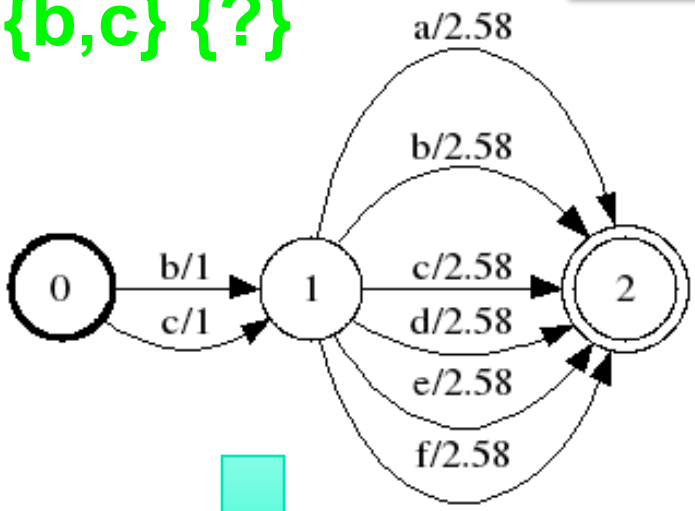
- When we're uncertain about the future, grammar + partial input can affect beliefs about what will happen
- With uncertainty of the past, grammar + future input can affect beliefs about *what has already happened*



grammar

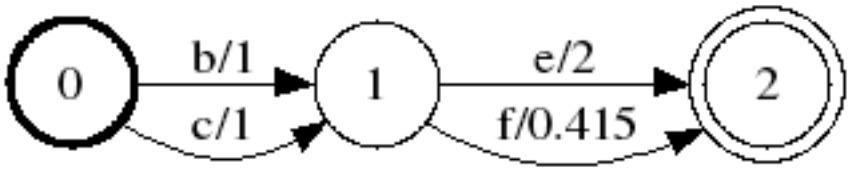
word 1

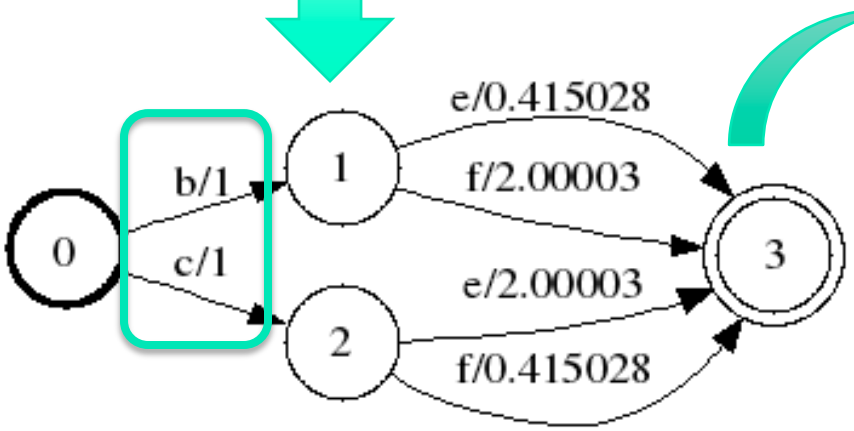
{b,c} {?}

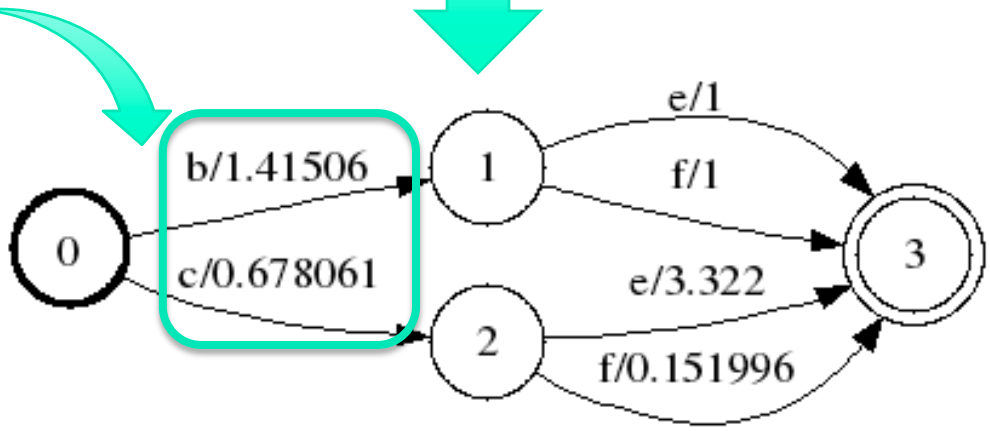


words 1 + 2

{b,c} {f,e}







Ingredients for the model

- To complete our rational model of local coherence effects, we need the following components:
  - A probabilistic **grammar**
  - A systematic **mapping** from sentences to noisy (pFSA) inputs
  - A **quantified signal** of the alarm about representations of the past that is induced by the current word
- I'll present these ingredients in the form of an **experiment** on the "classic" local-coherence sentence

1. Probabilistic Grammatical Knowledge

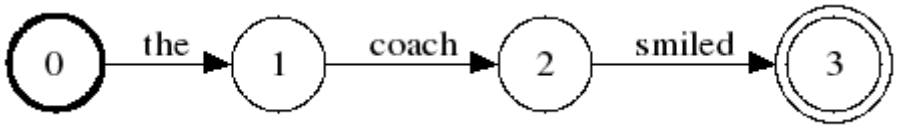
- We can transform a (strongly regular) PCFG into a weighted FSA
- We use the following grammar with surprisal values estimated from the parsed Brown corpus

```
7.30 S -> S-base Conj S-base
0.01 S -> S-base
0.00 S-base -> NP-base VP
3.71 NP -> NP-base RC
0.11 NP -> NP-base
0.00 NP-base -> Det N
2.02 VP -> V PP
0.69 VP -> V NP
2.90 VP -> V
0.00 PP -> P NP
```

```
0.47 RC -> WP S/NP
2.04 RC -> VP-pass/NP
4.90 RC ->
        WP FinCop VP-pass/NP
0.74 S/NP -> VP
1.32 S/NP -> NP-base VP/NP
3.95 VP/NP -> V NP
0.10 VP/NP -> V
2.18 VP-pass/NP -> VBN NP
0.36 VP-pass/NP -> VBN
```

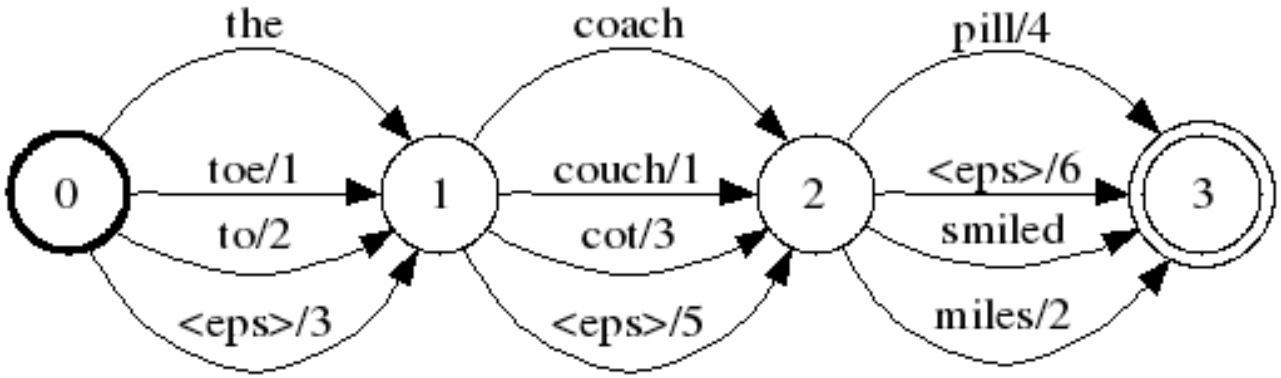
2. Sentence → noisy input mapping

- There are lots of possibilities here
- Our implementation: start with the sequence of actually observed words



- Make every lexical item (including <eps>) probable in proportion to ***Levenshtein*** *(string-edit)* ***distance***

```
Dist(dog, cat)     = 3    Dist(the, toe)    = 1
Dist(<eps>, toes) = 4    Dist(goth, hot)   = 2
```



3. Error identification signal (*EIS*)

- *Relative Entropy* (KL-divergence) is a natural metric of change in a probability distrib. (Levy, 2008; Itti & Baldi, 2005)
- In our case, the distributions in question are *probabilities over the previous words in the sentence*
- Call this distribution $P_i(w_{[0,j)})$
- The size of the change in this distribution induced by the *i*-th word is EIS_i, defined as

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

new distribution *old distribution*

Error identification signal: local coherences

- Full experiment:
  - Probabilistic grammar with rule probabilities estimated from parsed Brown corpus
  - Lexicon with all `<tag,word>` combinations of frequency >500 in parsed Brown corpus (plus sentence wds)
  - Error identification signal as defined above

  - The important part of the change is that ***at*** can be re-interpreted as ***and*** or other near-neighbors

But, you may protest…

- Most items in Tabor et al., 2004 did not involve the preposition ***at*** before the modified noun
- For example:

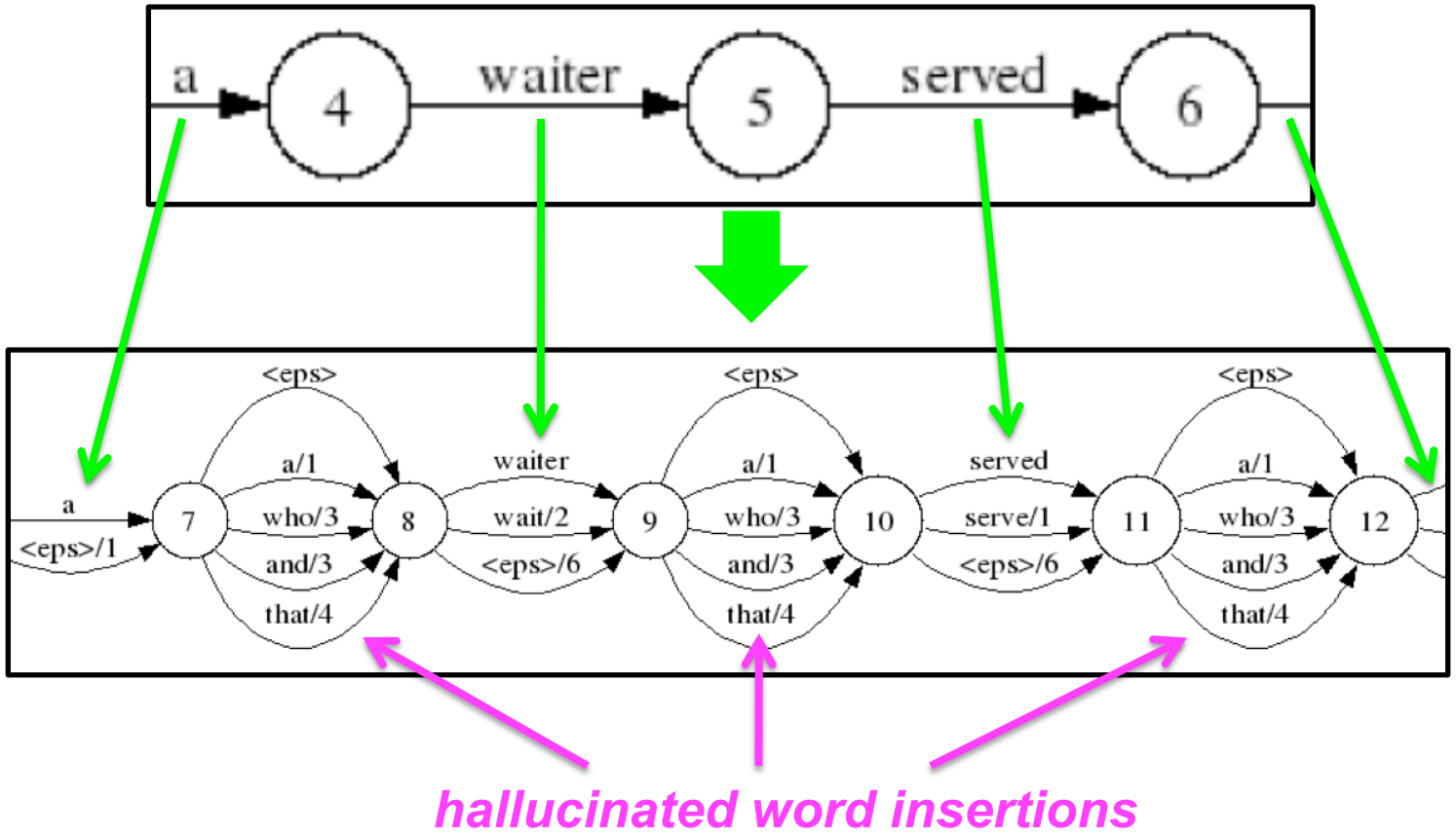
who/that/and

The manager watched a waiter served/given pea soup by a trainee.

- But these sentences can also involve revisions of past beliefs—specifically, *that a word has been missed*

Missed words

- Modeling beliefs about missed words requires only a minor modification to the noisy-input representation



Missed words (II)

- Experiment 2: stimulus without the preposition ***at***

EIS = 0.0168

The manager watched a waiter served…

The manager watched a waiter given…

EIS = 0.0117

- The difference in error-identification signal is much smaller, but we still get it

Other potential applications of theory

- "Good-enough" processing representations (Ferreira et al., 2002)

  While Anna dressed the baby played in the crib.

- "Morphological mismatch" processing effects in cases of superficial semantic mismatch (Kim & Osterhout, 2005)

  *The meal **devoured…***

- Modeling longer-distance regressions in reading of naturalistic text

What the model is still missing

- Lots of things! But a couple of things most sorely missed:
  - Trans-finite-state probabilistic rules (technical, not theoretical shortcoming)
  - Richer probabilistic information sources, such as *plausibility* of noun-verb match (statistical, not theoretical shortcoming)

more difficult

(Tabor et al., 2004)

*The bandits worried about the **prisoner** transported…*

*The bandits worried about the **gold** transported…*

less difficult

Acknowledgements

- Thank you for all *the people that* your help!
  - UCSD Computational Psycholinguistics Lab:
    - Adam Bickett
    - Klinton Bicknell
    - Gabriel Doyle
    - Albert Park
    - Nathaniel Smith
  - Cyril Allauzen and the rest of the developers of the OpenFst (*http://openfst.org*) and GRM libraries *(http://www.research.att.com/~fsmtools/grm/)*

UCSD

Thank you for listening!

http://idiom.ucsd.edu/~rlevy