computational psycholinguistics

CS 585, Fall 2019

Introduction to Natural Language Processing

http://people.cs.umass.edu/~miyyer/cs585/

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

some slides adapted from Roger Levy

computational psycholinguistics:

how do humans **comprehend**, **produce**, and **acquire** language?

how can computational methods help us learn more about these processes?

ok… how can computers help?

human behavior is super complicated! we don't understand how the brain really even works.

we can encode many simplifying assumptions in a computational model such that analyzing the model is much more tractable

let’s say we want to study *disfluencies*

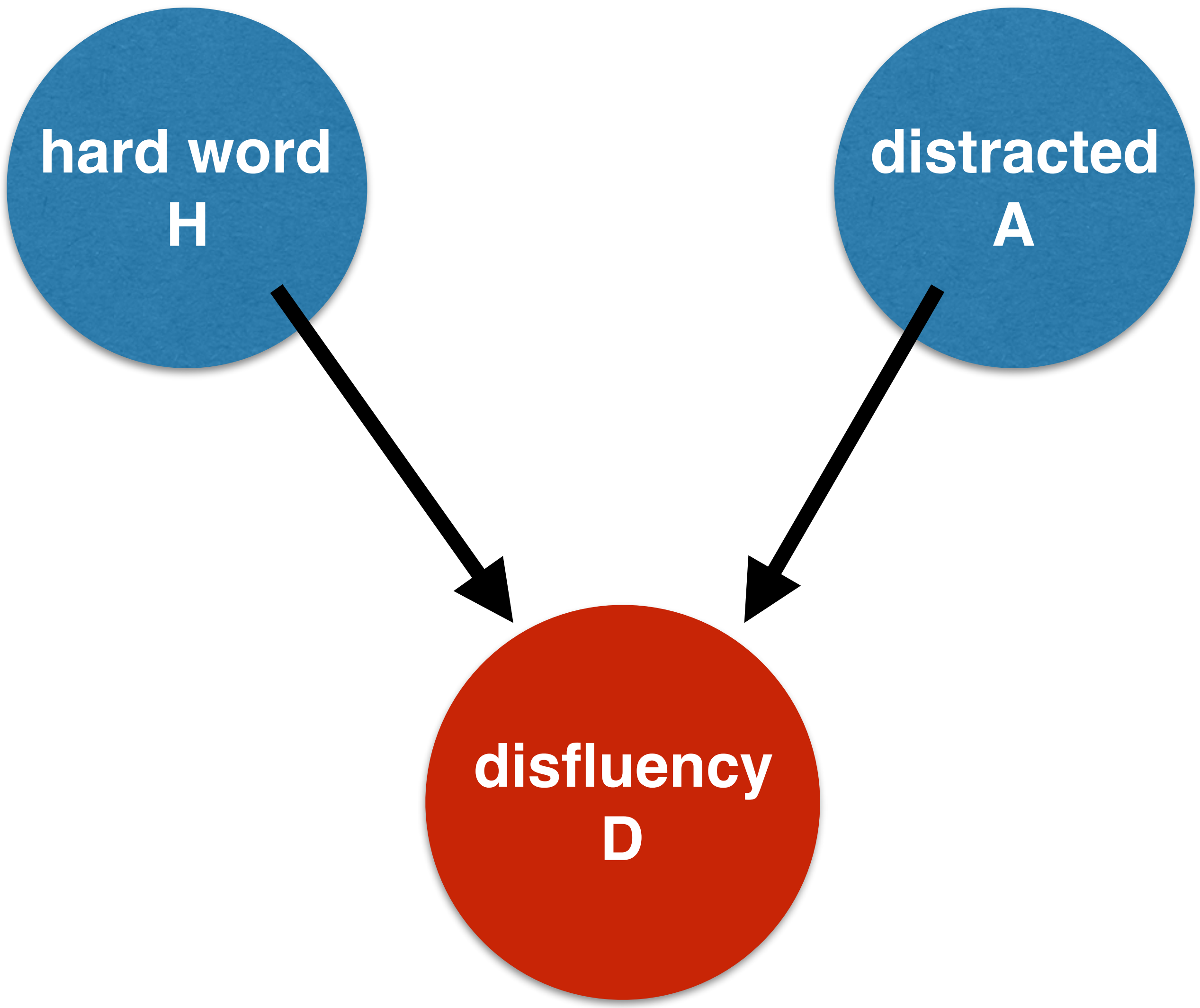
I read a book about, uh…

what could cause a person to produce disfluencies?

lots of reasons! let’s simplify:

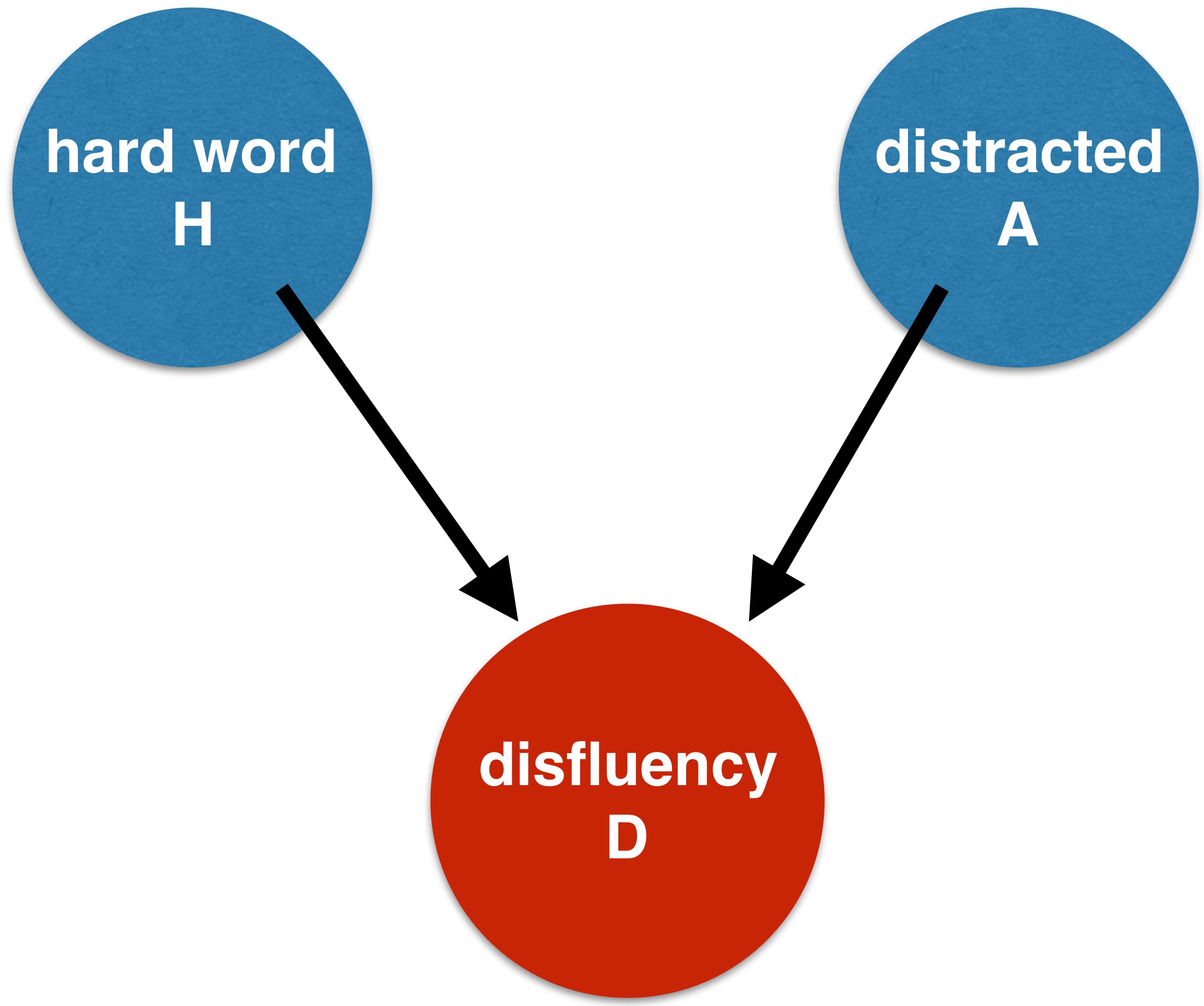
- disfluencies are caused by either:
    - the upcoming word being hard to produce, maybe because its long or low-frequency (e.g., *astrolabes*)
    - the speaker was distracted by something while they were in the middle of a sentence

a simple graphical model



$$P(H, A, D) = ???$$

a simple graphical model



$$P(H, A, D) = P(H)P(A)P(D \mid H, A)$$

design a human experiment

W	A	D = no disfluency	D=disfluency
easy	undistracted	0.99	0.01
easy	distracted	0.7	0.3
hard	undistracted	0.85	0.15
hard	distracted	0.4	0.6

can answer questions like:
if the speaker uttered a disfluency, what is the probability that the word was hard?

computational model of human sentence processing

- any such model must at least:
  - be robust to arbitrary inputs
  - figure out the most likely interpretation in cases of ambiguity
  - be able to do inference on incomplete inputs

computational model of human sentence processing

- any such model must at least:
  - be robust to arbitrary inputs
  - **figure out the most likely interpretation in cases of ambiguity**
  - be able to do inference on incomplete inputs

let's assume humans have a PCFG in their brains. what experiments can we use to test the parsing algorithm they use?

standard psycholinguistics experiments

- *behavioral* experiments:
  - what choices do people make in various language-producing and language-comprehending situations?
  - how long do they take to make these choices?
- *offline* experiments:
  - have people rate or complete sentences
- *online* experiments:
  - track eye movements, have people read aloud, have them read under time pressure, measure their brain activity with e.g., EEG, etc.

human sentence comprehension

- The women discussed the dogs on the beach

what does *on the beach* modify?

- The women kept the dogs on the beach.

what does *on the beach* modify?

human sentence comprehension

- The women discussed the dogs on the beach

what does *on the beach* modify?

dogs (90%), discussed (10%)

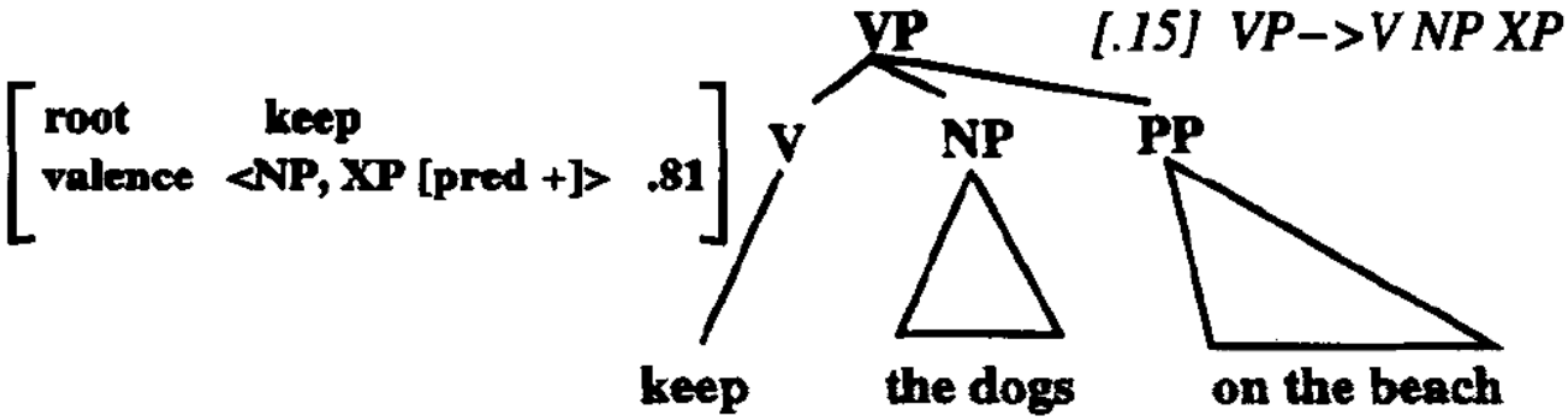
- The women kept the dogs on the beach.

what does *on the beach* modify?

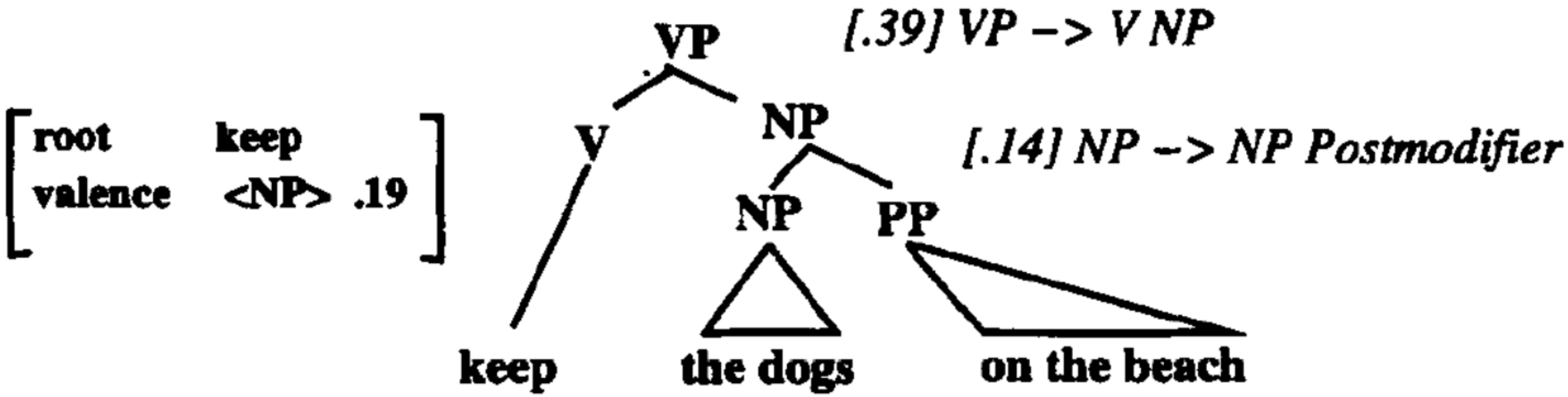
dogs (95%), kept (5%)

Ford et al., 1982

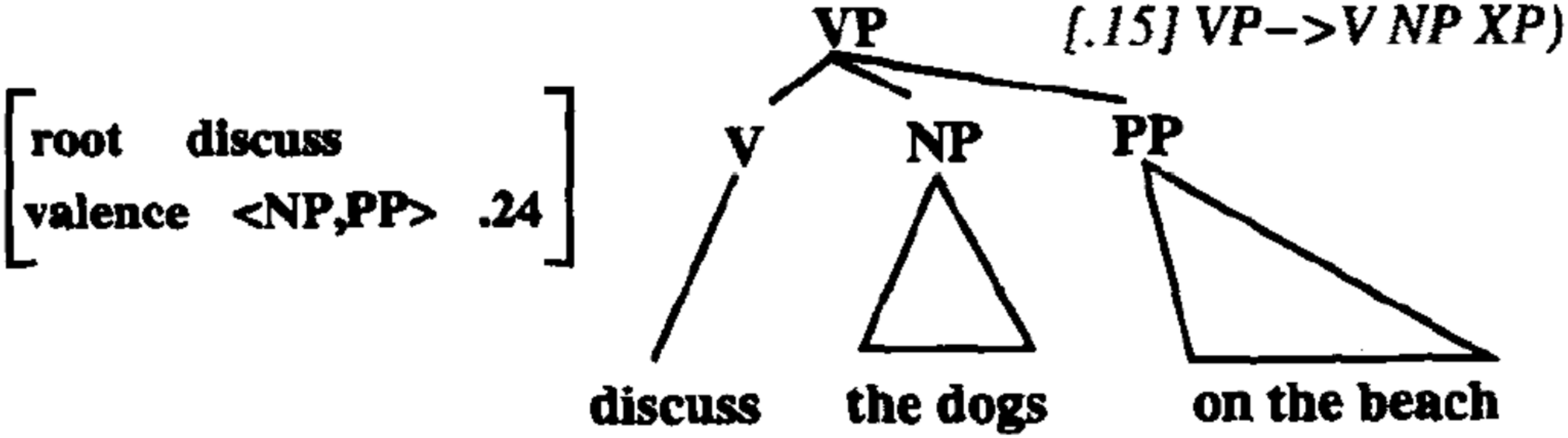
what does a parser think about these sentences?



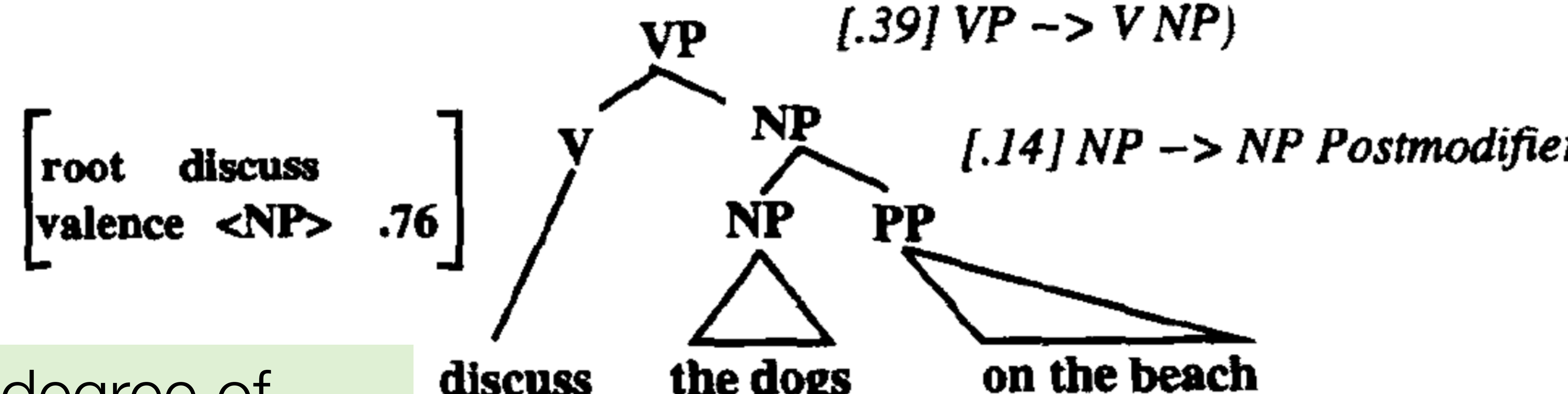
(a) .15 * .81 = .12 (prefered)



(b) .19 * .39 * .14 = .01 (disprefered)



(a) .15 * .24 = .036 (disprefered)



(b) .76* .39 * .14 = .041 (prefered)

degree of preference not matched!

exercise!

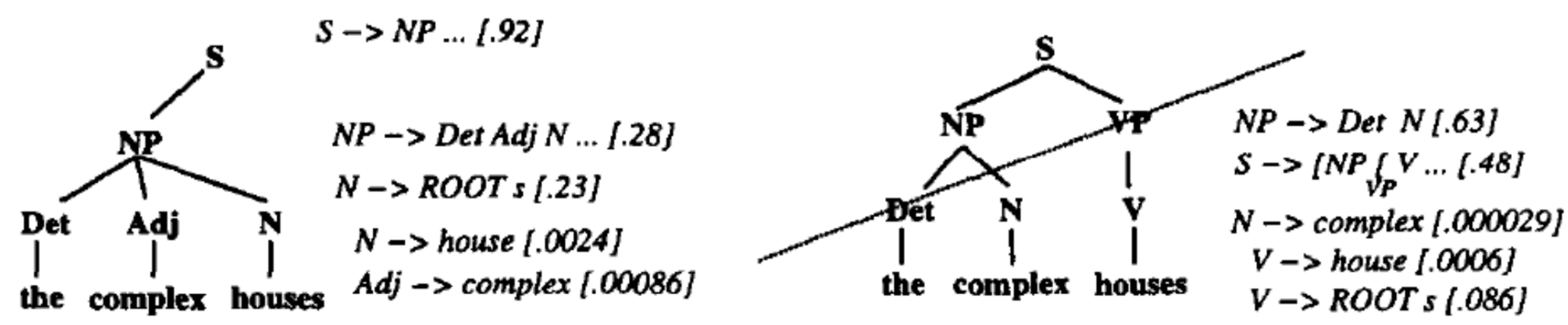
garden path sentences provide a way to test human parser processing

- how many parses does a human keep in memory while reading a sentence?
  - **full serial**: keep only one parse at all times
  - **full parallel**: keep all possible parses
  - **limited parallel**: keep some but not all parses

does this sound similar to any algorithms that we've discussed?

garden path effects can arise in the limited-parallel setting!

- The complex houses married and single students and their families.
  - **[S [NP The complex] [VP houses…] …] discarded :(**
  - **[S [NP The complex houses …] …] kept**



(a) (prefered) $1.2 \cdot 10^{-7}$ (b) (disprefered) $4.5 \cdot 10^{-10}$

human brains react differently to surprising and predictable words

The squirrel stored some nuts in the { tree / fridge }

Predictable words are **read faster** and have distinctive EEG responses

Kutas & Hillyard, 1980

Ehrlich & Rayner, 1981

The squirrel stored some nuts in the $\left\{ \begin{array}{l} \text{tree} \\ \text{fridge} \end{array} \right.$

how do we computationally quantify "surprisal"?

use a language model!

$$\text{surprisal}(w_i) = \log \frac{1}{P(w_i \mid w_{1 \ldots i-1})}$$

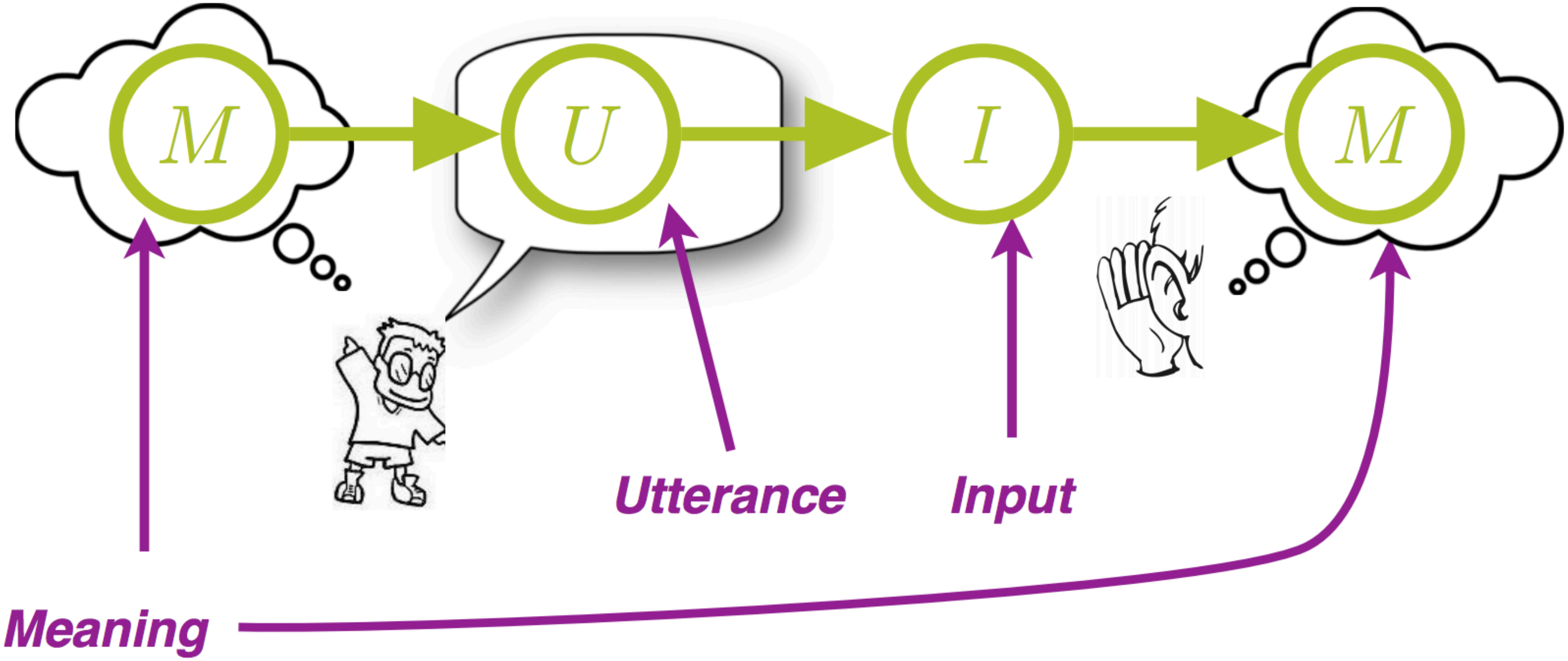
comprehension > production

- comprehension:

$$P(\text{meaning} \mid \text{input}, \text{context})$$

- production:

$$\min \text{cost}(\text{utterance} \mid \text{meaning}, \text{context})$$



what factors determine the “cost” of an utterance?

what factors determine the "cost" of an utterance?

- utterance should convey the intended meaning
- utterance should be succinct to avoid wasting time
- minimize effort on both the speaker and listener

intended meaning:
i'd like a beer!

i'd like a beer

where can i get a beer?

[mime beer drinking]

it's Miller time!

i'm in Germany!

Garr!!!!!

how do we decide between multiple plausible utterances?

- Terry gave the exhausted traveller from France a silver dollar.
- Terry gave a silver dollar **to** the exhausted traveller from France.

- The least we should do is make it as much fun as possible.
- The least we should do is **to** make it as much fun as possible.

let's look closely at the *dative alternation*

prepositional dative structure:	... *gave* [*toys*] [*to the children*]	V NP PP
double object structure:	... *gave* [*the children*] [*toys*]	V NP NP

why should we use one over the other?

theory 1: subtly different semantics

- Prepositional dative signals *transfer of location*
- Double object signals *transfer of possession*
  - I sent storage a book (double object, *storage* is animate)
  - I sent a book to storage (dative, *storage* is inanimate)
  - That movie gave me the creeps
  - That movie gave the creeps to me

the rom gorped the blick to the dax

how likely is *gorping* to involve moving something?

Kako, 2006

theory 1: subtly different semantics

- Prepositional dative signals *transfer of location*
- Double object signals *transfer of possession*
  - I sent storage a book (double object, *storage* is animate)
  - I sent a book to storage (dative, *storage* is inanimate)
  - That movie gave me the creeps
  - That movie gave the creeps to me

the rom gorped the blick to the dax

how likely is *gorping* to involve moving something?

the rom gorped the dax the blick

what about now?

Kako, 2006

theory 2: processing preferences

- Every context causes a different alignment of various preferences, which affect what kind of construction we end up producing (dative vs double object)
  - discourse-given vs. discourse-new
  - short vs long
  - definite vs indefinite
  - animate vs inanimate
  - pronoun vs full NP

Collins, 1995

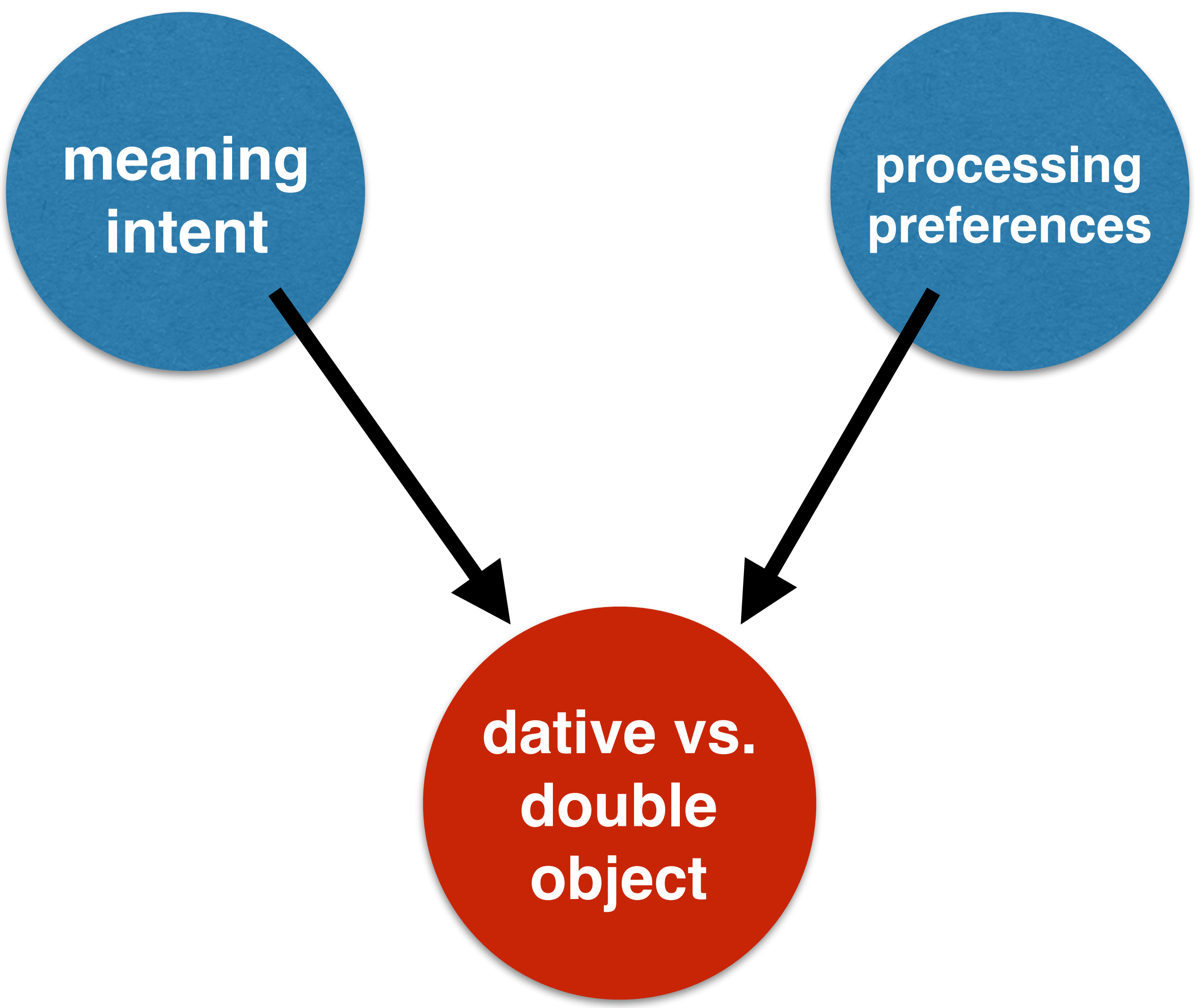
corpus analysis kinda supports that all of these factors are important

Predictor x_i	***Coefficient β_i***
log Recipient Length	1.31
log Theme Length	-1.17
Recipient Animacy	2.14
Theme Animacy	-0.92
Recipient Discourse Status	1.33
Theme Discourse Status	-1.76
Recipient Pronominality	-1.54
Theme Pronominality	2.2
Recipient Definiteness	0.8
Theme Definiteness	-1.09

Bresnan et al., 2007

how do we decide which theory is “more correct”?

- what if both are right???



let's do a controlled human experiment!

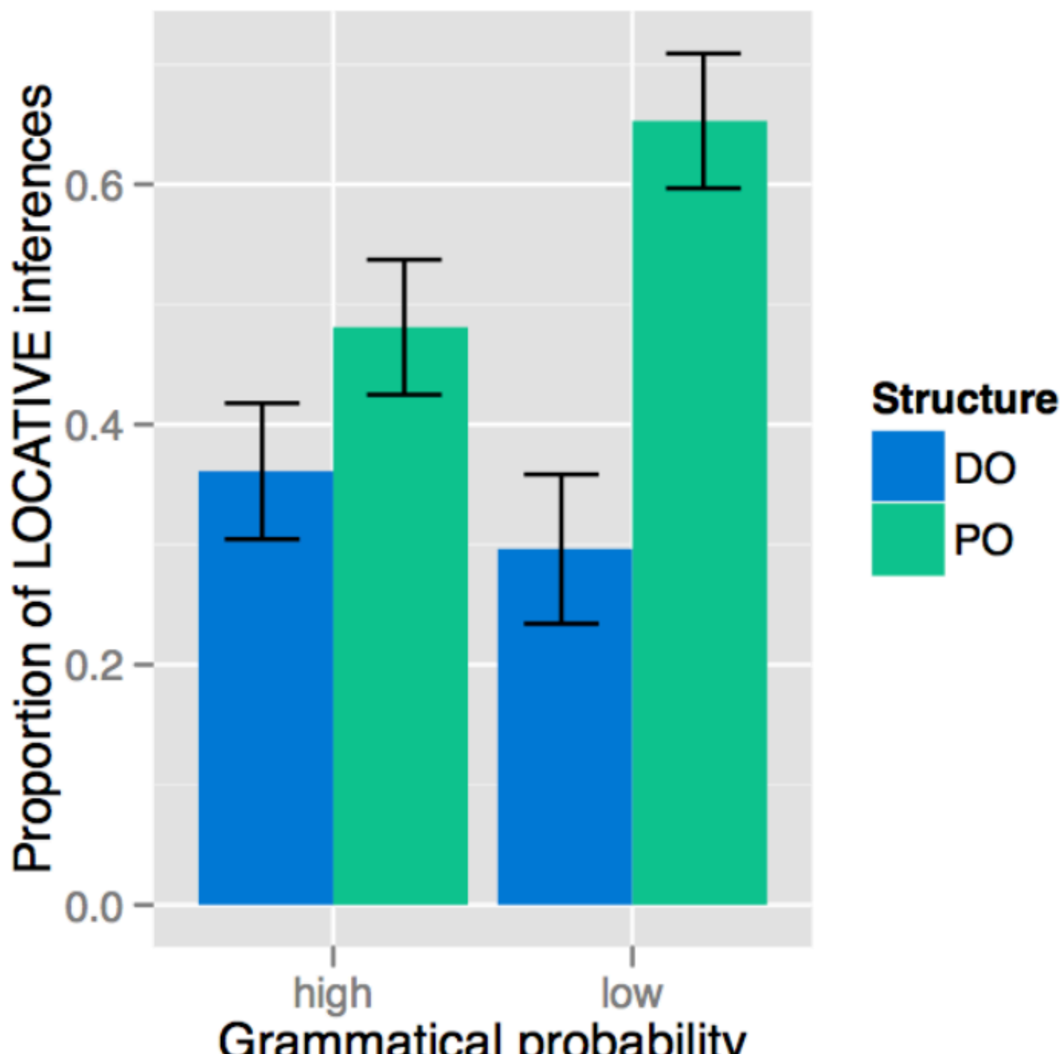
The zarg prolted the cherid to a really gromious flig .

Which is more likely?

○	The cherid is in a new place.	LOCATIVE inference
○	The cherid has a new owner.	POSSESSIVE inference

Myslin & Levy, 2015

Sentence	S	$P(S\|G)$
The zarg prolted [the cherid] to [a really gromious flig].	PO	high
The zarg prolted [the flig] [a really gromious cherid].	DO	high
The zarg prolted [a really gromious cherid] to [the flig].	PO	low
The zarg prolted [a really gromious flig] [the cherid].	DO	low



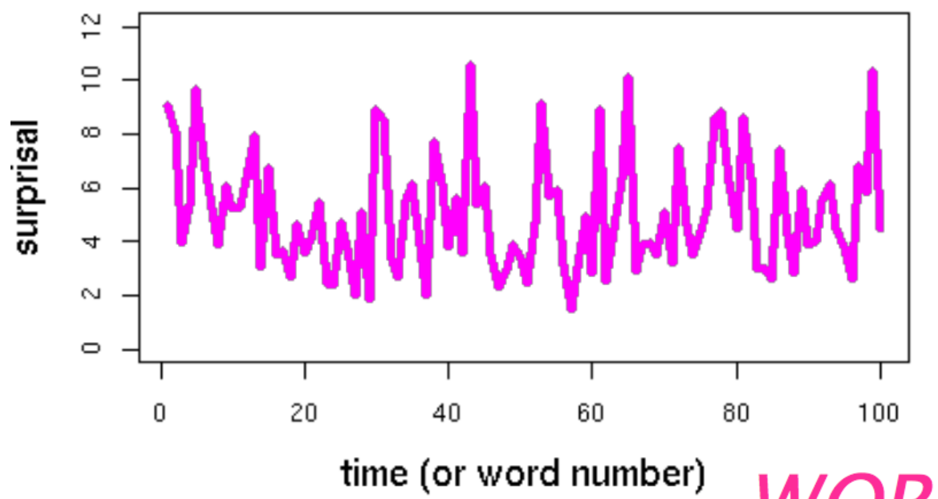
results:
there are *both* subtle meaning differences and processing preferences

producing language in adverse conditions

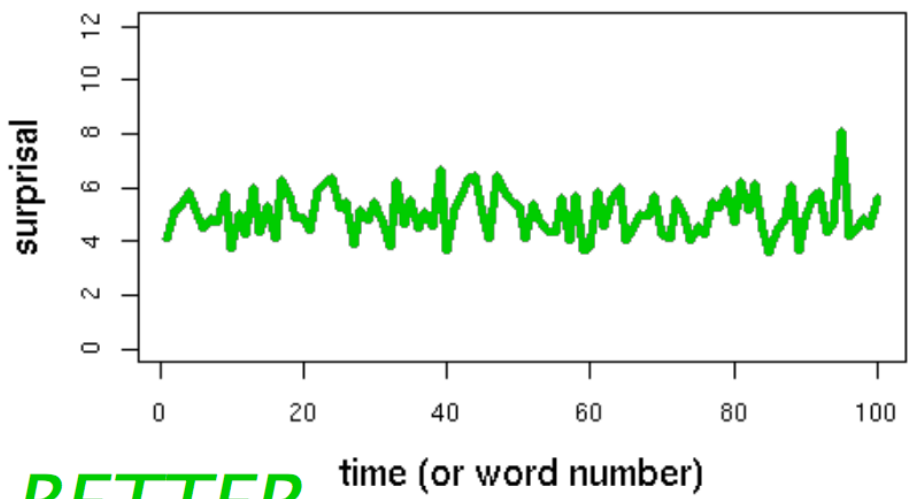
- often we cannot control the environment in which we produce language.
  - in addition to noise / external distractions, people have limited attention spans and you may not know the person you're speaking with very well
- despite this, we still manage to communicate pretty well most of the time…. how do we manage this? how do we achieve *redundancy* in such conditions?

uniform information density

- spreading out information evenly in a sentence minimizes total comprehension difficulty!



WORSE



BETTER

Levy & Jaeger, 2007

why do we use *that* sometimes?

Certain types of *relative clauses* (RC) in English are optionally introduced by the “meaningless” word *that*

How big is the family (that) you cook for __ ?

RC

modifies the noun family

“you cook for the family”

in a relative clause without *that*, the first word of the RC has two functions:

How big is the family **you** …

1. it signals that an RC has begun

2. it provides some information about the content of the RC

inserting *that* separates these two things

under what conditions should we use *that*?

how do humans acquire language?

- two extremes:
    - “we’re born with it”: we have a built-in mechanism in our brains that allows us to rapidly pick up language
    - “we learn it from scratch”: language is entirely learned from hearing and imitating the environment

- if the latter, then “how does something come out that does not go in?”

Chomsky’s universal grammar

- a theory that all humans are born with the genetic capacity to acquire language
- children have a “language acquisition device” (LAD) in their brains. once the LAD is triggered by input (any language/speech a child hears), a child will begin the linguistic stages of development.
  - All children (with the exception of special cases of children who were isolated from speech as infants) will develop language regardless of the kind of input they receive.

some arguments for UG

- all of the world’s languages share many properties
- despite each child observing totally different inputs growing up, we all rapidly converge to approximately the same grammar

poverty of the stimulus

- a child does not receive enough data from the environment to completely learn a grammar

(1) I like this ball and you like that one.
(2) I like this red ball and you like that one.

- in (1), “one” refers to “ball”. in (2), “one” means “red ball” but could also refer to “ball” in general.
- Like adults, **18-month-olds** show that they prefer the “red ball” interpretation

poverty of the stimulus

- Binding theory

  (1) While he was dancing, the Ninja Turtle ate pizza.
  (2) He ate pizza while the Ninja Turtle was dancing.

- in (1), *he* can refer to *Ninja Turtle*, whereas in (2) this interpretation is invalid

- both sentences were shown to preschoolers after a puppet show (either with a Ninja Turtle eating pizza or someone else eating pizza)

what does a UG look like?

- is there a dictionary and grammar encoded in our brains from birth?
- is it an *inductive bias* on our learning algorithm?
- does it even exist???

no one knows :(