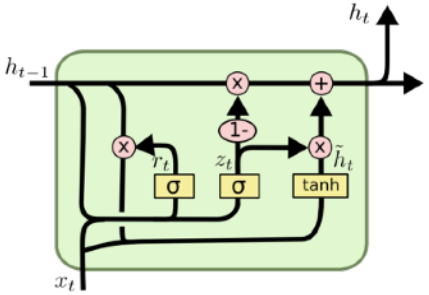
A brief introduction to human language processing and computational psycholinguistics

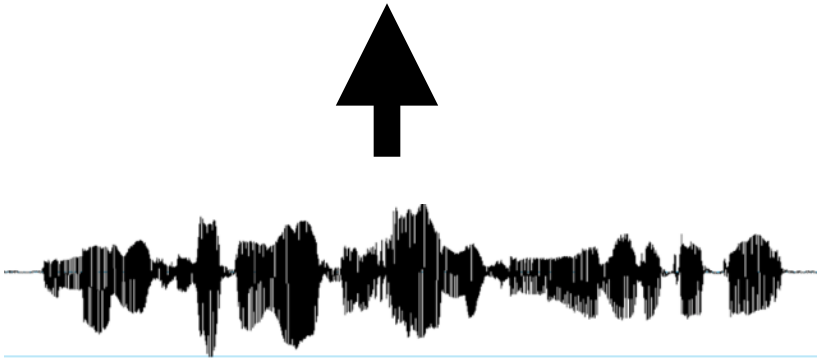
Roger Levy
Director, Computational Psycholinguistics Laboratory
Department of Brain & Cognitive Sciences

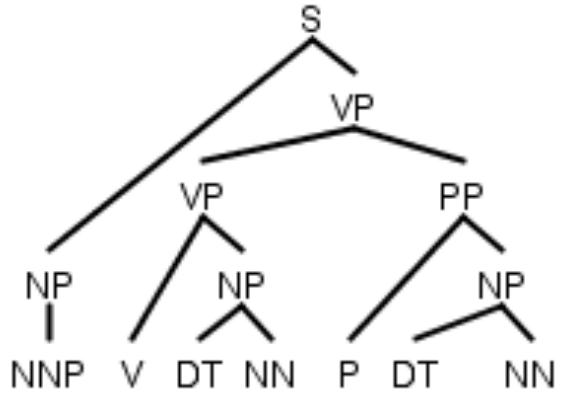
29 April 2021

Natural language processing



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

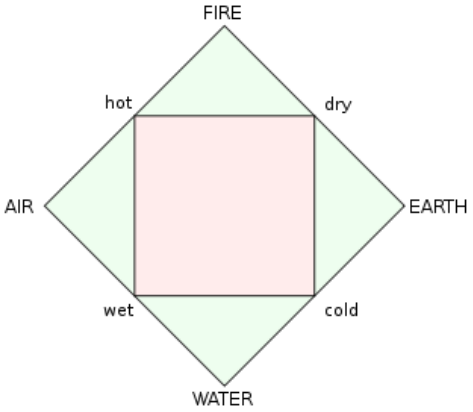


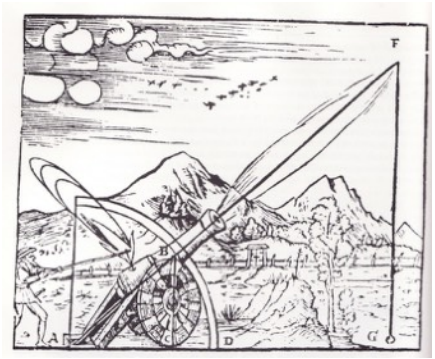


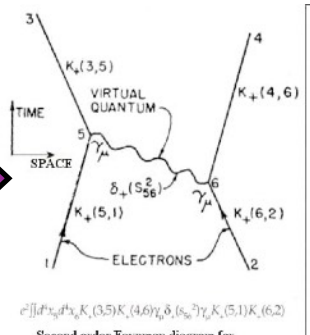
Keith saw the man with the telescope.

Progression of scientific knowledge

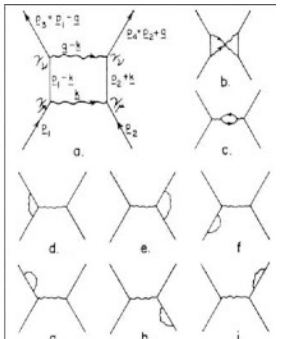
- Physics





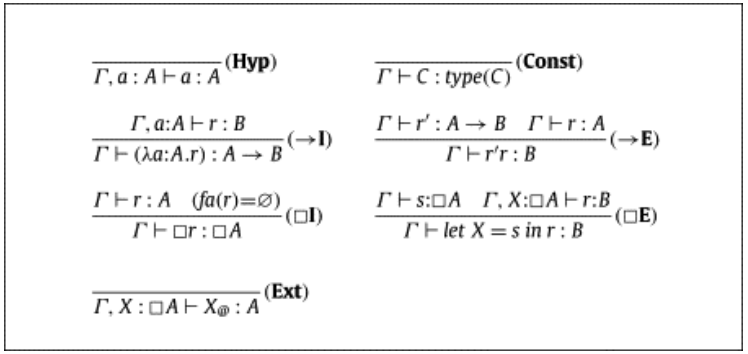


Second order Feynman diagram for electron-electron scattering and the corresponding mathematical expression


Fourth order Feynman diagrams for electron-electron scattering

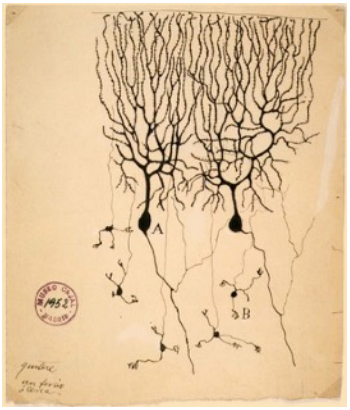
- Computer science

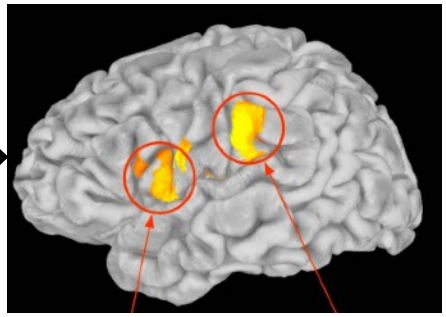
```
C     * spawn worker tasks *
      call pvmfspawn('worker', PVMDEFAULT, '', NPROCS, task_ids, numt)

      num_data = SIZE/NPROCS
C     * send data to worker tasks *
      do 20 i = 1,NPROCS
         call pvmfinitsend(PVMDEFAULT, info)
         call pvmfpack(INTEGER4, num_data, 1, 1, info)
         call pvmfpack(INTEGER4, a(num_data*(i-1)+1), num_data, 1, info)
         call pvmfsend(task_ids(i), 4, info)
 20   continue

C     * wait and gather results *
      msgtype = 7
      sum = 0
      do 30 i = 1,NPROCS
         call pvmfrecv(task_ids(i), msgtype, info)
         call pvmfunpack(INTEGER4, results(i), 1, 1)
         sum = sum + results(i)
 30   continue
      print *, "The sum is ",sum
      call pvmfexit()
      stop
```

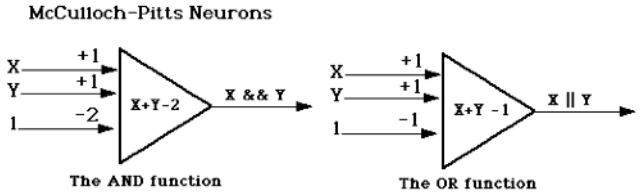


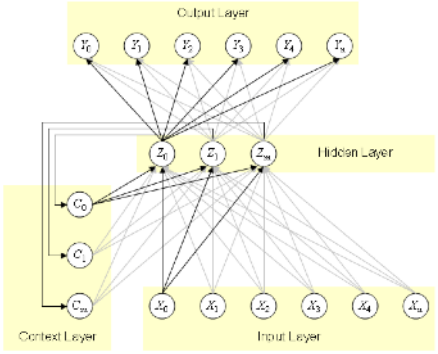
Progression of scientific knowledge

- Neuroscience







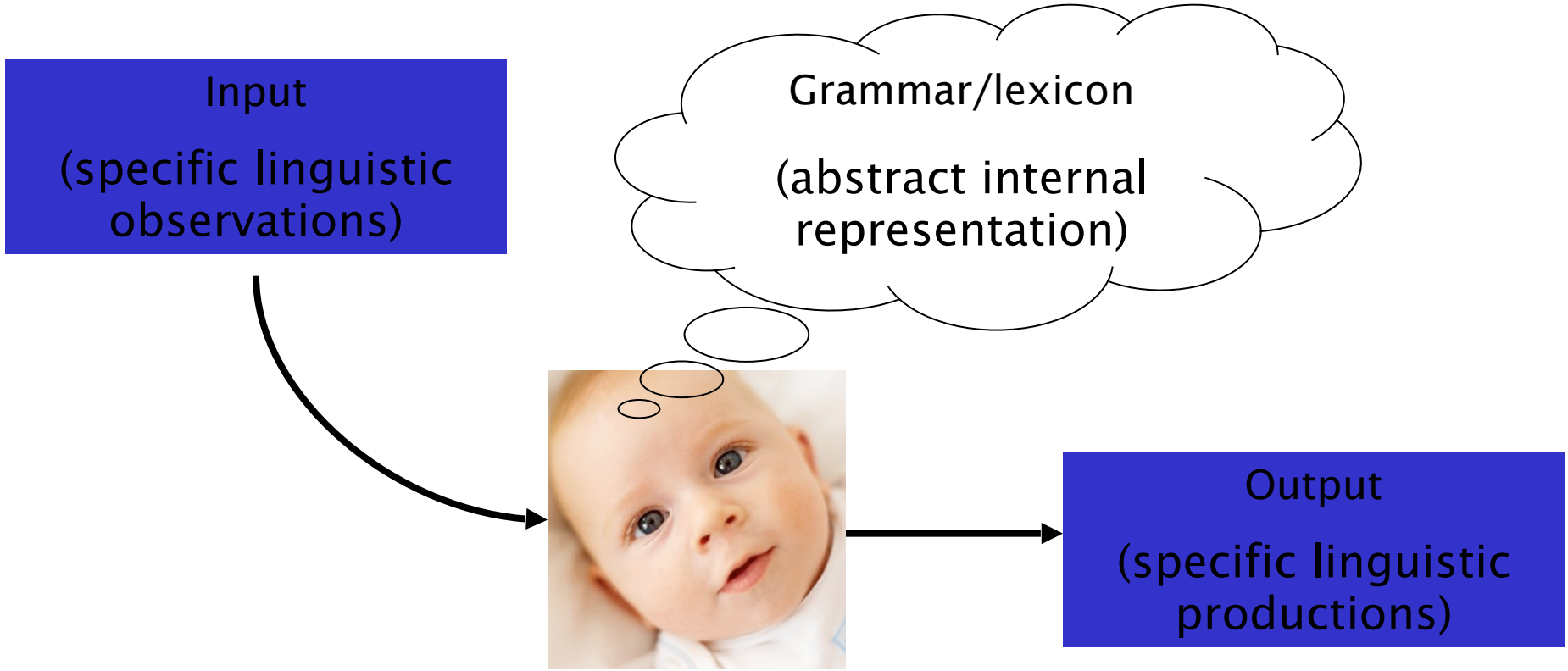


The study of language

- Of all the outward manifestations of the symbolic capacities of human cognition, language is the most discrete and measurable
- Hence we should expect that the forward progression of the science of language will lead the way in the development of formal, precise theories of the mind
- *Computational modeling of human language* is precisely that development

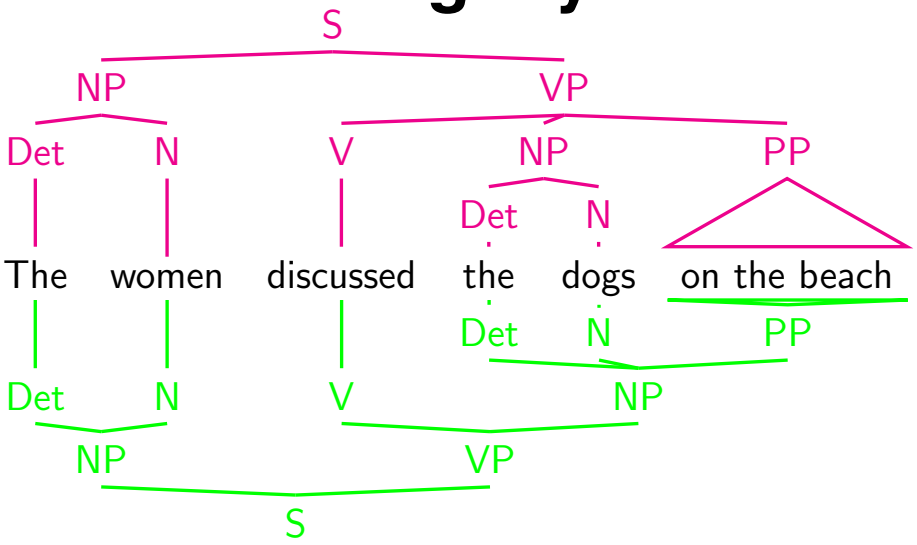
Computational psycholinguistics

- How do *humans* acquire linguistic knowledge, represent it mentally, and deploy it to speak and understand each other?



Challenges for efficient linguistic communication

Ambiguity



Environmental noise

Memory Limitations

Incomplete knowledge of one's interlocutors

If you'd like to learn more...

- 9.19: Computational Psycholinguistics, taught by me!
  - Current syllabus:

    https://canvas.mit.edu/courses/7745
- Other relevant courses:
  - 9.59J/24.905J: Laboratory in Psycholinguistics (Professor Edward Gibson, BCS)
  - 9.66: Computational Cognitive Science (Professor Joshua Tenenbaum, BCS & CSAIL)
  - 24.904: Language Acquisition (Professor Athulya Aravind, Linguistics)

Three vignettes

- Language acquisition
- Language comprehension
- Something funny

Experiments with infant-directed speech

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

SCIENCE • VOL. 274 • 13 DECEMBER 1996

Stimulus from Saffran et al. 1996 type experiment

What do you think the words are?

pigola

golatu

daropi

tudaro

https://www.youtube.com/watch?v=EFIxiflDk_o, 5:34 in

Where are the word boundaries?

badi
pokute
seginekute
pokutefa

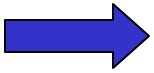
whatsthat
thedoggie
canyousaydoggie
thedoggieishungry

Transition	Probability
ba→di	1
di→*END*	1
po→ku	1
ku→te	1
te→*END*	0.67
se→gi	1
gi→ne	1
ne→ku	1
te→fa	0.33

Word segmentation

- Given a corpus of fluent speech or text (no word boundaries), we want to identify the words.

whatsthat
thedoggie
yeah
wheresthedoggie



whats that
the doggie
yeah
wheres the doggie

- Early language acquisition task for human infants.

[slide courtesy of Sharon Goldwater]

(Goldwater, Griffiths, & Johnson, 2006, 2007, 2009)

The Dirichlet process

- Our generative model for word segmentation assumes data (words!) arise from clusters. Defines
  - A distribution over the number and size of the clusters.
  - A distribution P_0 over the parameters describing the distribution of data in each cluster.
- Clusters = frequencies of different words.
- Cluster parameters = identities of different words.

[slide courtesy of Sharon Goldwater]

The Chinese restaurant process

- In the DP, the number of items in each cluster is defined by the Chinese restaurant process:
  - Restaurant has an infinite number of tables, with infinite seating capacity.
  - The table chosen by the ith customer, z_i, depends on the seating arrangement of the previous $i-1$ customers :

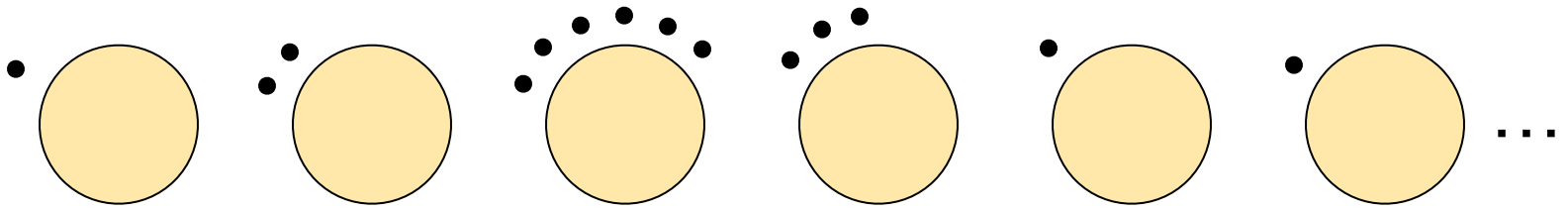
$$P(z_i = k|\mathbf{z}_{-i}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if the } k\text{-th table is occupied} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is the next unoccupied table} \end{cases}$$

- CRP produces a power-law distribution over cluster sizes.

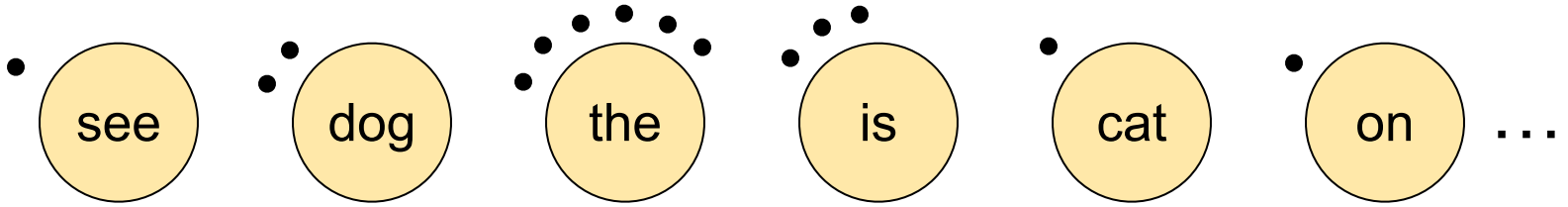
[slide courtesy of Sharon Goldwater]

The two-stage restaurant

1. Assign data points to clusters (tables).



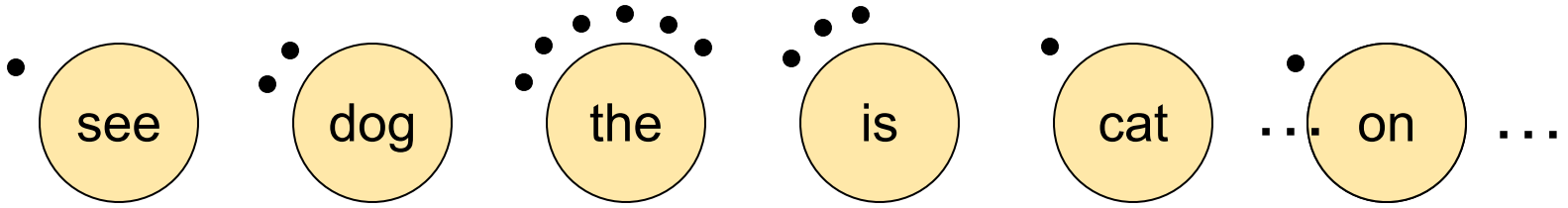
2. Sample labels for tables using P_0.



[slide courtesy of Sharon Goldwater]

Alternative view

- Equivalently, words are generated sequentially using a cache model: previously generated words are more likely to be generated again.



… the dog is on

[slide courtesy of Sharon Goldwater]

Unigram model

- DP model yields the following distribution over words:

$$P(w_i = w \mid \mathbf{w}_{-i}) = \frac{n_w + \alpha P_0(w)}{i - 1 + \alpha}$$

with $P_0(w = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)$ for characters $x_1 \ldots x_m$.

  - P_0 favors shorter lexical items.
  - Words are not independent, but are exchangeable: a unigram model: $P(w_1, w_2, w_3, w_4) = P(w_2, w_4, w_1, w_3)$

- Input corpus contains utterance boundaries. We assume a geometric distribution on utterance lengths.

[slide courtesy of Sharon Goldwater]

Unigram model, in more detail

- How a corpus comes into being…
- First, a probability distribution over corpus length *N*:

$$P(N = n) = (1 - p_{C_\#})^n p_{C_\#}$$

- Next, a probability distribution for the type identity of each new word:

$$P(z_i = k | z_{1\ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

- Finally, a probability distribution over the phonological form of each word type

$$P(form(k) = w_k) = \frac{1}{V^{length(w_k)+1}}$$

Advantages of DP language models

- Solutions are sparse, yet grow as data size grows.
  - Smaller values α_0 of lead to fewer lexical items generated by $P_{0.}$
- Models lexical items separately from frequencies.
  - Different choices for P_0 can infer different kinds of linguistic structure.
- Amenable to standard search procedures (e.g., Gibbs sampling).

[slide courtesy of Sharon Goldwater]

Gibbs sampling

- Compare pairs of hypotheses differing by a single word boundary:

```
whats.that
the.doggie
yeah
wheres.the.doggie
…
```

```
whats.that
the.dog.gie
yeah
wheres.the.doggie
…
```

- Calculate the probabilities of the words that differ, given current analysis of all other words.
- Sample a hypothesis according to the ratio of probabilities.

[slide courtesy of Sharon Goldwater]

Experiments

- Input: same corpus as Brent (1999), Venkataraman (2001).
  - 9790 utterances of transcribed child-directed speech.
  - Example input:

```
youwanttoseethebook
looktheresaboywithhishat
andadoggie
youwanttolookatthis
...
```

- Using different values of α_0, evaluate on a single sample after 20k iterations.

[slide courtesy of Sharon Goldwater]

Example results

```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

[slide courtesy of Sharon Goldwater]

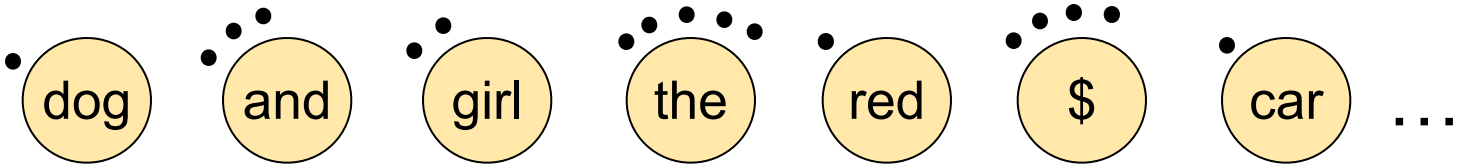
Improving the model

- By incorporating context (using a bigram model), perhaps we can improve segmentation…

[slide courtesy of Sharon Goldwater]

Hierachical Dirichlet process

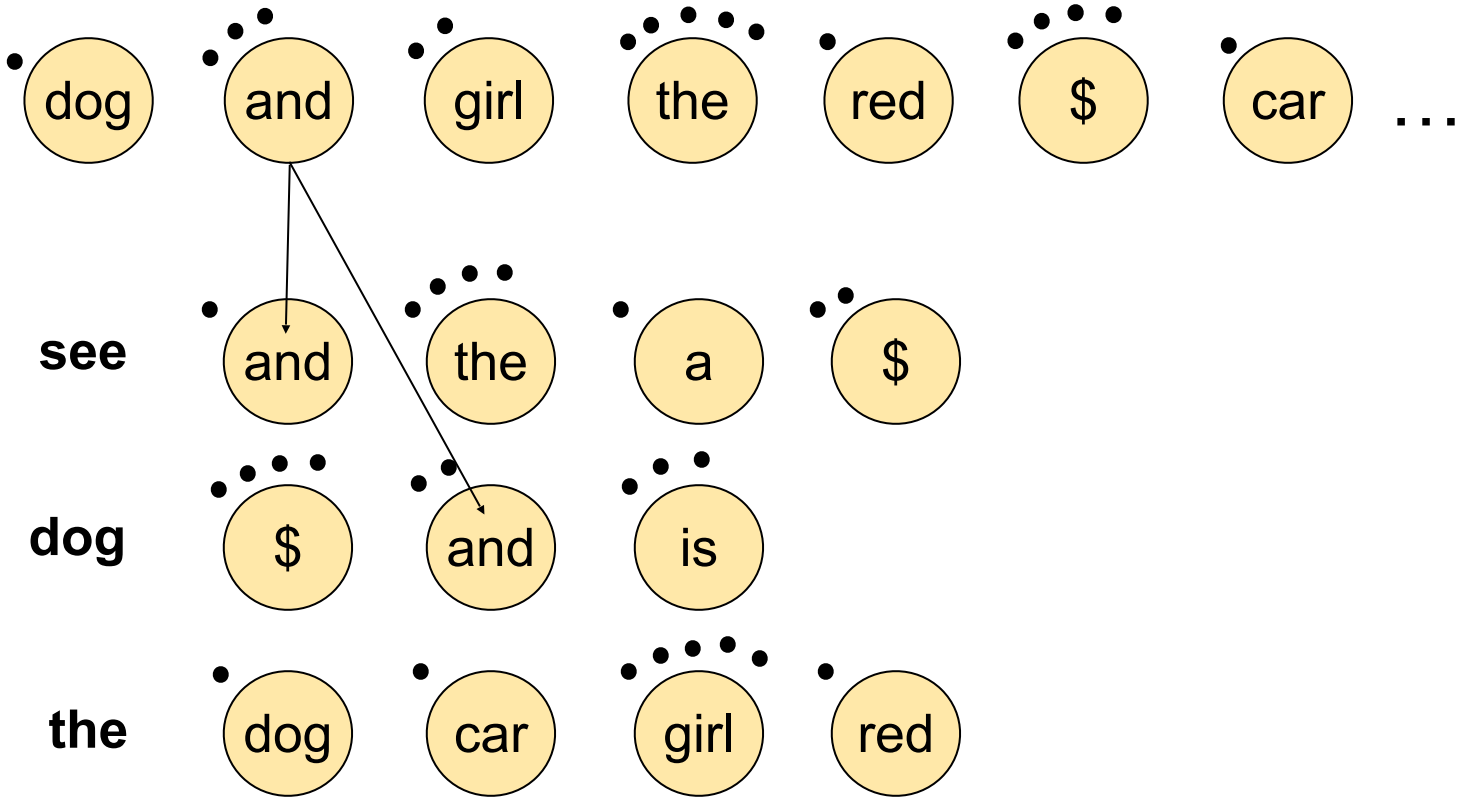
1. Generate G, a distribution over words, using $\mathrm{DP}(\alpha_0, P_0)$.



[slide courtesy of Sharon Goldwater]

Adding bigram dependencies

2. For each word in the data, generate a distribution over the words that follow it, using $\mathrm{DP}(\alpha_1, G)$.



[slide courtesy of Sharon Goldwater]

Example results

```
you want to see the book
look theres a boy with his hat
and a doggie
you want to lookat this
lookat this
have a drink
okay now
whats this
whats that
whatis it
look canyou take it out
...
```

[slide courtesy of Sharon Goldwater]

Quantitative evaluation

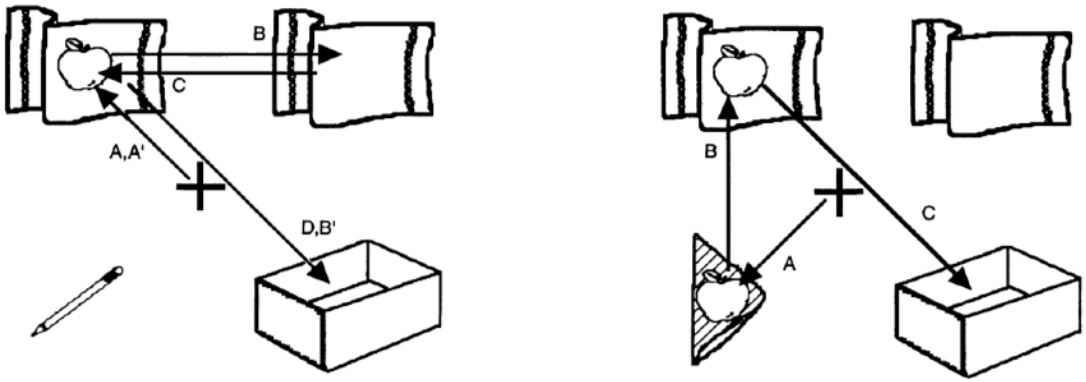
	Boundaries		Word tokens		Lexicon	
	Prec	Rec	Prec	Rec	Rec	Prec
DP (unigram)	**92.4**	62.2	61.9	47.6	57.0	**57.5**
Venk. (bigram)	81.7	82.5	68.1	68.6	54.5	57.0
HDP (bigram)	89.9	**83.8**	**75.7**	**72.1**	**63.1**	50.3

- With appropriate hyperparameter choices for α & β,
  - Boundary precision nearly as good as unigram, recall much better.
  - F-score (avg. of prec, rec) on all three measures outperforms all previously published models.
- More generally, the model provides insight into how word learning can fall out of a simple generative model of language

[slide courtesy of Sharon Goldwater]

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
- **Question:** how do we reconcile these information sources?



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

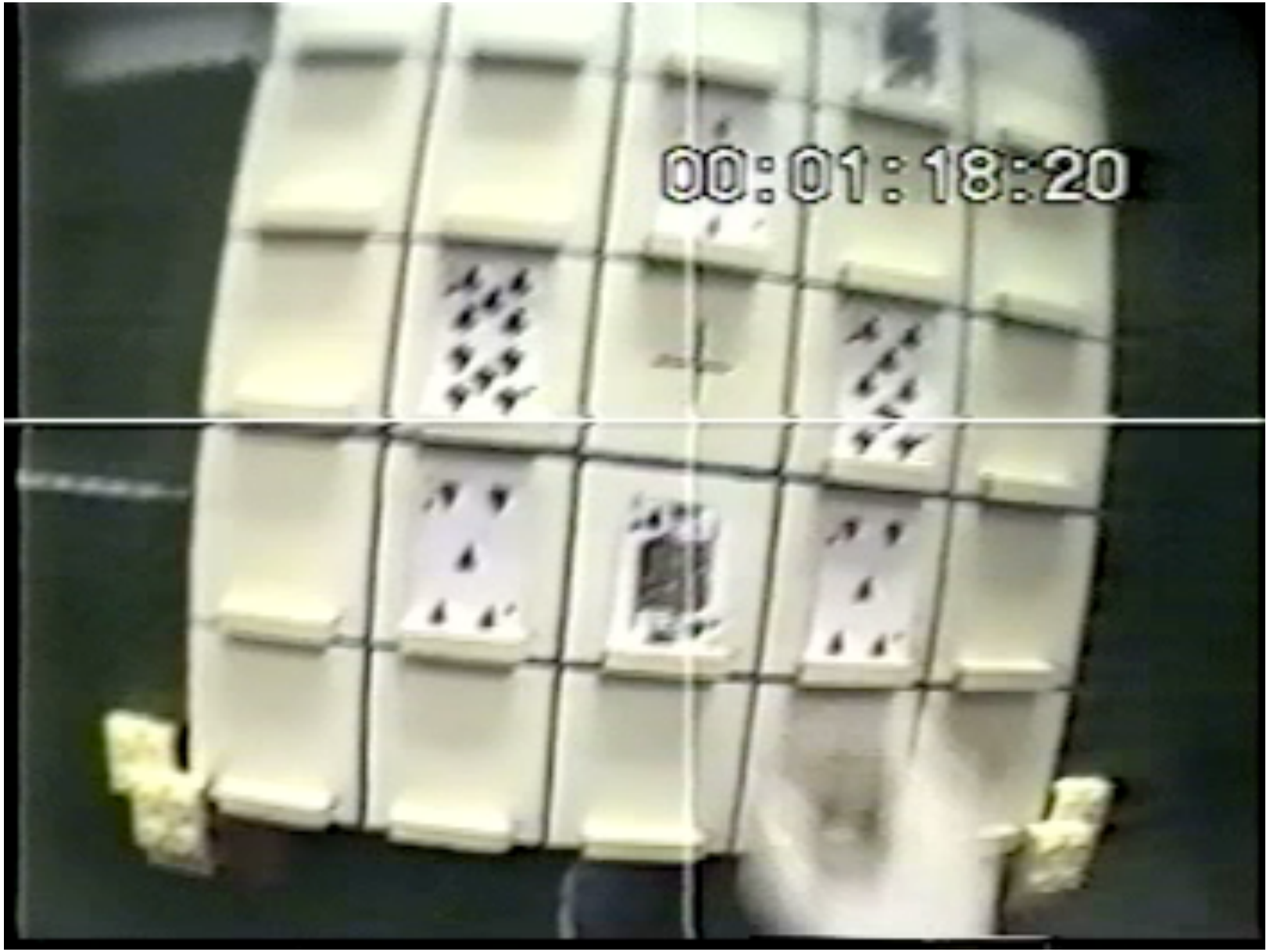
Measuring human incremental processing state

- Eye movements in the visual world
- Word-by-word reading times
  - Self-paced reading
  - Eye movements during natural reading
- Recordings of brain activity
  - Electrophysiological (EEG/ERP)
  - Magneto-encephalography (MEG)
  - functional Magnetic Resonance Imaging (fMRI)
  - Electrocorticography (ECoG)

Behavioral

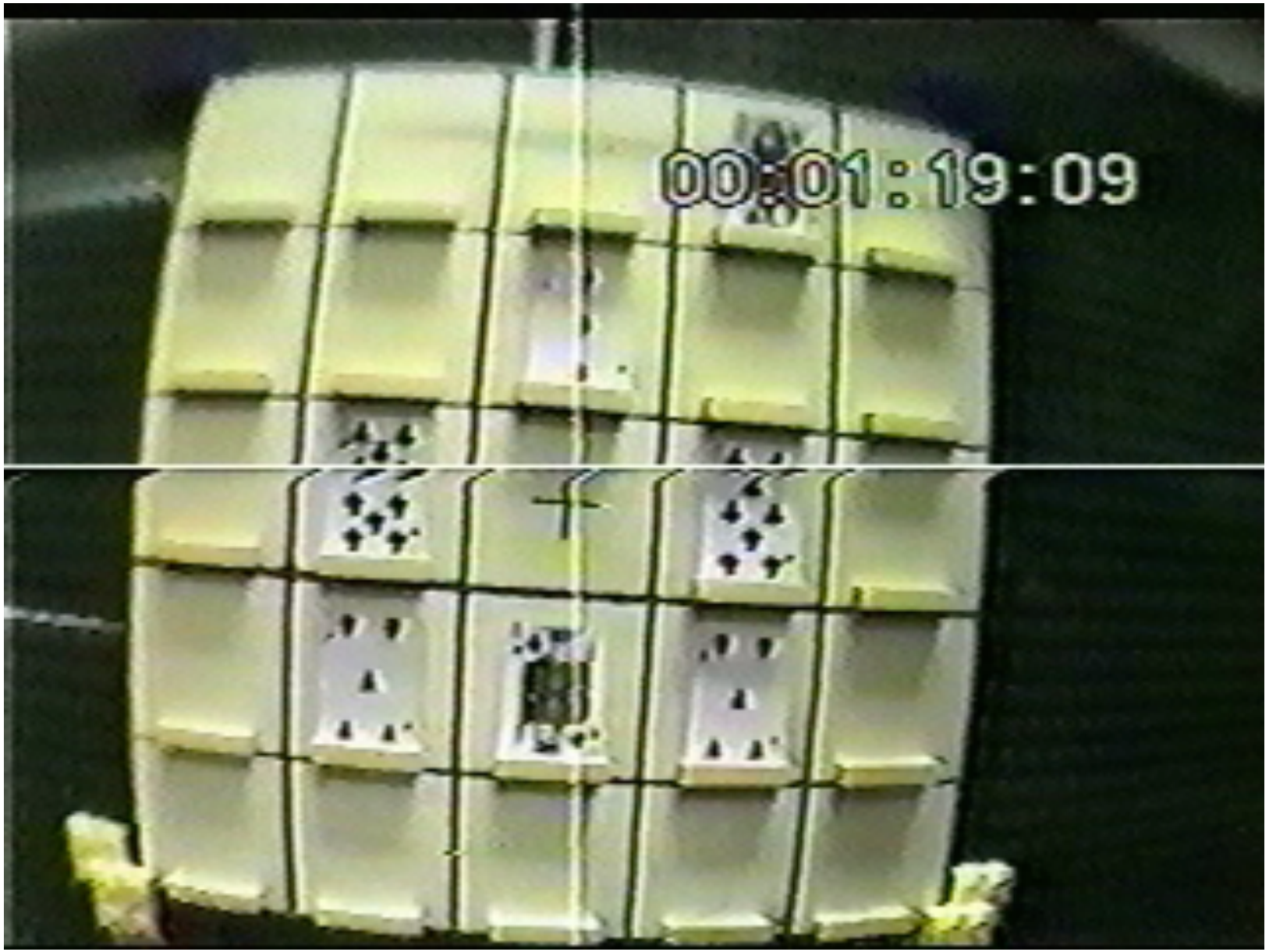
Neural

Eye movements in the visual world



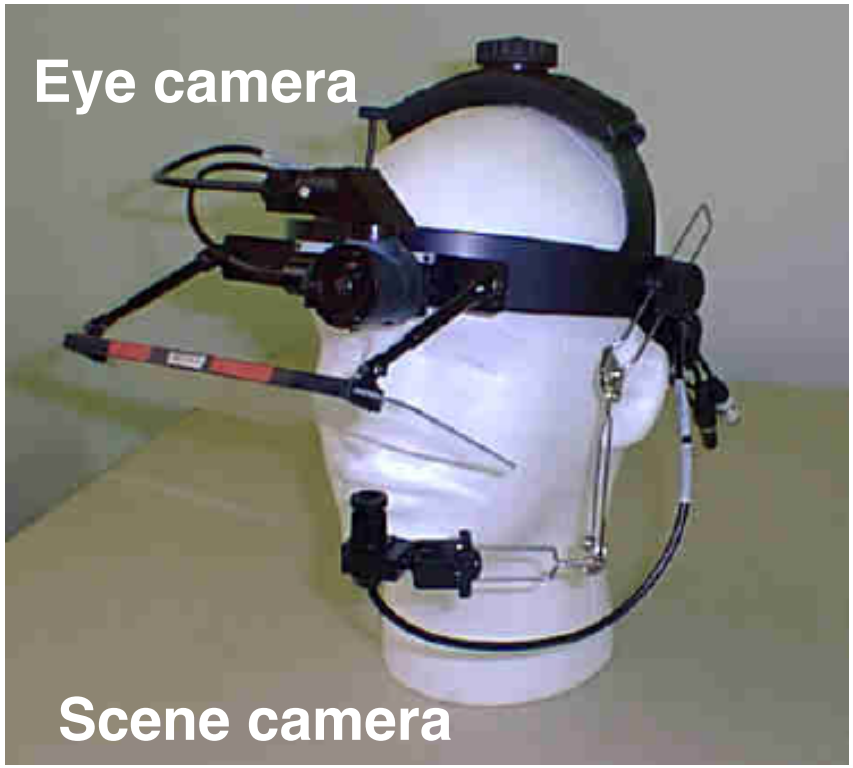
(Video courtesy of Mike Tanenhaus)

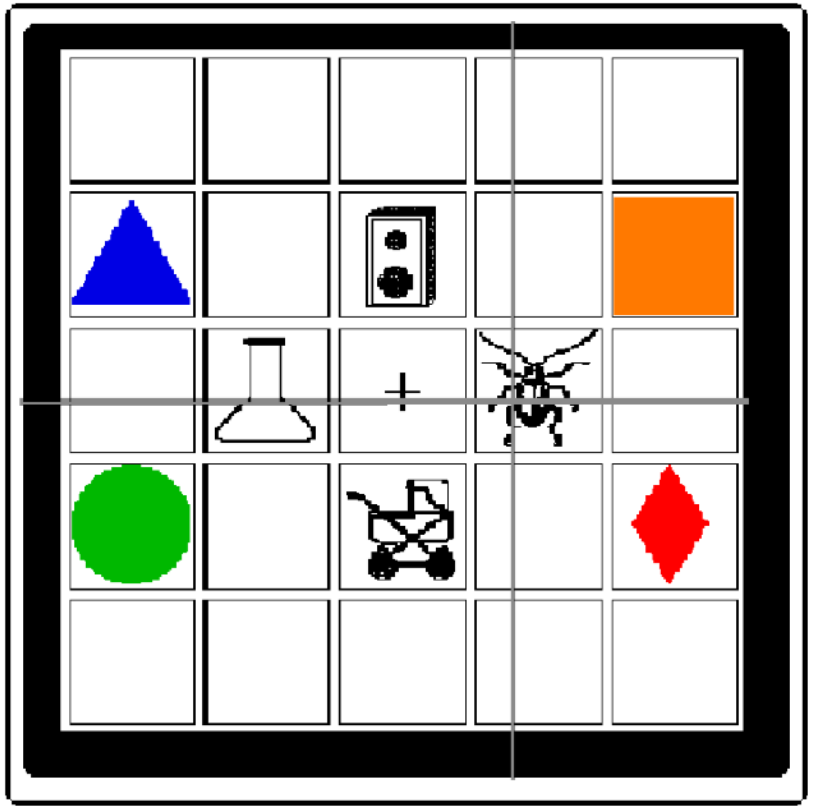
Eye movements in the visual world



(Video courtesy of Mike Tanenhaus)

A visual world experiment





Instruction to experimental participant:

"Pick up the beaker"

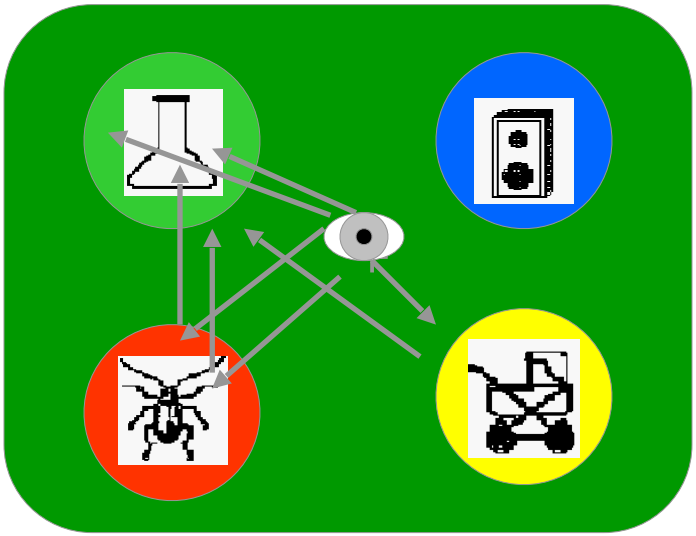
(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)

Data from human eye movements

"Look at the cross."

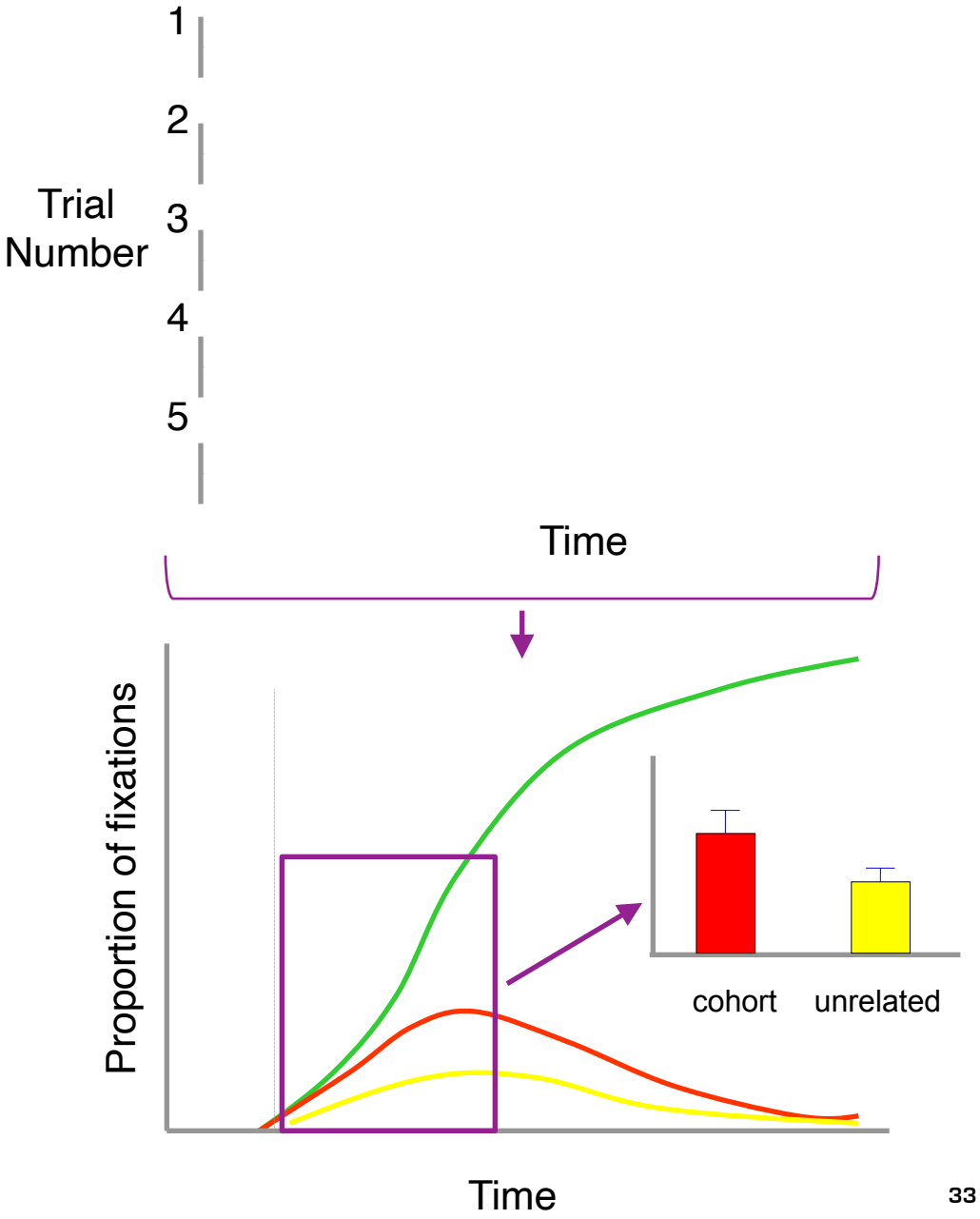
"Pick up the beaker."



Target = beaker

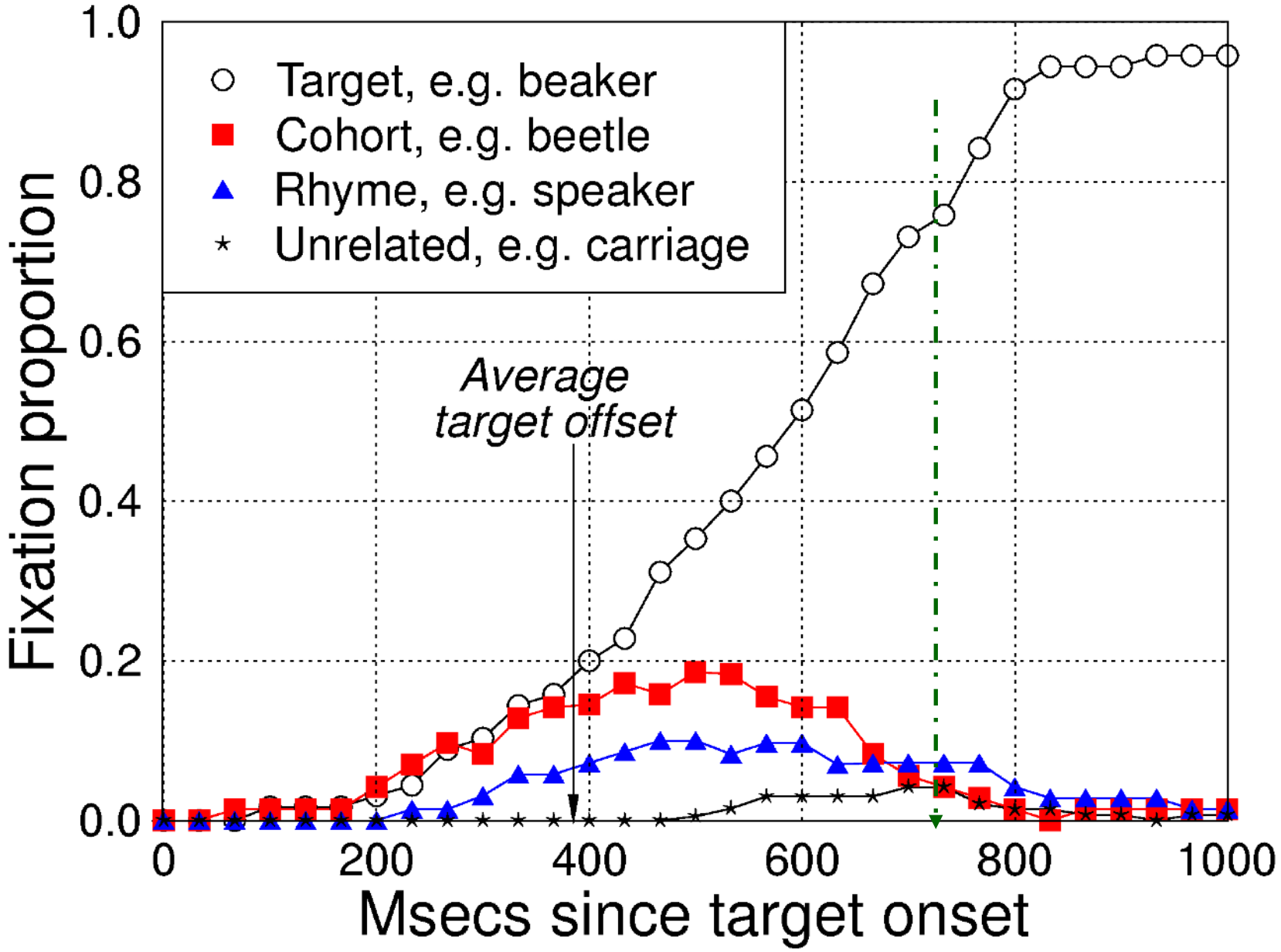
Cohort = beetle

Unrelated = carriage



(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)



(Slide courtesy of Mike Tanenhaus)

Generalizing incremental disambiguation

- Uncertainty in predictions about upcoming material

The old man stopped and stared at the *statue? dog? view? woman?*

The squirrel stored some nuts in the *tree*

- This is uncertainty about *what has not yet been said*
- Reading-time (Ehrlich & Rayner, 1981) and EEG (Kutas & Hillyard, 1980, 1984) evidence shows this affects processing rapidly
- A good model should account for expectations about how this uncertainty will be resolved

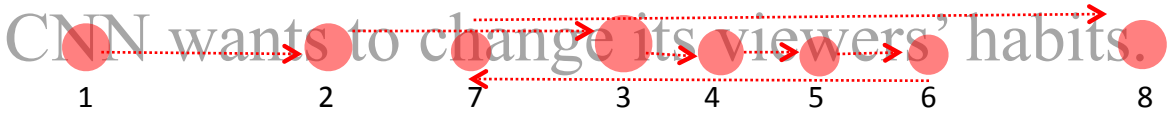
*(the **Cloze procedure**; Taylor, 1953)*

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?



(slide courtesy Yevgeni Berzak)

How do people read?

Fixations

(slide courtesy Yevgeni Berzak)

How do people read?

Saccades

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

Eye movements in reading

(movie by Piers Cornelissen)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

What do you see during a fixation?

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

Perceptual span

What do you see during a fixation?

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

What do you see during a saccade?

(slide courtesy Yevgeni Berzak)

How do people read?

What do you see during a saccade?

Nothing

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

CNN wants to change its viewers' habits.

Forward
Saccade

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

How do people read?

(slide courtesy Yevgeni Berzak)

Eye movement measures

- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

(slide courtesy Yevgeni Berzak)

Eye movement measures

- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

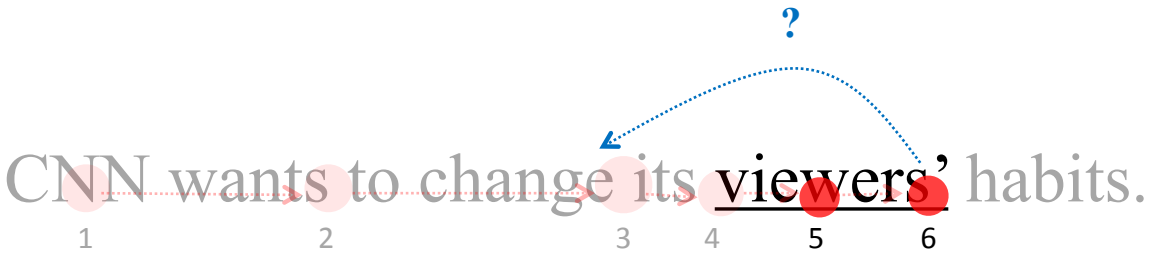
(slide courtesy Yevgeni Berzak)

Eye movement measures

- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

(slide courtesy Yevgeni Berzak)

Eye movement measures



- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

(slide courtesy Yevgeni Berzak)

Eye movement measures

- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

(slide courtesy Yevgeni Berzak)

Eye movement measures

- Skips (also skip rate / fixation probability)
- First fixation duration
- First pass duration (or Gaze duration)
- First pass regression rate
- Go-past duration
- Total fixation duration

(slide courtesy Yevgeni Berzak)

Linguistic Expectations

- Linguistic expectations can be studied with eye tracking for reading.
- Reading times (across different eye movement measures) reflect how contextual predictability affects linguistic processing.

(slide courtesy Yevgeni Berzak)

Rayner & Well 1996

The hikers slowly climbed up the _____

Equal word length & frequency

mountain (95%)

hillside (3%)

(slide courtesy Yevgeni Berzak)

Rayner & Well 1996

The hikers slowly climbed up the mountain to get a better view.

The hikers slowly climbed up the hillside to get a better view.

		Fixation Time		
Constraint	Fixation Probability	First Fixation	Gaze Duration	Total Time
High	0.78	239	261	294
Low	0.90	250	281	360

(slide courtesy Yevgeni Berzak)

Staub 2011

While the professor lectured the students walked across the quad.

(slide courtesy Yevgeni Berzak)

Staub 2011

???

While the professor lectured the students walked across the quad.

(slide courtesy Yevgeni Berzak)

Staub 2011

???

[While the professor [lectured the students]] walked across the quad.

Subj V Obj

(slide courtesy Yevgeni Berzak)

Staub 2011

???

[While the professor [lectured the students]] walked across the quad.

Subj V Obj

[While the professor lectured] [the students walked across the quad.]

Subj V Subj

(slide courtesy Yevgeni Berzak)

Staub 2011

??? <u>walked</u>

[While the professor [lectured the students]] <u>walked</u> across the quad.

Subj V Obj

[While the professor lectured] [the students <u>walked</u> across the quad.]

Subj V Subj

(slide courtesy Yevgeni Berzak)

Staub 2011

???
[While the professor [lectured the students]] walked across the quad.
Subj V Obj

While the professor lectured, the students walked across the quad.

(slide courtesy Yevgeni Berzak)

Staub 2011

??? <u>walked</u>

[While the professor [lectured the students]] <u>walked</u> across the quad.

Subj V Obj

[While the professor lectured,] [the students <u>walked</u> across the quad.]

Subj V Subj

(slide courtesy Yevgeni Berzak)

Staub 2011

While the professor lectured the students walked across the quad.
(ambiguous)

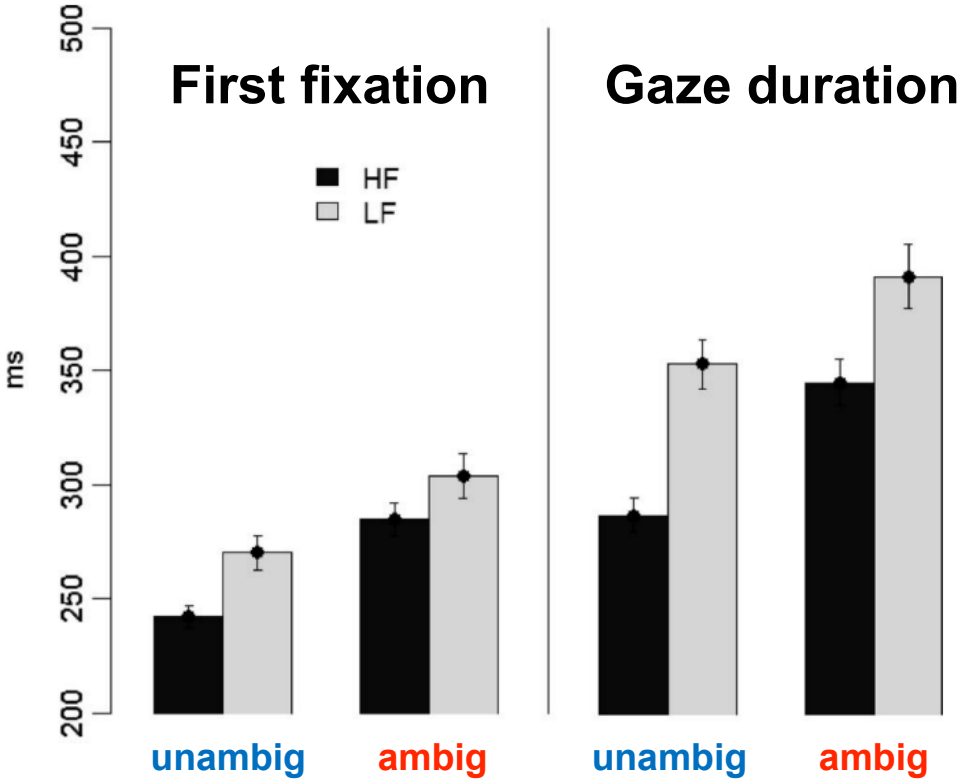
While the professor lectured, the students walked across the quad.
(unambiguous)

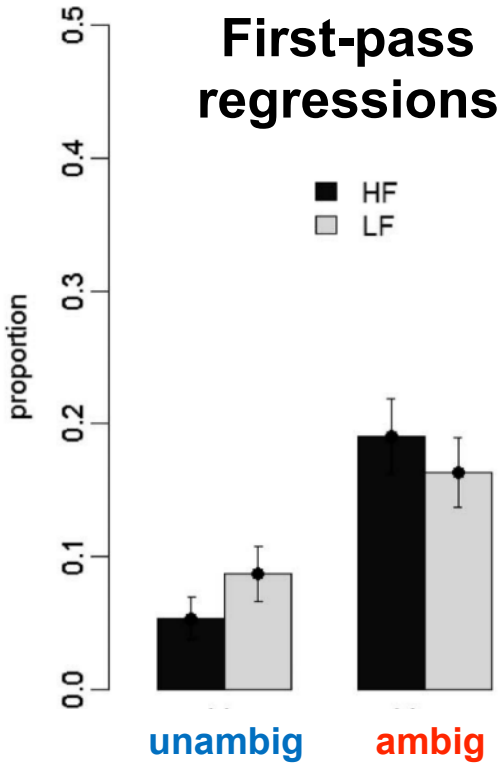
High Frequency

ambled

ambled

Low Frequency





(slide courtesy Yevgeni Berzak)

Psycholinguistic methodology (2)

- A lower-tech method: **self-paced reading**
- Reveal each consecutive word with a button press

```
While the clouds crackled, above the glider soared ----------
```

- Readers aren't allowed to backtrack
- We measure time between button presses and use it as a proxy for incremental processing difficulty

Psycholinguistic methodology (3)

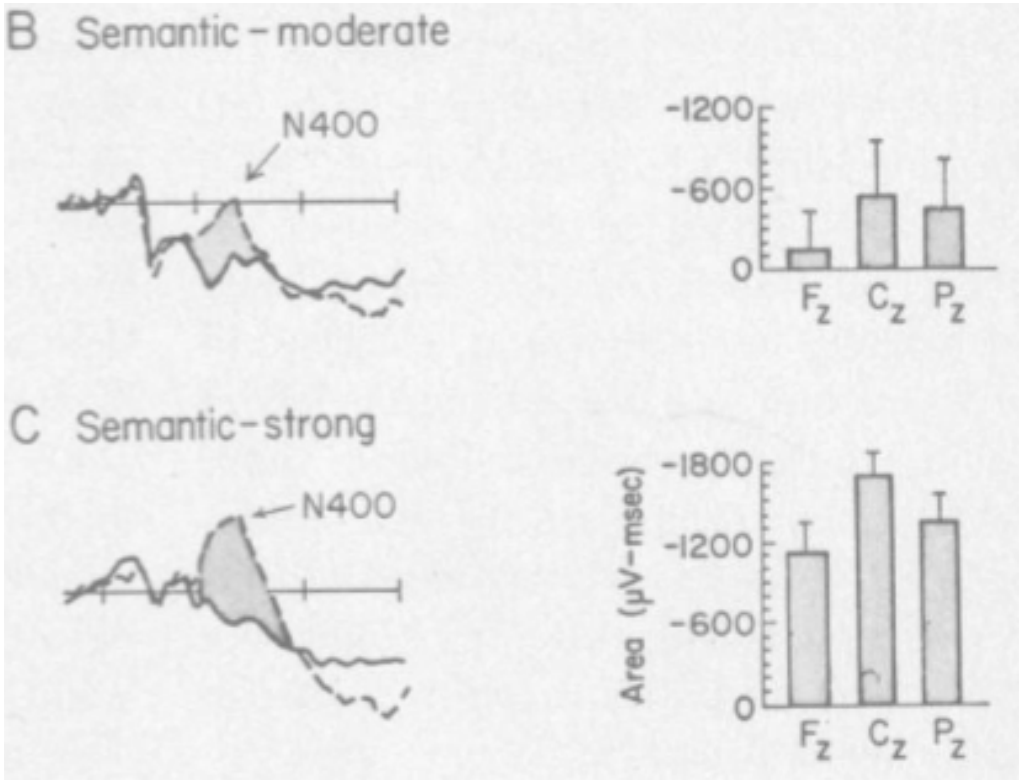
- Also: *neurolinguistic* experimentation more and more widely used to study language comprehension
  - methods vary in temporal and spatial resolution
  - people are more passive in these experiments: sit back and listen to/read a sentence, word by word
  - strictly speaking *not* behavioral measures
  - the question of “what is difficult” becomes a little less straightforward

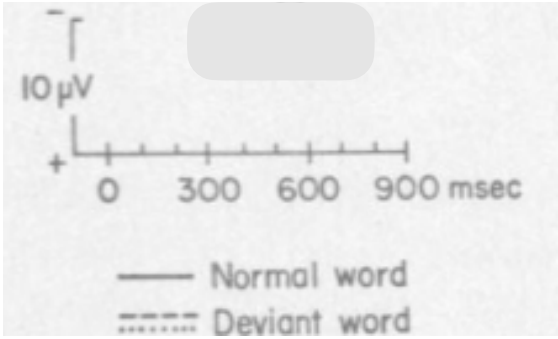
Electrophysiological responses

Rapid Serial Visual Presentation

The N400 in language comprehension

- Differing degrees of semantic congruity:
  - He took a sip from the *drink*. (normal)
  - He took a sip from the *waterfall*. (moderate incongruity)
  - He took a sip from the *transmitter*. (strong incongruity)



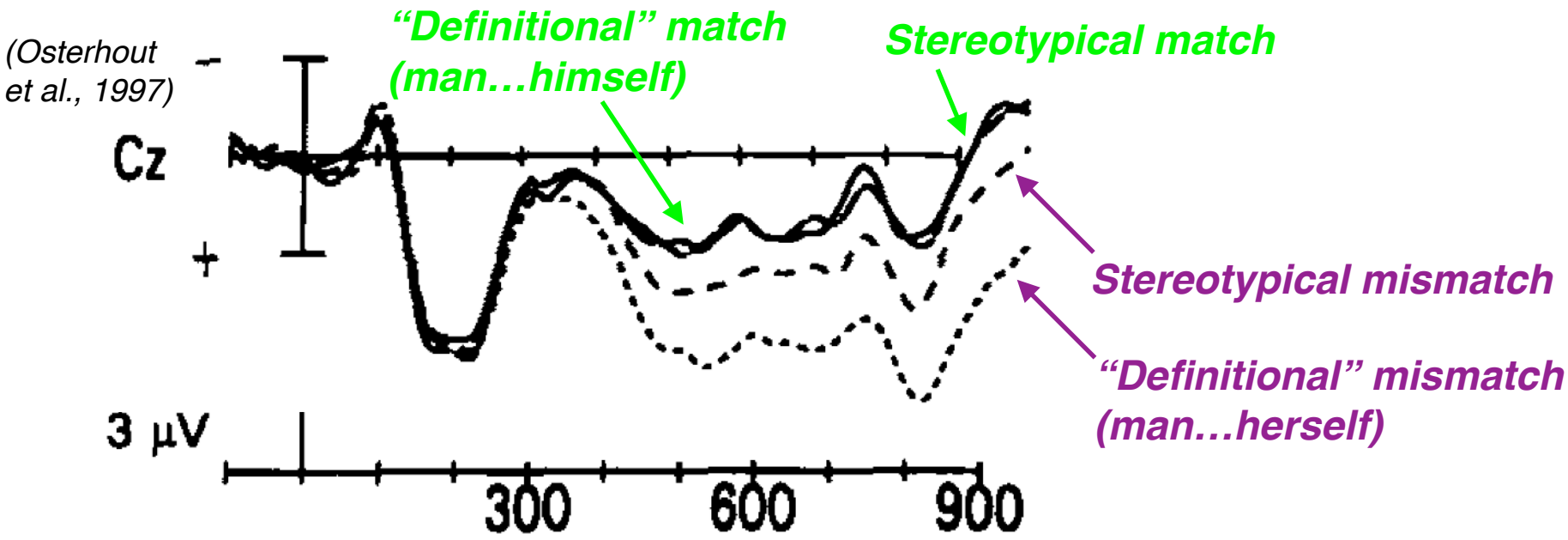


(Kutas & Hillyard, 1980, 1984)

The P600 ERP component in language comprehension

- Mismatches to lexically specified (*definitional**) semantic properties induce measurable expectation violations

  The man prepared herself for the interview.



- Mismatches to *stereotypical* semantic properties induce similar violations

  The nurse prepared himself for the operation.

(Osterhout et al., 1997; see also reading time studies by Sturt, 2003; Duffy & Keir, 2004, inter alia)

fMRI recordings during comprehension

- MRI measures changes in brain associated with blood flow
- Slow, but good *spatial resolution* for which parts of the brain are active in processing

Sentences condition

A	RUSTY	LOCK	WAS	FOUND	IN	THE	DRAWER	+	LOCK/ PEAR	+

Nonwords condition

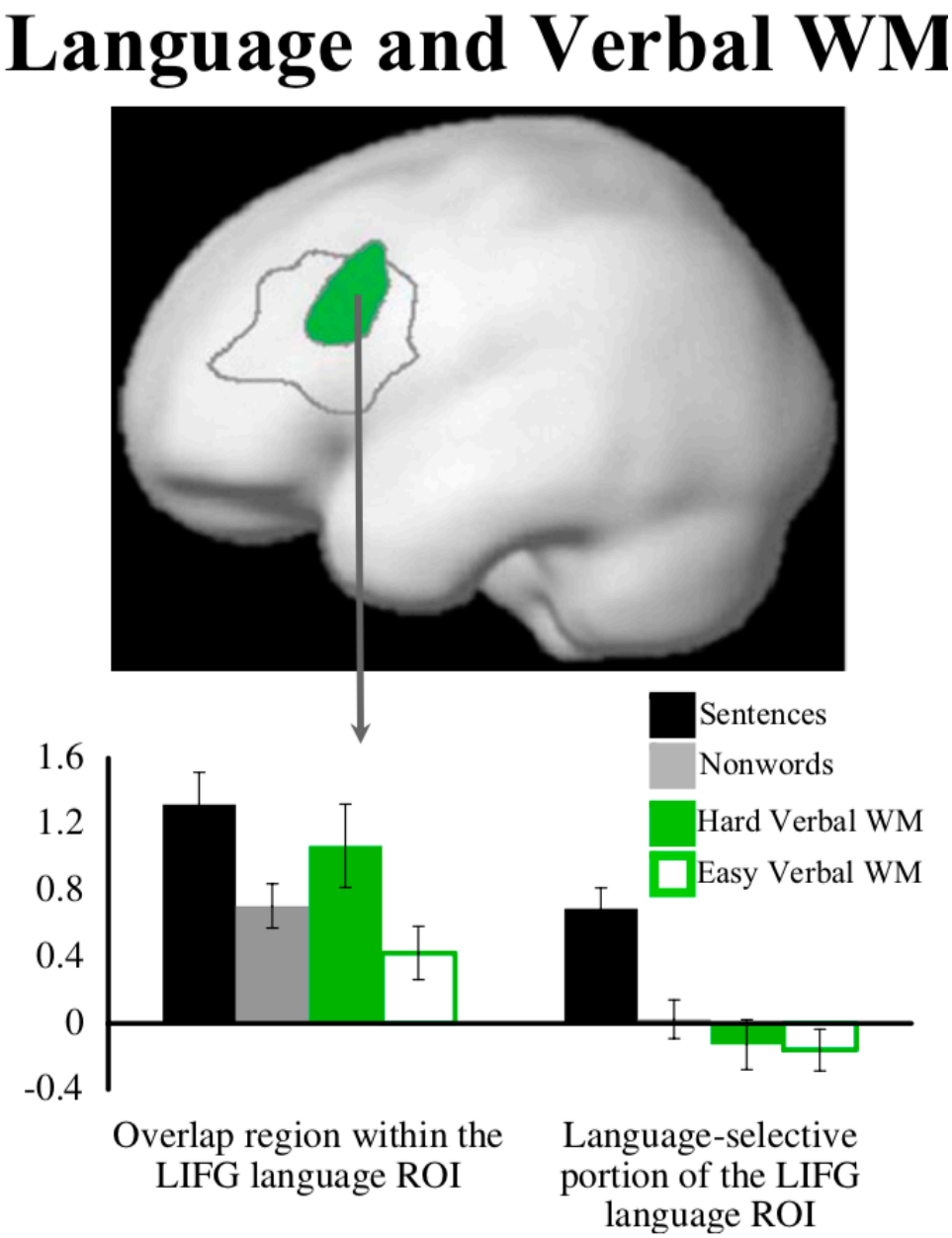
DAP	DRELLO	SMOP	UB	PLID	KAV	CRE	REPLODE	+	DRELLO/ NUZZ	+

Expt 3 (Verbal WM): Sample trial (hard condition)

+	three six	two four	one eight	five three	Response: 36241853 36248153	Feedback: ✓/✗	+

(Fedorenko et al., 2011)

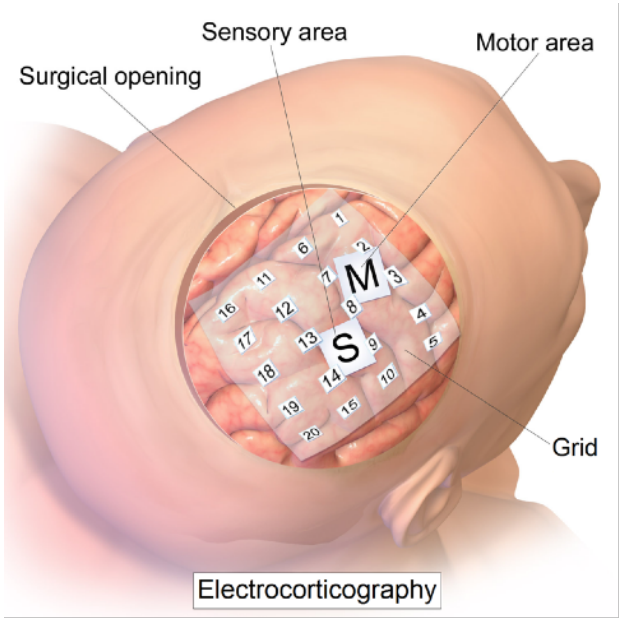
Functional brain specificity for language



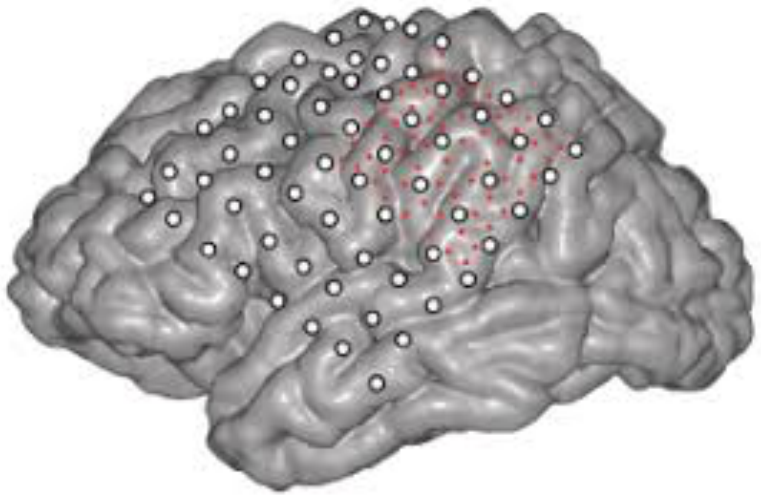
(Fedorenko et al., 2011)

Electrocorticography

- Pre-surgical epilepsy patients get electrode arrays directly implanted on the surface of the cortex



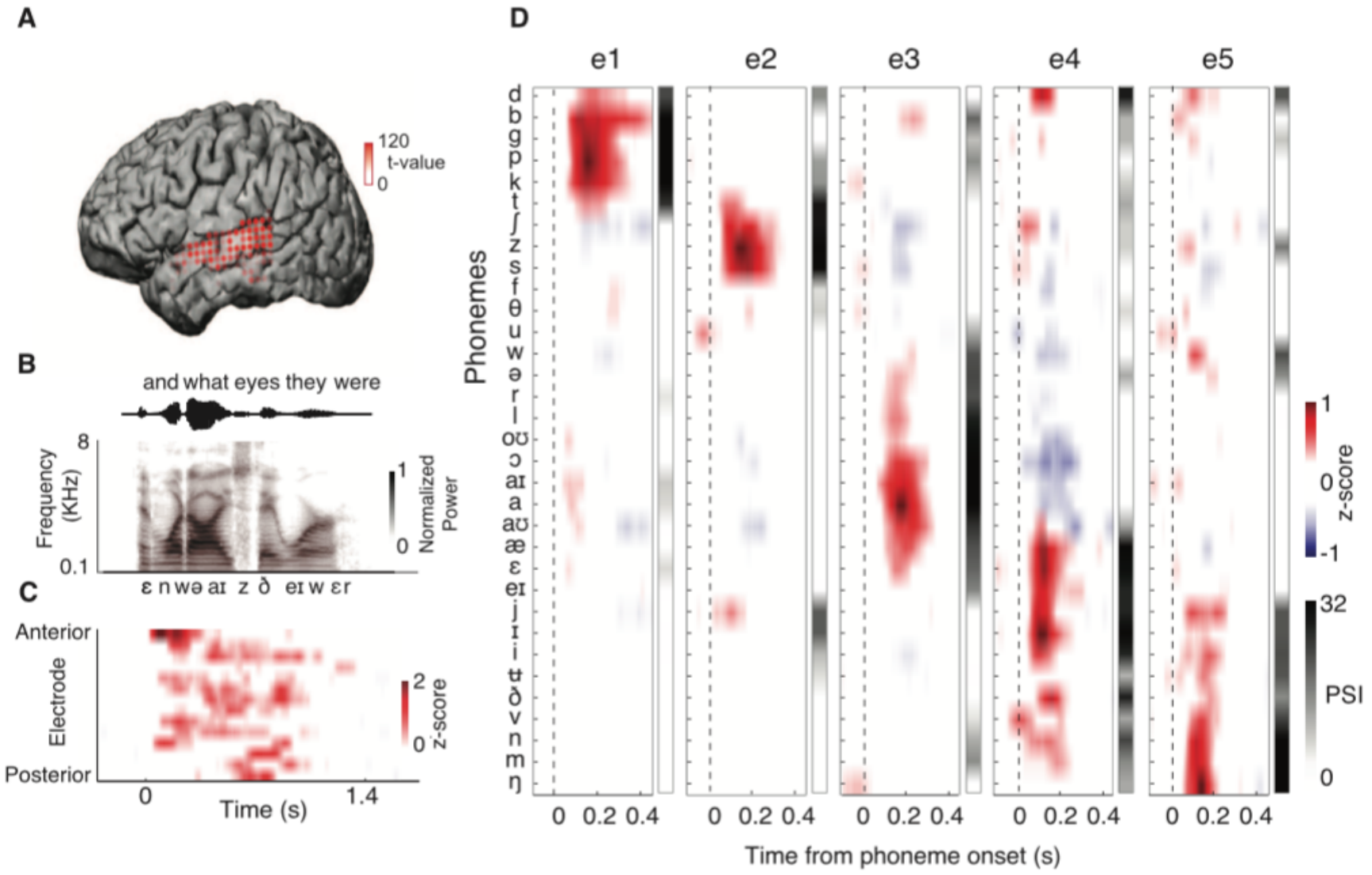
https://commons.wikimedia.org/wiki/File:Intracranial_electrode_grid_for_electrocorticography.png



http://med.stanford.edu/neurosurgery/research/NPTL/research2/_jcr_content/main/panel_builder/panel_0/text_image.img.620.high.png

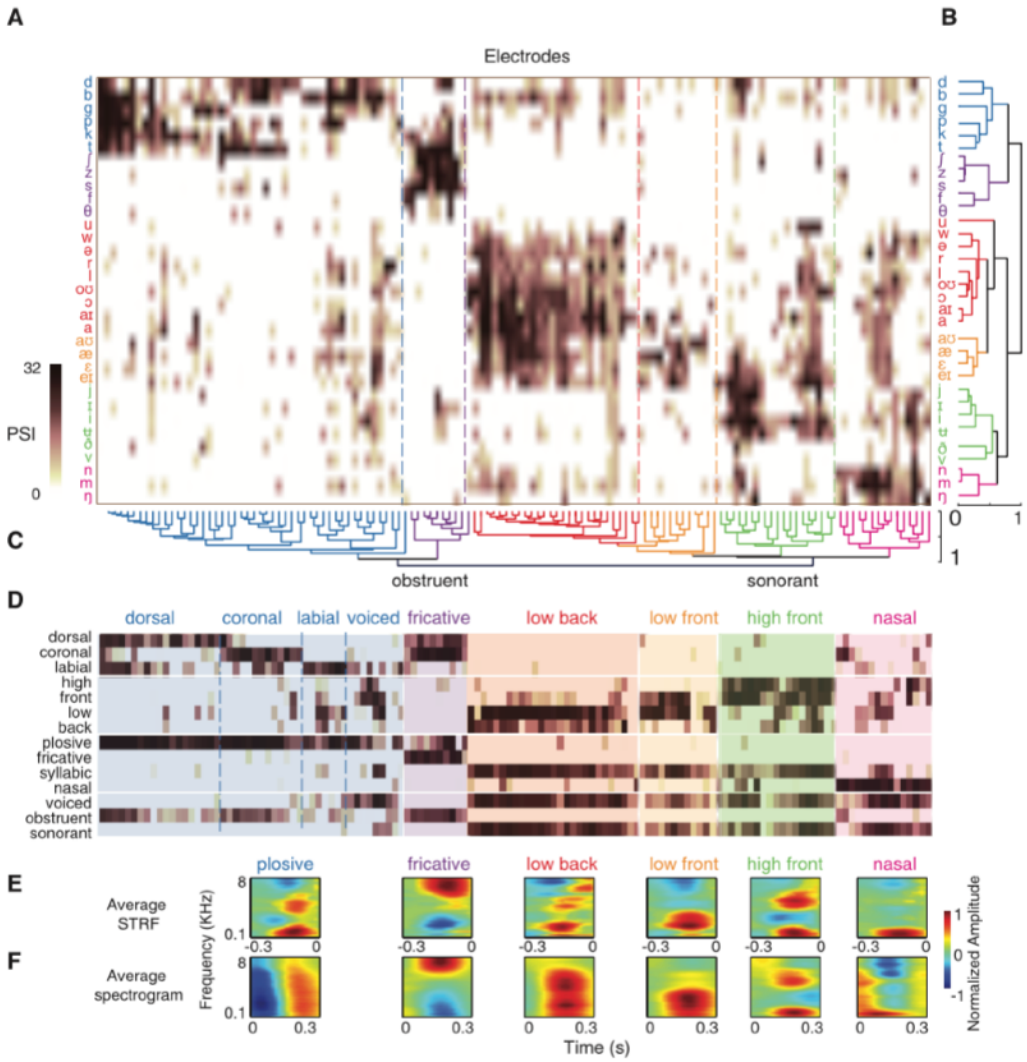
- During pre-surgical monitoring many patients generously donate their energy & attention for experiments

Neural phonemic representations



(Mesgarani et al., 2014, Science)

Neural consonant representations



(Mesgarani et al., 2014, Science)

Scientific opportunity:

Comprehensive theory to account for patterns of human language use & representation

Engineering opportunity:

Better prediction of human language understanding, and more human-like AI language-using agents

Language comprehension

- Desiderata for a satisfactory theory of sentence comprehension:
  - Robustness to arbitrary input
  - Accurate disambiguation
  - Inference on the basis of incomplete input *(incrementality)*
  - Processing difficulty is differential and localized

Not all sentences are equally easy to understand, nor are all parts of a given sentence are equally easy to understand

- I primarily focus on the relationship of the last of these desiderata to the rest

Quantifying probabilistic online processing difficulty

- Let a word's difficulty be its *surprisal* given its context:

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

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

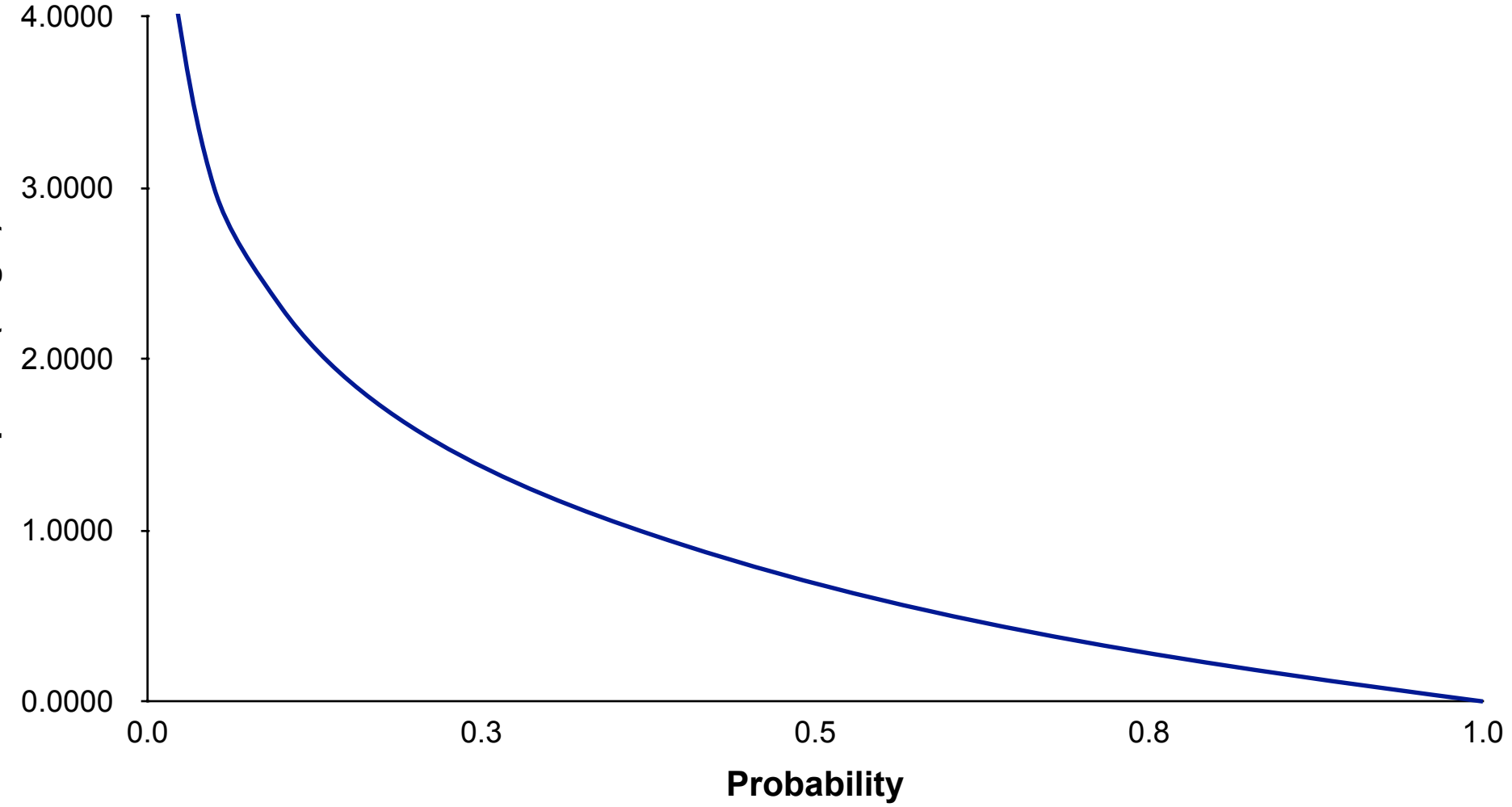
- Captures the *expectation* intuition: the more we expect an event, the easier it is to process
  - Brains are ***prediction*** engines!

    my brother came inside to… *chat? wash? get warm?*

    the children went outside to… *play*

  - Predictable words are read faster (Ehrlich & Rayner, 1981) and have distinctive EEG responses (Kutas & Hillyard 1980)
- Combine with probabilistic grammars to give *grammatical expectations* *(Hale, 2001, NAACL; Levy, 2008, Cognition)*

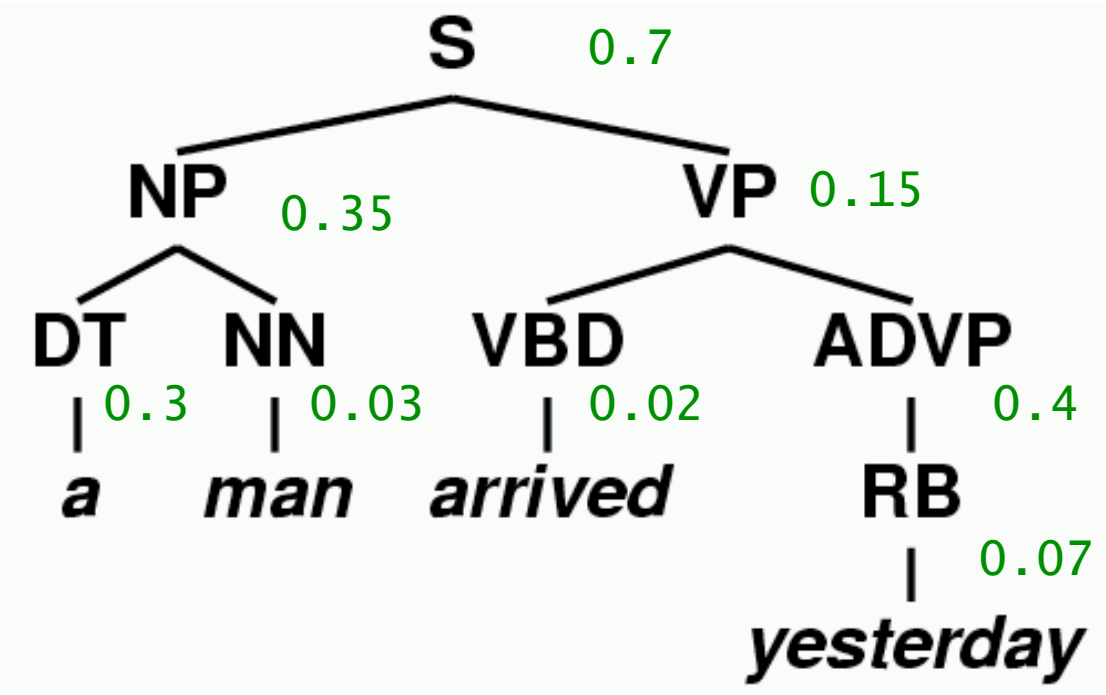
The surprisal graph



a man arrived yesterday

0.3 S → S CC S
0.7 S → NP VP
0.35 NP → DT NN

0.15 VP → VBD ADVP
0.4 ADVP → RB
...

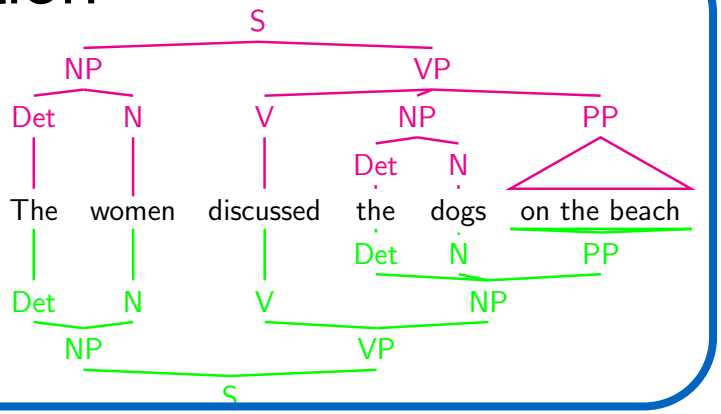


Total probability: 0.7*0.35*0.15*0.3*0.03*0.02*0.4*0.07= 1.85×10^{-7}

Algorithms by Lafferty and Jelinek (1992), Stolcke (1995) give us $P(w_i|context)$ from a PCFG

Three problems unified by surprisal

- Ambiguity resolution



- Prediction

my brother came inside to…

the children went outside to… *play*

- Syntactic complexity

This is the malt that the rat that the cat that the dog chased killed ate.

Garden-pathing and surprisal

- Here's a ***local syntactic ambiguity:***

When the dog scratched the vet and his new assistant removed the muzzle.

difficulty here
(68ms/char)

- Compare with:

When the dog scratched, the vet and his new assistant removed the muzzle.

When the dog scratched its owner the vet and his new assistant removed the muzzle.

easier
(50ms/char)

(Frazier & Rayner, 1982)

A small PCFG for this sentence type

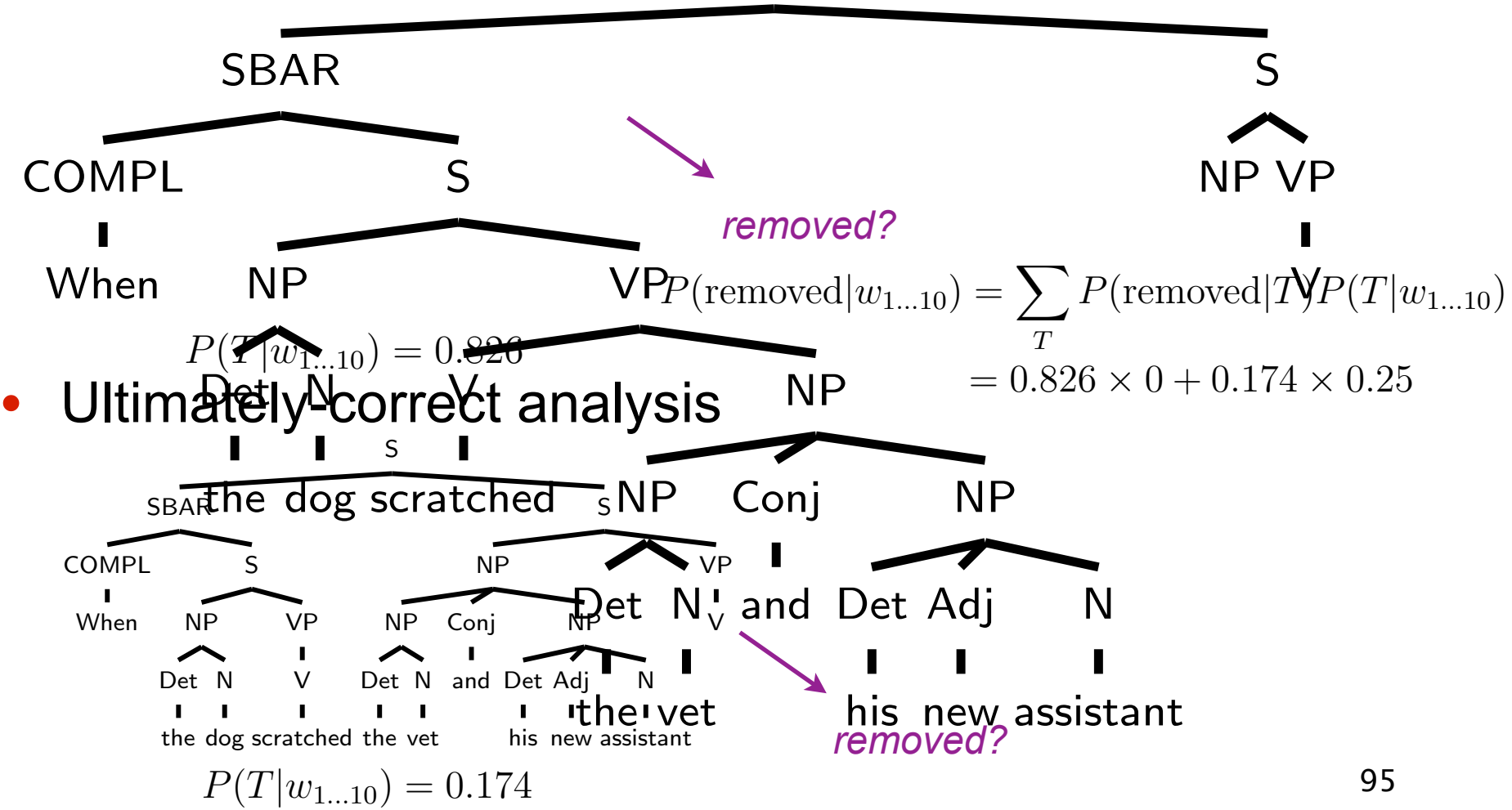
S	→ SBAR S	0.3	
S	→ NP VP	0.7	
SBAR	→ COMPL S	0.3	
SBAR	→ COMPL S COMMA	0.7	
COMPL	→ When	1	
NP	→ Det N	0.6	
NP	→ Det Adj N	0.2	
NP	→ NP Conj NP	0.2	
Conj	→ and	1	
Det	→ the	0.8	
Det	→ its	0.1	
Det	→ his	0.1	
N	→ dog	0.2	
N	→ vet	0.2	
N	→ assistant	0.2	
N	→ muzzle	0.2	
N	→ owner	0.2	
Adj	→ new	1	
VP	→ V NP	0.5	
VP	→ V	0.5	
V	→ scratched	0.25	
V	→ removed	0.25	
V	→ arrived	0.5	
COMMA	→ ,	1	

(analysis in Levy, 2011)

Two incremental trees

- "Garden-path" analysis:

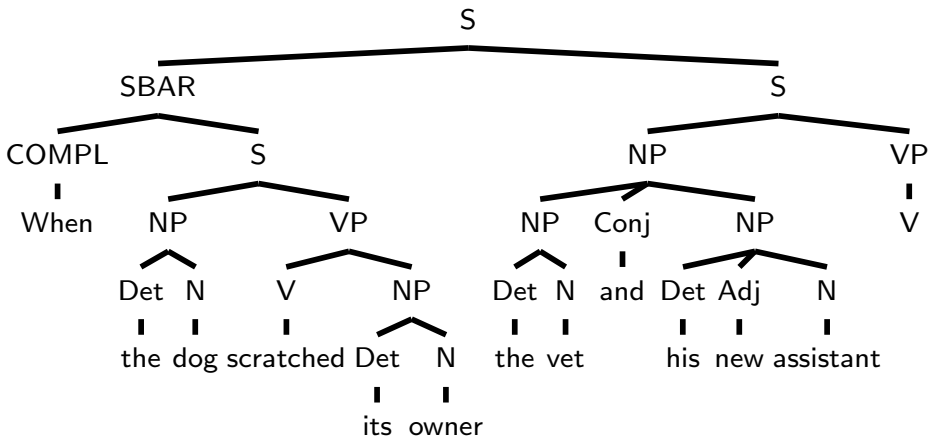
Disambiguating word probability marginalizes over incremental trees:



- Ultimately-correct analysis

Preceding context can disambiguate

- *"its owner"* takes up the object slot of *scratched*



Condition	Surprisal at Resolution
NP absent	4.2
NP present	2

Sensitivity to verb argument structure

- A superficially similar example:

When the dog arrived the vet and his new assistant removed the muzzle.

But harder here!

Easier here

(c.f. When the dog scratched the vet and his new assistant removed the muzzle.)

(Staub, 2007)

Modeling argument-structure sensitivity

S	→ SBAR S	0.3	Conj	→ and	1	Adj	→ new	1
S	→ NP VP	0.7	Det	→ the	0.8	VP	→ V NP	0.5
SBAR	→ COMPL S	0.3	Det	→ its	0.1	VP	→ V	0.5
SBAR	→ COMPL S COMMA	0.7	Det	→ his	0.1	V	→ scratched	0.25
COMPL	→ When	1	N	→ dog	0.2	V	→ removed	0.25
NP	→ Det N	0.6	N	→ vet	0.2	V	→ arrived	0.5
NP	→ Det Adj N	0.2	N	→ assistant	0.2	COMMA	→ ,	1
NP	→ NP Conj NP	0.2	N	→ muzzle	0.2			
			N	→ owner	0.2			

- The "context-free" assumption doesn't preclude relaxing probabilistic locality:

VP	→ V NP	0.5
VP	→ V	0.5
V	→ scratched	0.25
V	→ removed	0.25
V	→ arrived	0.5

Replaced by

⇒

VP	→ Vtrans NP	0.45
VP	→ Vtrans	0.05
VP	→ Vintrans	0.45
VP	→ Vintrans NP	0.05
Vtrans	→ scratched	0.5
Vtrans	→ removed	0.5
Vintrans	→ arrived	1

(Johnson, 1999; Klein & Manning, 2003)

Result

When the dog arrived the vet and his new assistant removed the muzzle.

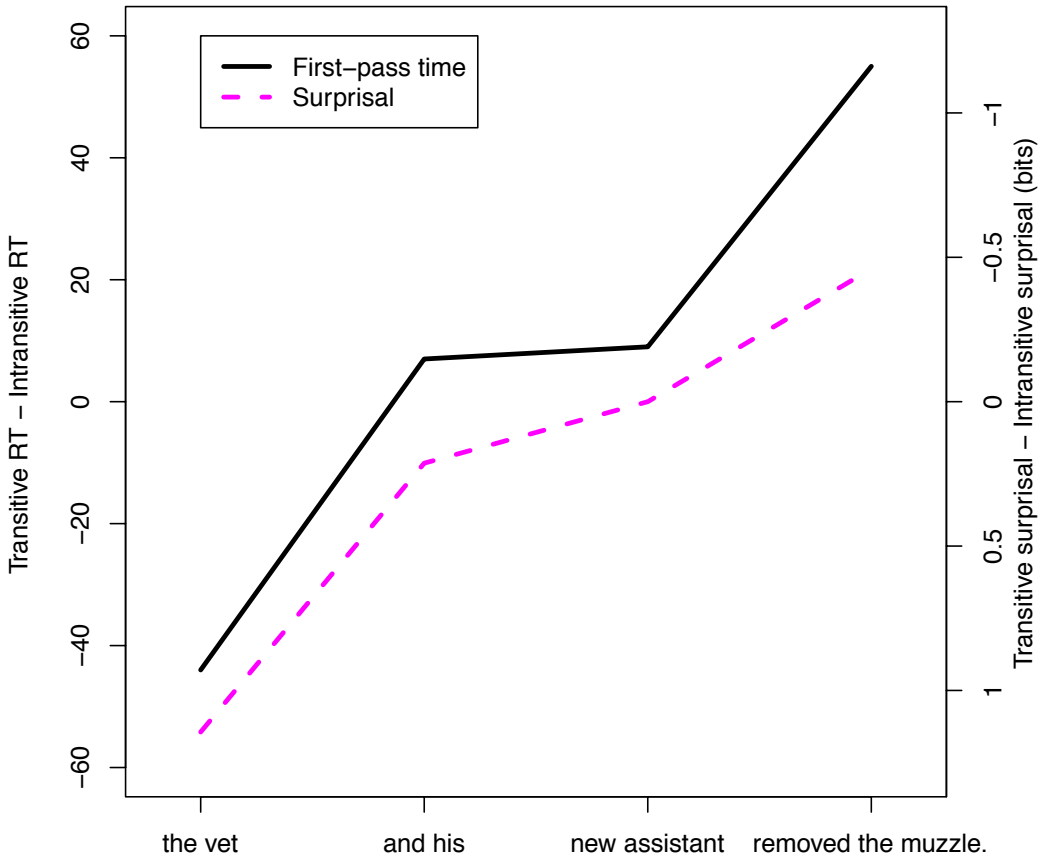
When the dog scratched the vet and his new assistant removed the muzzle.

Transitivity-distinguishing PCFG

Condition	Ambiguity onset	Resolution
Intransitive (arrived)	2.11	3.20
Transitive (scratched)	0.44	8.04

Move to broad coverage

- Instead of the pedagogical grammar, a “broad-coverage” grammar from the parsed Brown corpus (11,984 rules)
- Relative-frequency estimation of rule probabilities (“vanilla” PCFG)



Surprisal vs. predictability in general

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

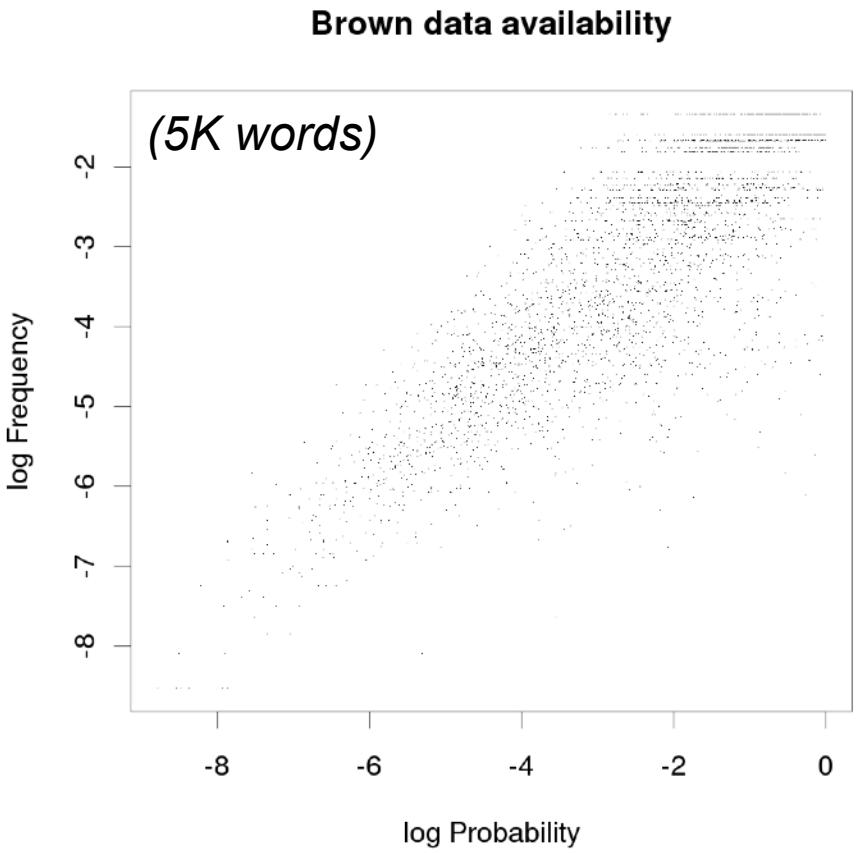
$$\left[\approx \log \frac{1}{P(w_i|w_{1\ldots i-1})} \right]$$

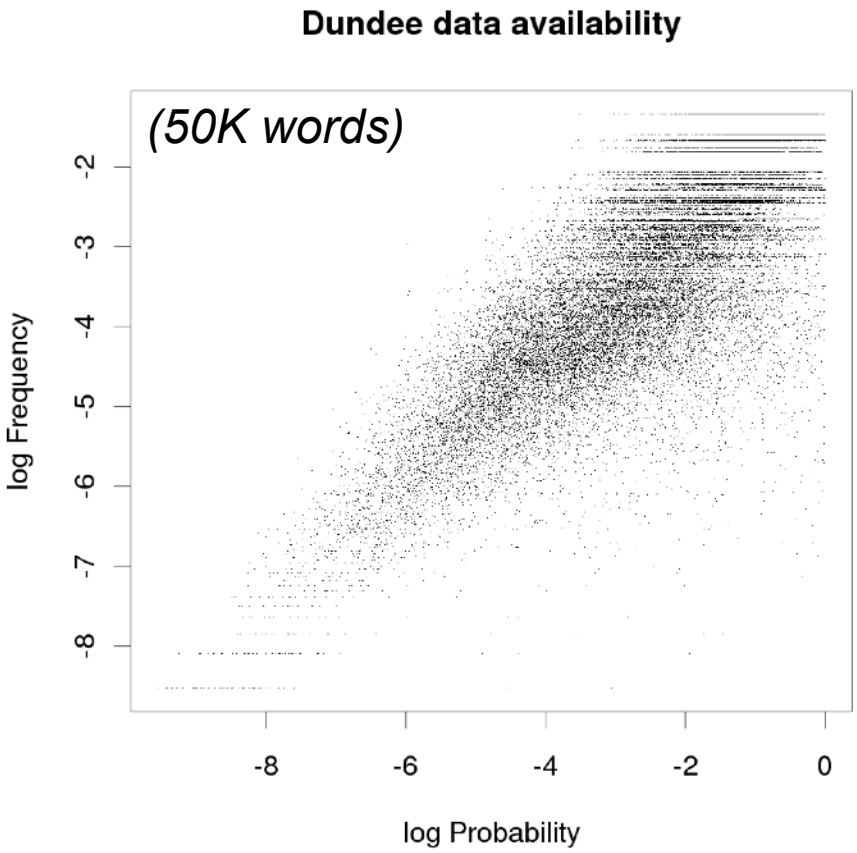
- But is there evidence for *surprisal* as the specific function relating probability to processing difficulty?

(Smith & Levy, 2013)

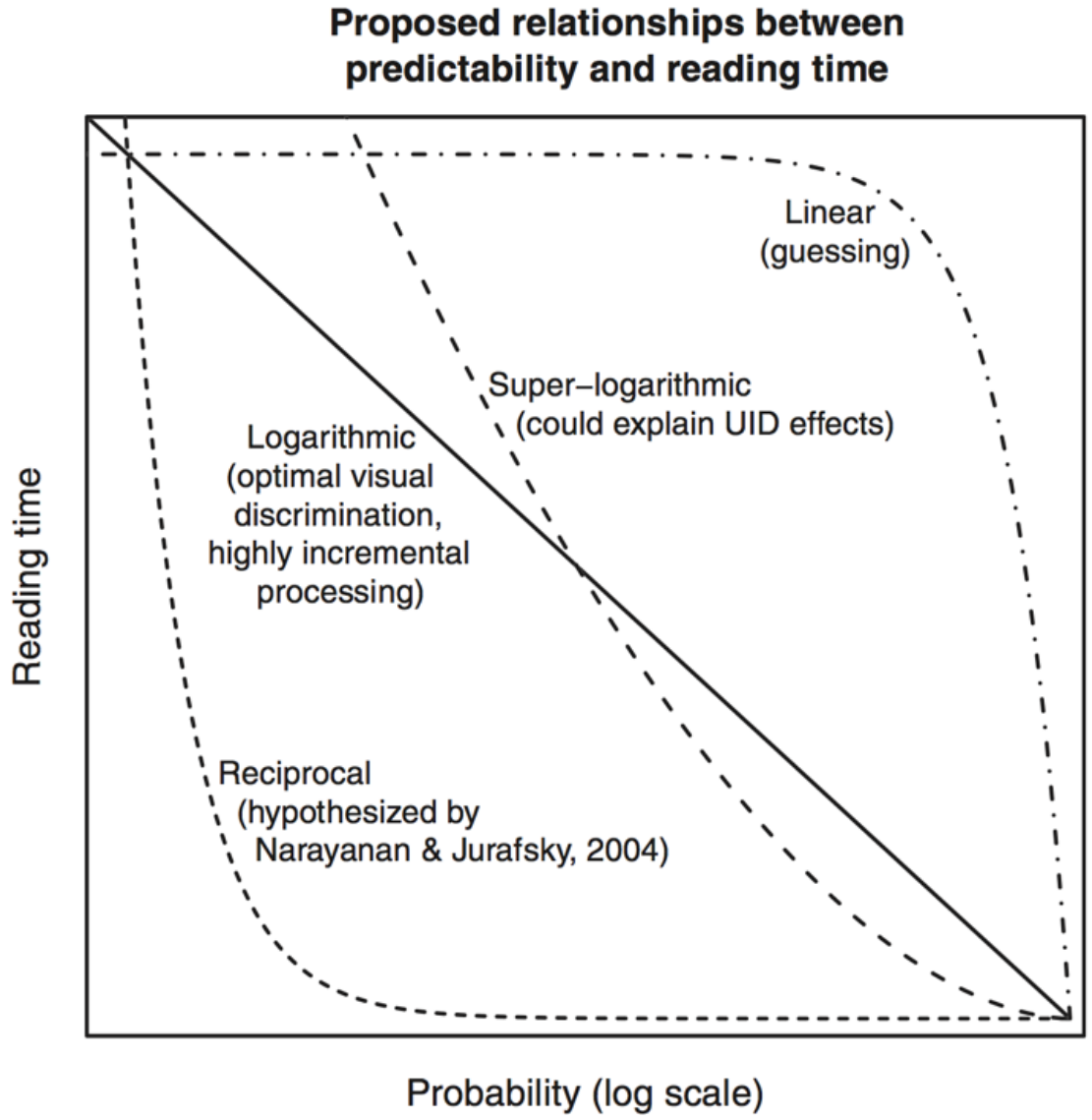
Estimating probability/time curve shape

- As a proxy for “processing difficulty,” reading time in two different methods: self-paced reading & eye-tracking
- Challenge: we need big data to estimate curve shape, but probability correlated with confounding variables



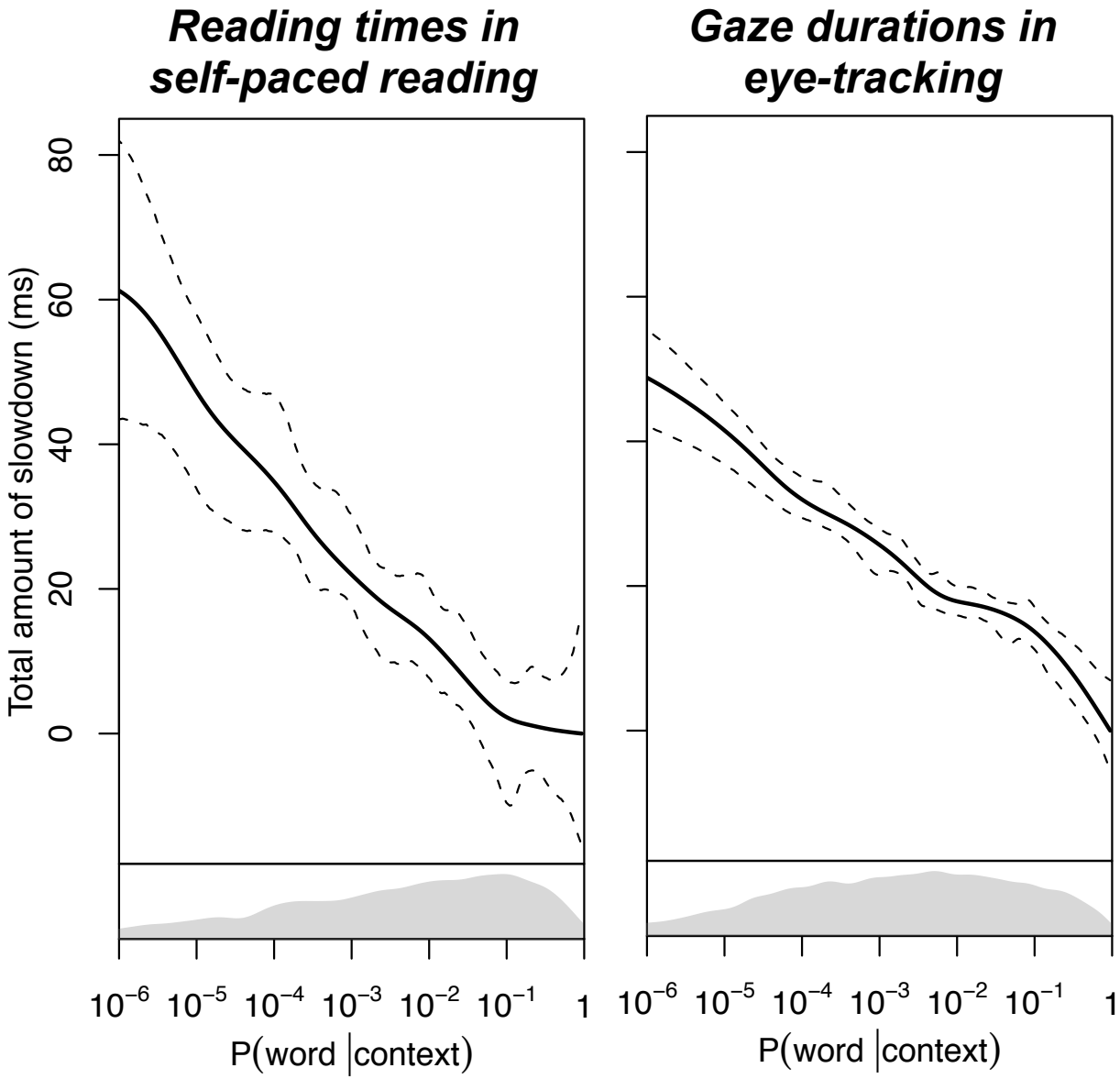


Hypothesized curve shapes



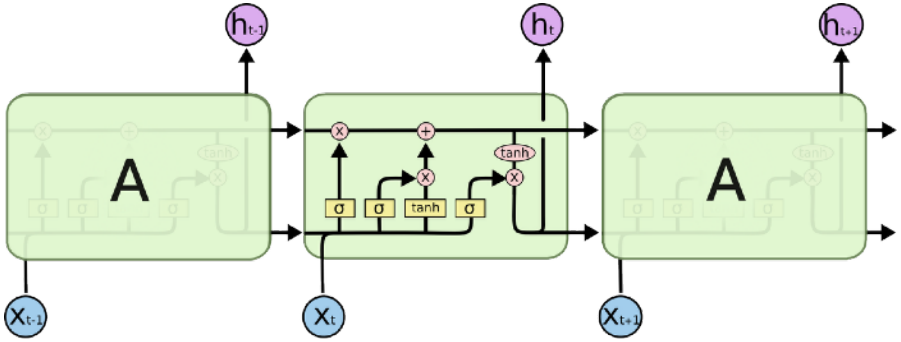
Estimating probability/time curve shape

- GAM regression: total contribution of word (trigram) probability to RT near-linear over 6 orders of magnitude!



(Smith & Levy, 2013)

What has state-of-the-art AI learned as “English”?



The girl who the newspaper.now calls his girlfriend has really been hateful . ✓

The monologue that the actor who the movie industry.likes made silent was being uploaded . ✓ (!)

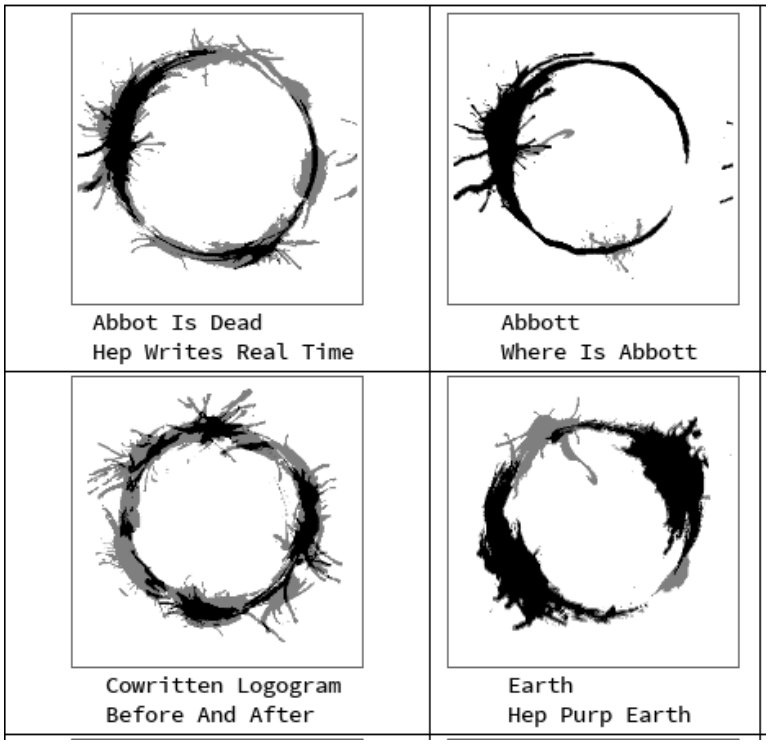
The man who the car.has gazed longingly at for years . ✗

The athlete who the restaurant.would justify decided to add the main West Coast restaurants to his menu and who hadn 't upgraded from his previous suite , into a more <UNK> steakhouse in New York . ? ? ?

(Jozefowicz et al., 2016; trained on ~800m words)

Technical question:

What generalizations are these models learning?



Testing syntax in neural language models with surprisal

Model	Architecture	Training data	Data size (tokens)	Reference
JRNN	LSTM	One Billion Word	$\sim$ 800 million	Jozefowicz et al. (2016)
GRNN	LSTM	Wikipedia	$\sim$ 90 million	Gulordava et al. (2018)
RNNG	RNN Grammar	Penn Treebank	$\sim$ 1 million	Dyer et al. (2016)
TinyLSTM	LSTM	Penn Treebank	$\sim$ 1 million	—

- LSTMs have no explicit syntactic state representations.
- RNN Grammars do, but it is not always clear how they use them in making predictions.

Simplest syntactic hierarchy: subordination

$$-\log P(\text{Completion}|\text{Context})$$

"No-matrix" variants
(No subsequent matrix clause)

Context | **Completion**

✓ *The doctor studied the textbook .* ◯

✗ *As the doctor studied the textbook .* ◯

Surprisal difference (should be *positive*)

"Matrix" variants
(There is a subsequent matrix clause)

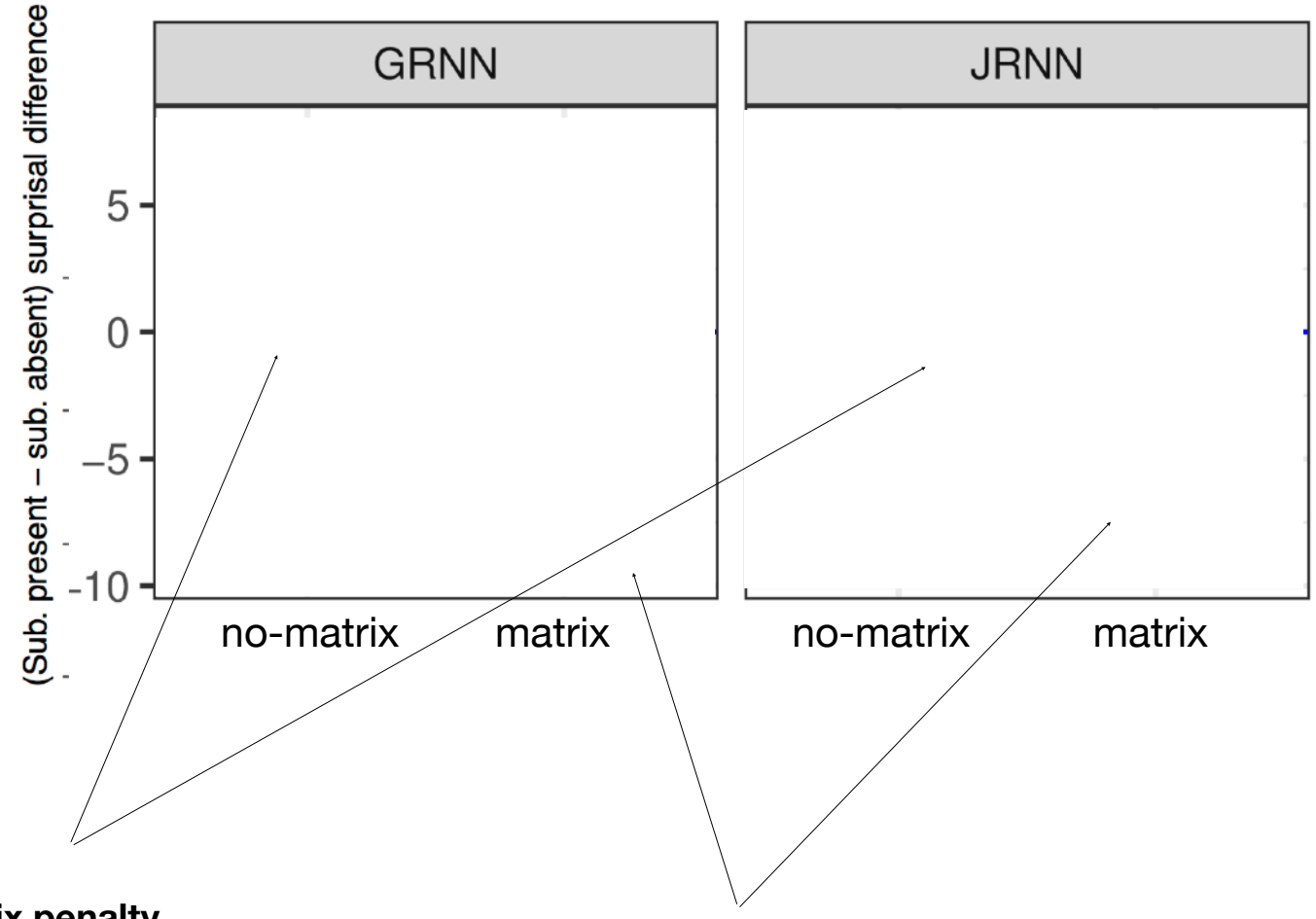
? *The doctor studied the textbook , the nurse walked into the office .*

✓ *As the doctor studied the textbook , the nurse walked into the office .*

Surprisal difference (should be *negative*)

(Futrell et al. 2019, NAACL)

Effects of Subordinate Clauses

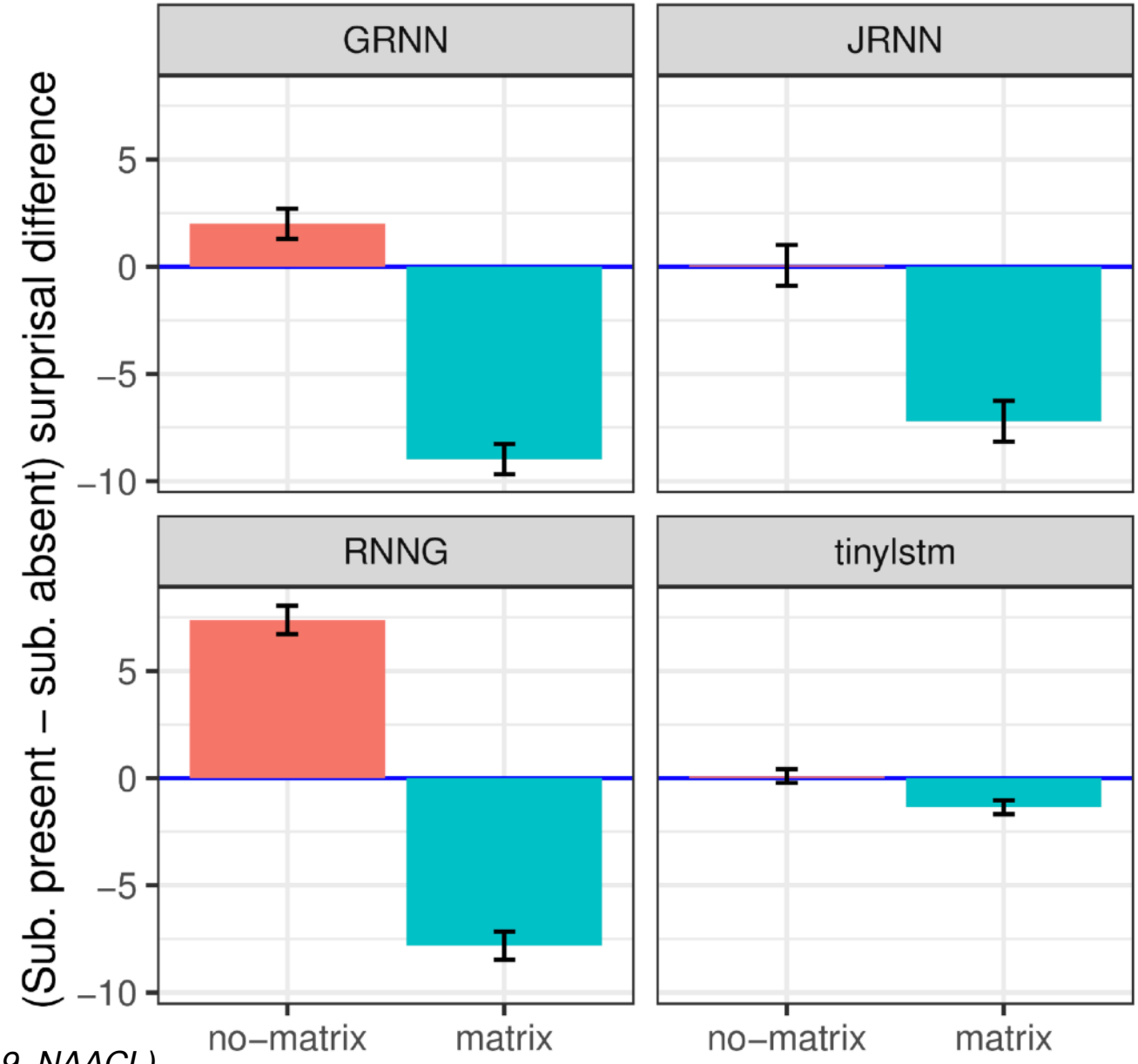


No-matrix penalty

Matrix licensing

(Futrell et al. 2019, NAACL)

Subordination: results



(Futrell et al. 2019, NAACL)

Subordination: summary

- All models learned *something* about the contingency between initial subordinator & need for a second clause
- Explicit representation of grammatical structure substantially sharpened that contingency

(Futrell et al. 2018, ArXiv; Futrell et al. 2019, NAACL)

Garden-pathing

(a) [transitive, −comma]

When the dog scratched the vet with his new assistant removed the muzzle.

(b) [transitive, +comma]

When the dog scratched, the vet with his new assistant removed the muzzle.

difficulty here

easier

(c) [intransitive, −comma]

When the dog arrived the vet with his new assistant removed the muzzle.

(d) [intransitive, +comma]

When the dog arrived, the vet with his new assistant removed the muzzle.

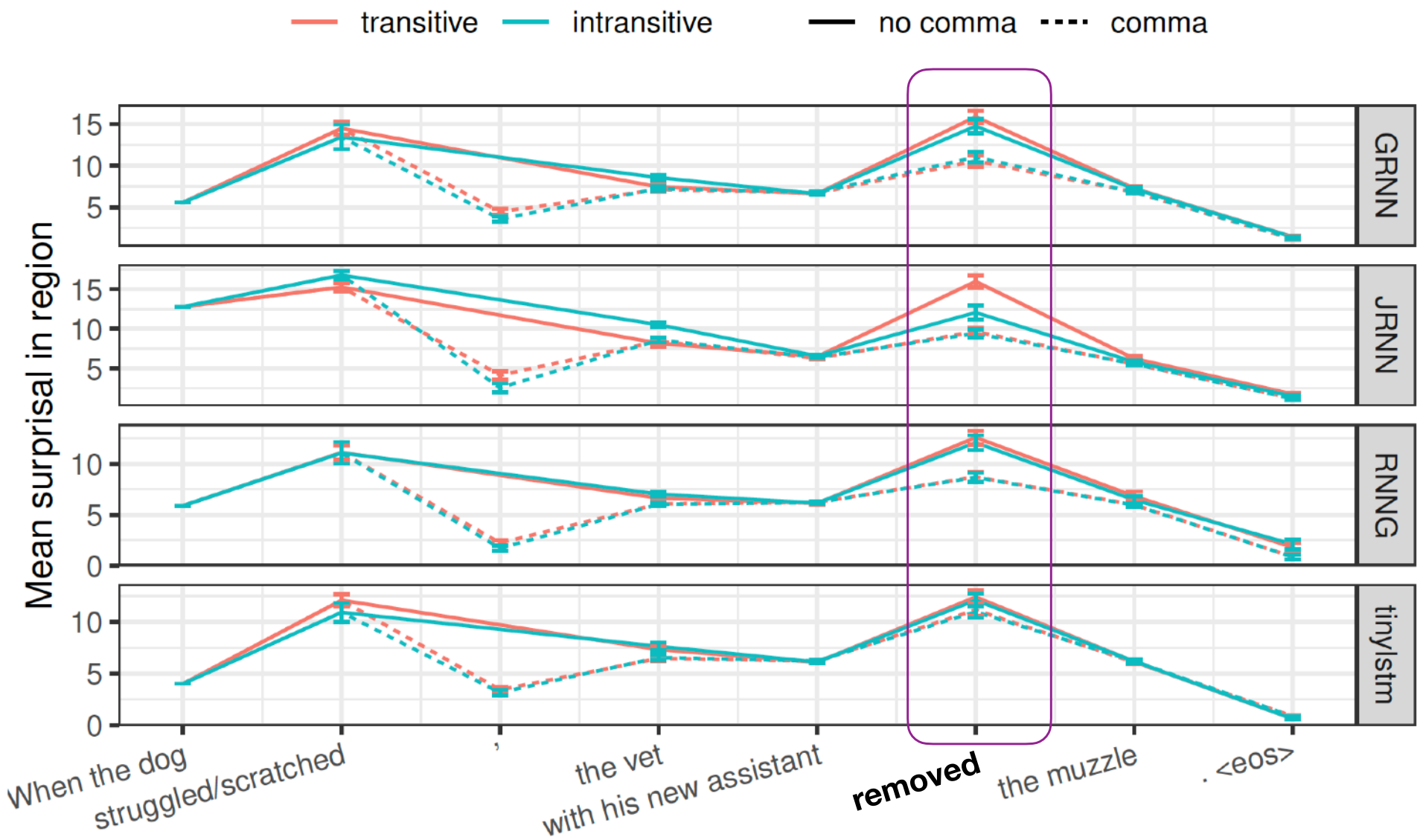
$$S(x) = -\log P(\text{removed} \mid \text{Context of version } x)$$

(i) $S(a) > S(b)$

(ii) $S(a) - S(b) > S(c) - S(d)$

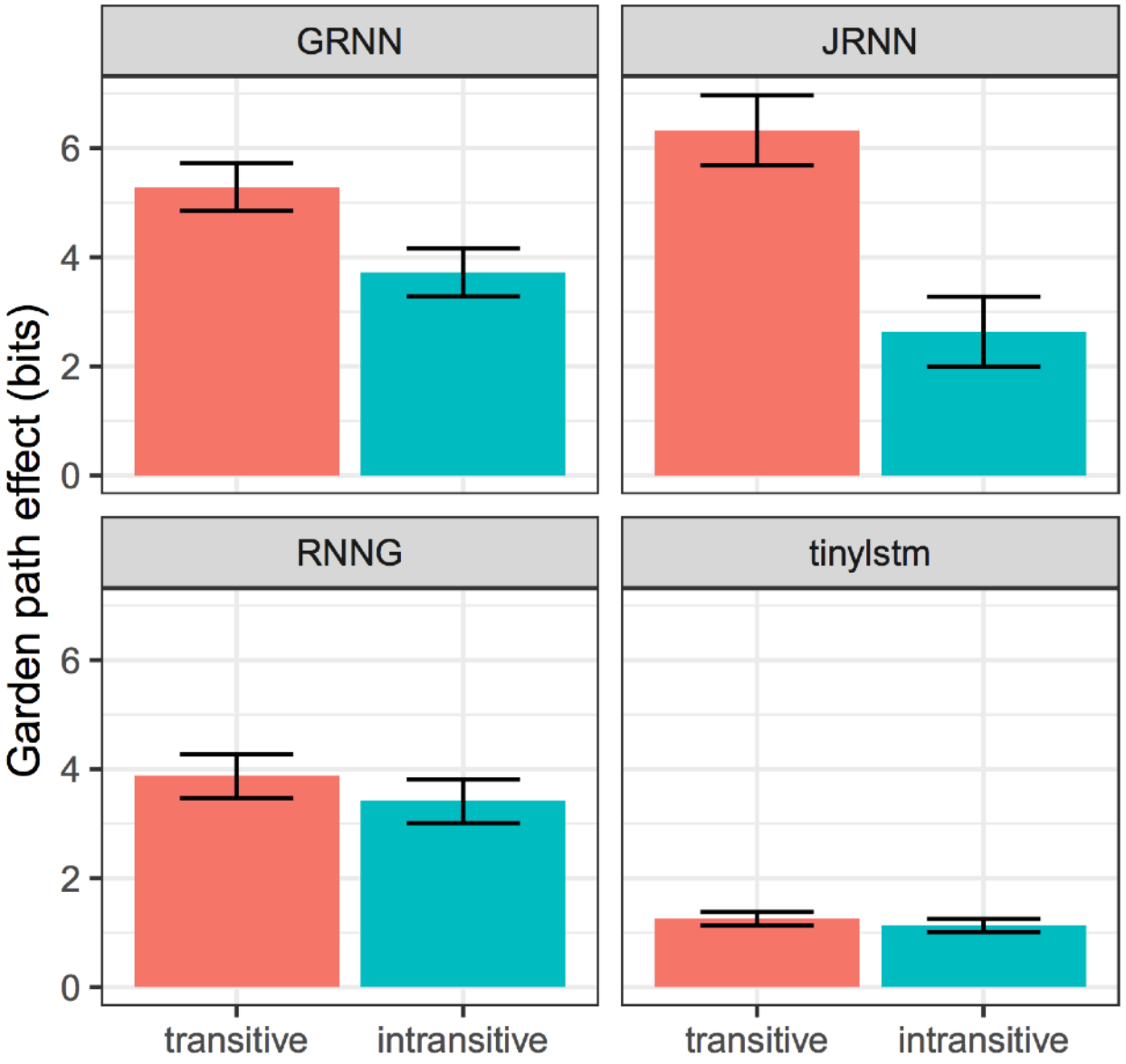
(Frazier & Rayner, 1982; Adams et al., 1998; Mitchell, 1987; van Gompel & Pickering, 2001; Staub, 2007)

Region-by-region surprisal profiles



(Futrell et al. 2019, NAACL)

NP/Z Garden Path Results



(Futrell et al. 2019, NAACL)

NP/Z Garden Paths: Degradation Over Time

(a) [short, −comma] As the author studying Babylon in ancient times wrote the book **grew**.

(b) [short, +comma] As the author studying Babylon in ancient times wrote, the book **grew**.

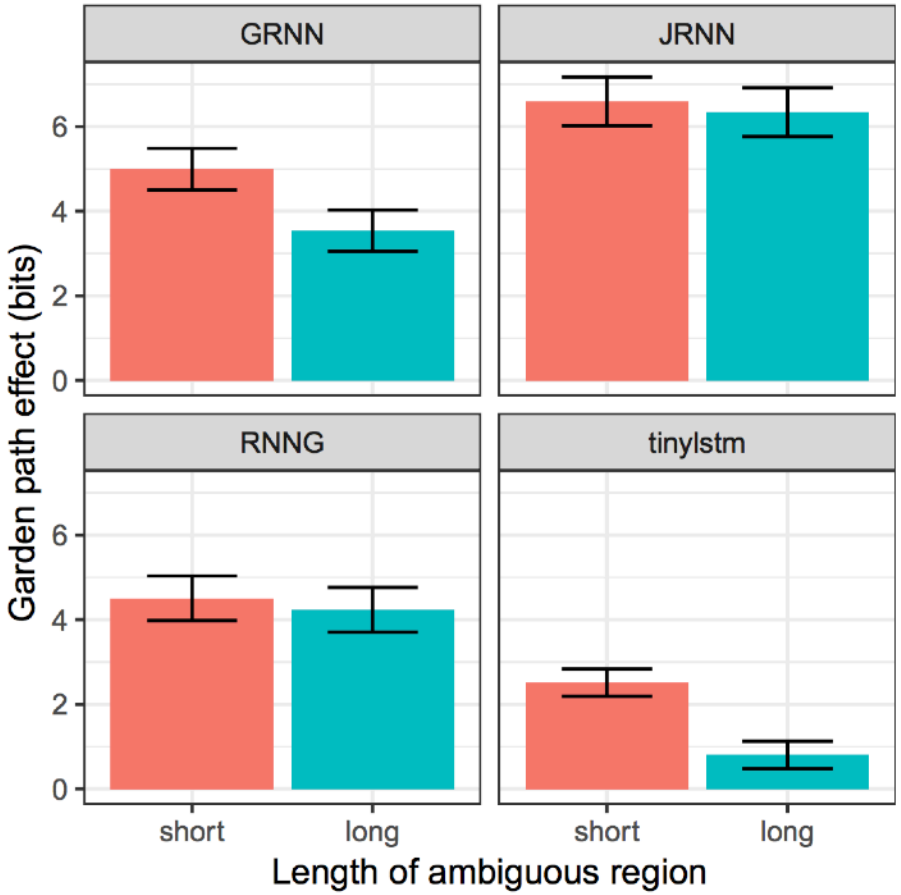
(c) [long, −comma] As the author wrote the book studying Babylon in ancient times **grew**.

(d) [long, +comma] As the author wrote, the book studying Babylon in ancient times **grew**.

(Warner & Glass, 1987; Ferreira & Henderson, 1991; Tabor & Hutchins, 2004; Levy et al., 2009)

Prediction:

$$S(a) - S(b) \approx S(c) - S(d) > 0$$



(Futrell et al. 2019, NAACL)

"Digging in" in human NP/Z garden-pathing

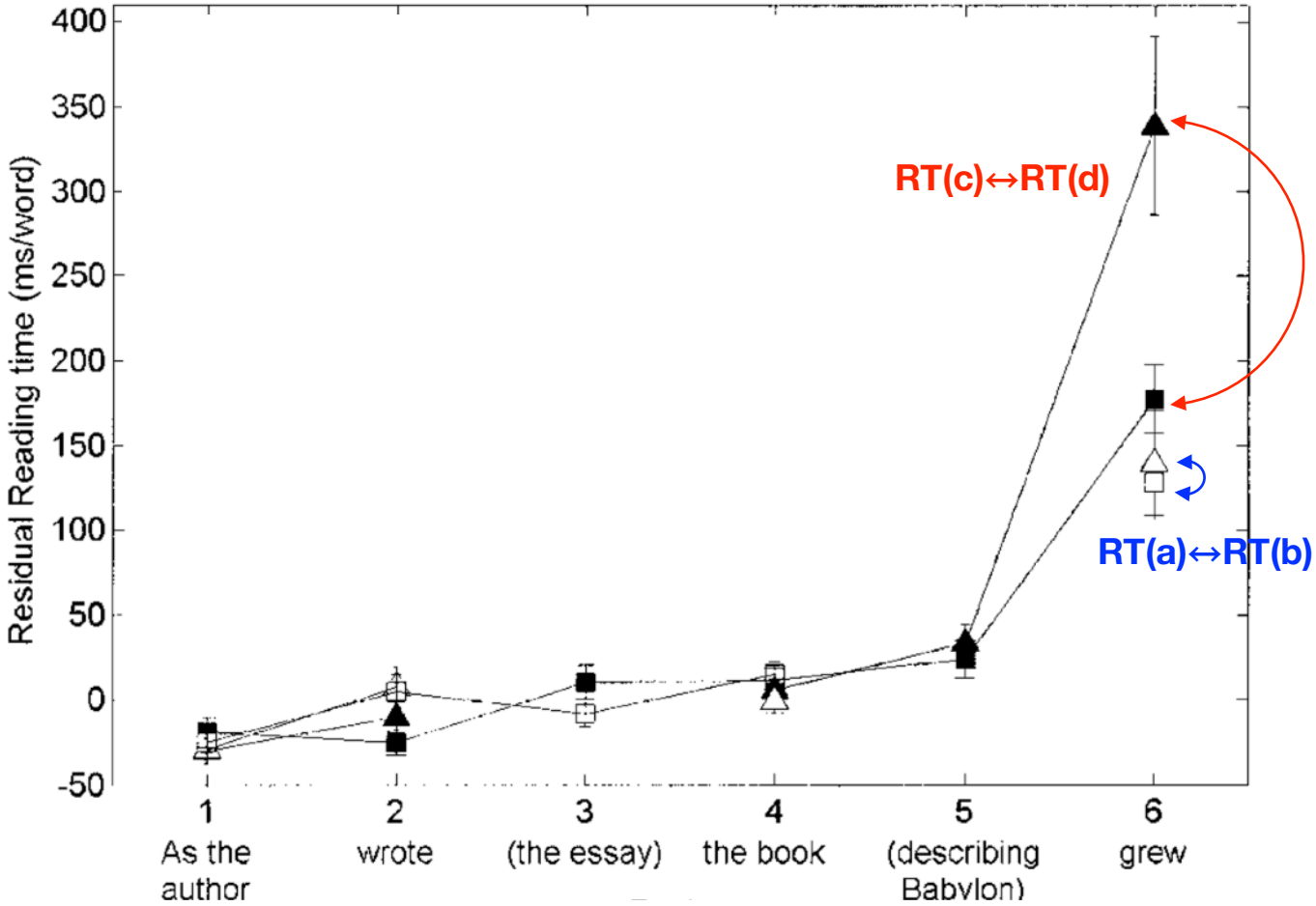
△(a) [short, −object] As the author wrote the book **grew**.

□(b) [short, +object] As the author wrote the essay the book **grew**.

▲(c) [long, −object] As the author wrote the book describing Babylon **grew**.

■(d) [long, +object] As the author wrote the essay the book describing Babylon **grew**.

Surprisal in neural language models doesn't capture the human "digging-in" effect

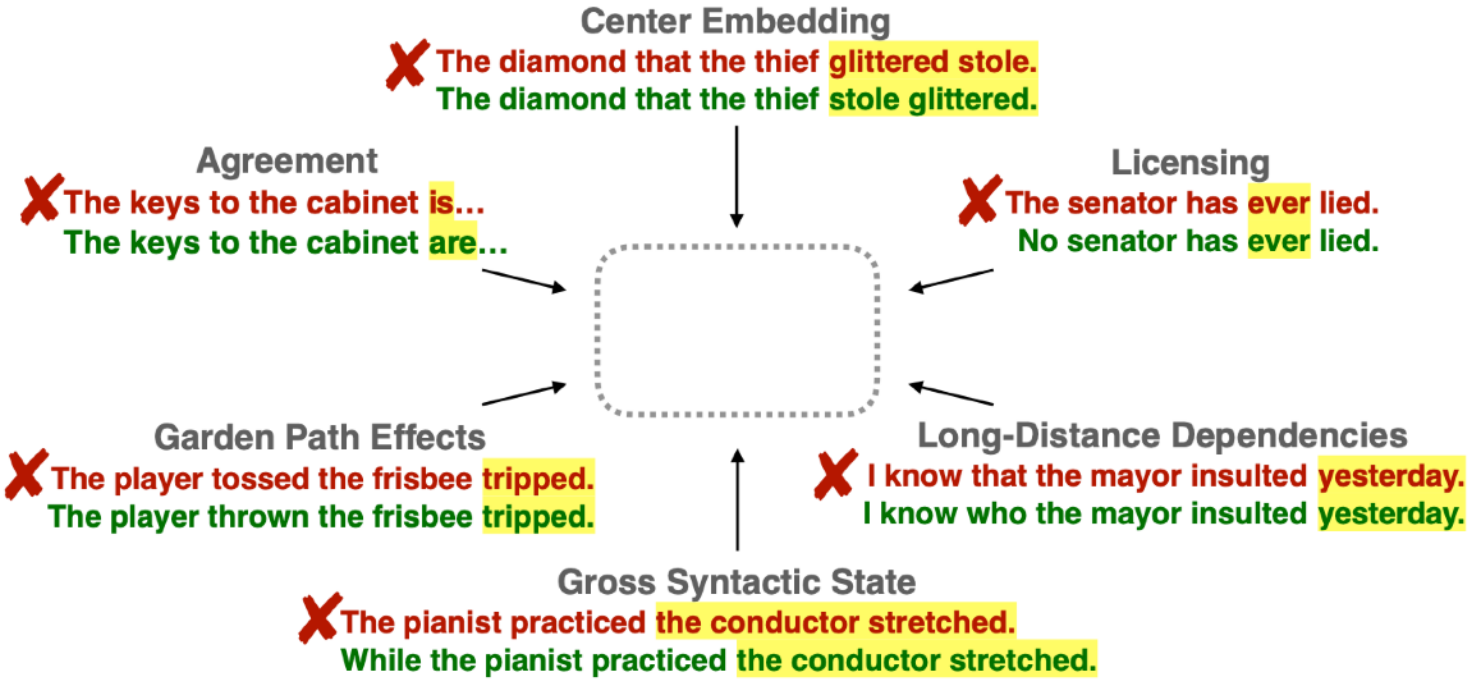


(Tabor & Hutchins, 2004)

NP/Z garden pathing: summary

- All models show evidence of a syntactic garden path that can be blocked by a comma
- Only models with larger amounts of data show verb transitivity-based garden-path modulation
- Not all models robustly maintain syntactic state-like distinctions over long stretches of intervening material
  - Explicit grammatical representations seem to help with this

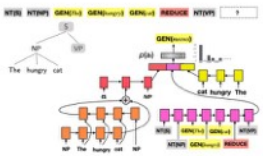
Scaling up syntactic evaluation

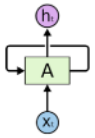


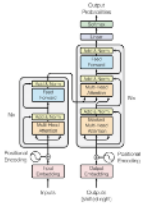
(Hu et al., 2020)

Controlled experiments

- What model properties affect syntactic generalization?







BLLIP-LG (42M)

BLLIP-MD (14M)

BLLIP-SM (5M)

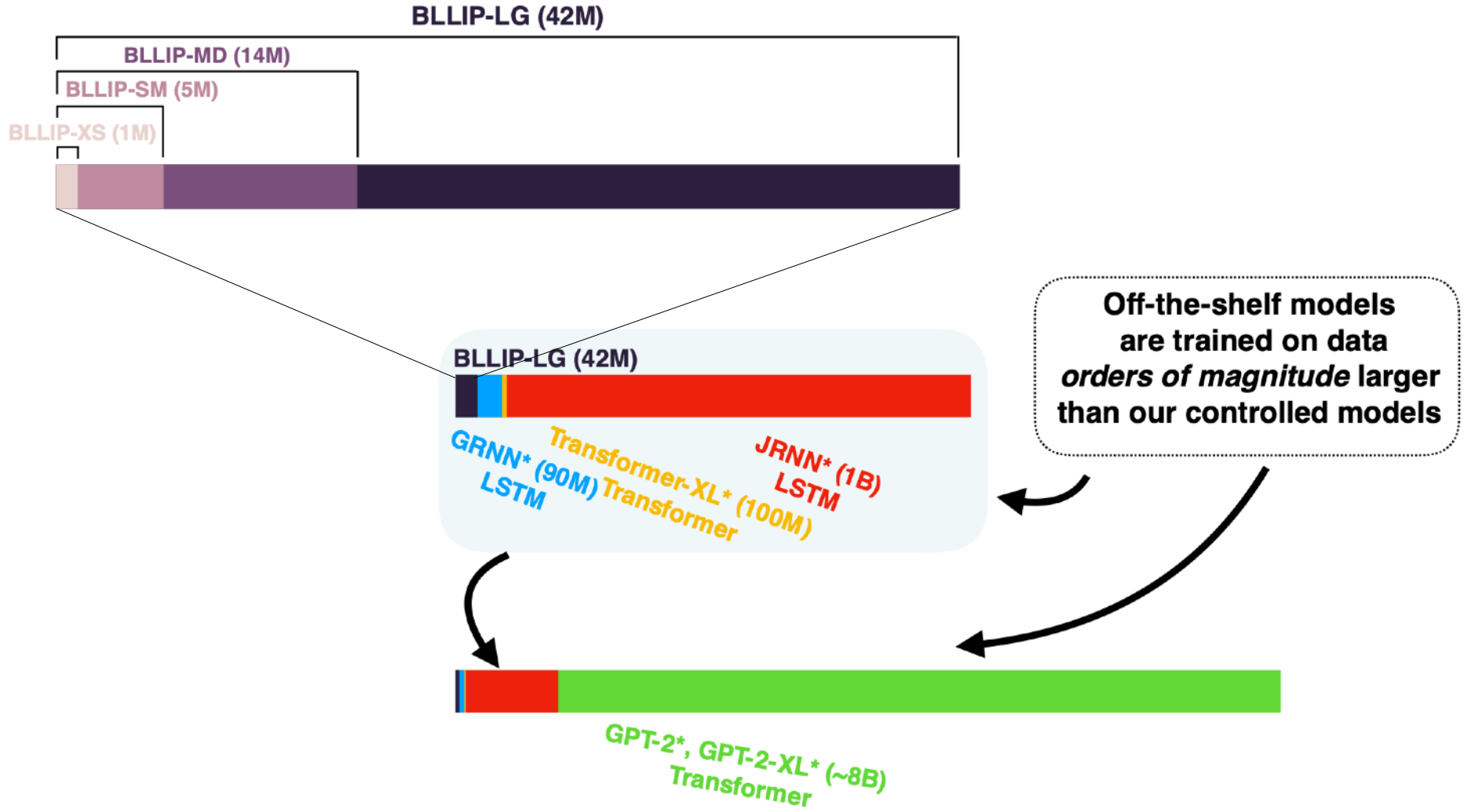
BLLIP-XS (1M)

n-gram
LSTM
ON-LSTM
RNNG
GPT-2

- **5 architectures x 4 training data sizes = 20 model types**
- **Multiple random seeds**

- Additionally test 5 off-the-shelf models for comparison

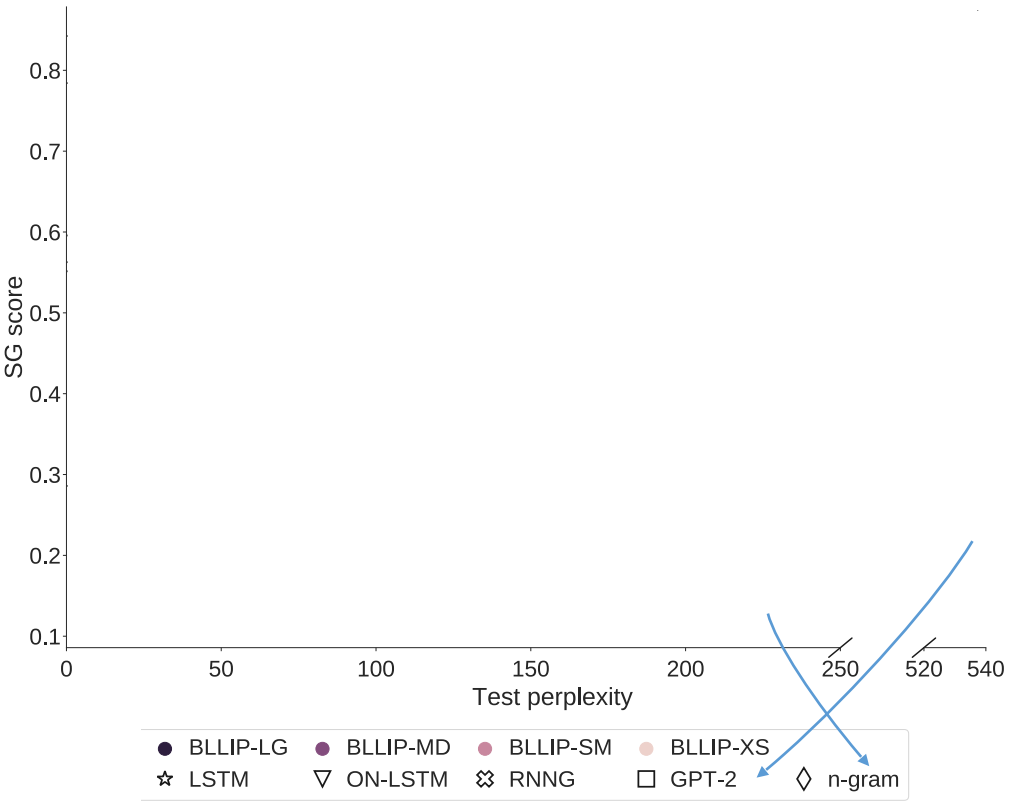
(Hu et al., 2020)


BLLIP-LG (42M)
BLLIP-MD (14M)
BLLIP-SM (5M)
BLLIP-XS (1M)
BLLIP-LG (42M)
GRNN* (90M) LSTM
Transformer-XL* (100M) Transformer
JRNN* (1B) LSTM
Off-the-shelf models are trained on data orders of magnitude larger than our controlled models
GPT-2*, GPT-2-XL* (~8B) Transformer

(Hu et al., 2020)

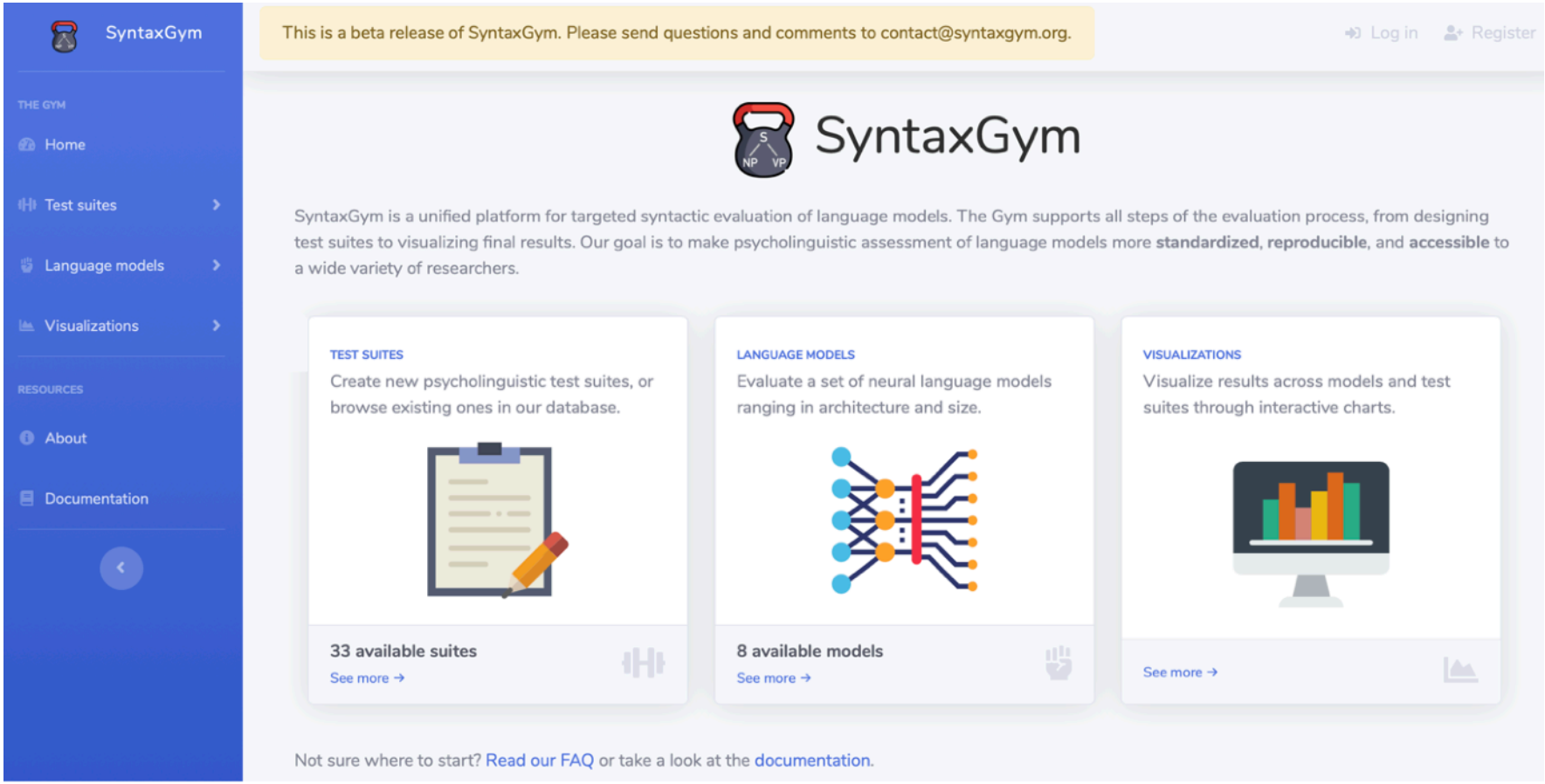
Scaled-up evaluation results

Dissociation between test set perplexity and syntactic generalization



(Hu et al., 2020)

Psycholinguistic tests of AI language models



http://syntaxgym.org

What makes a pun funny?

Researchers showed the robot ten puns, hoping that one of them would make it laugh. Unfortunately, no pun in ten did.

What makes a pun funny?

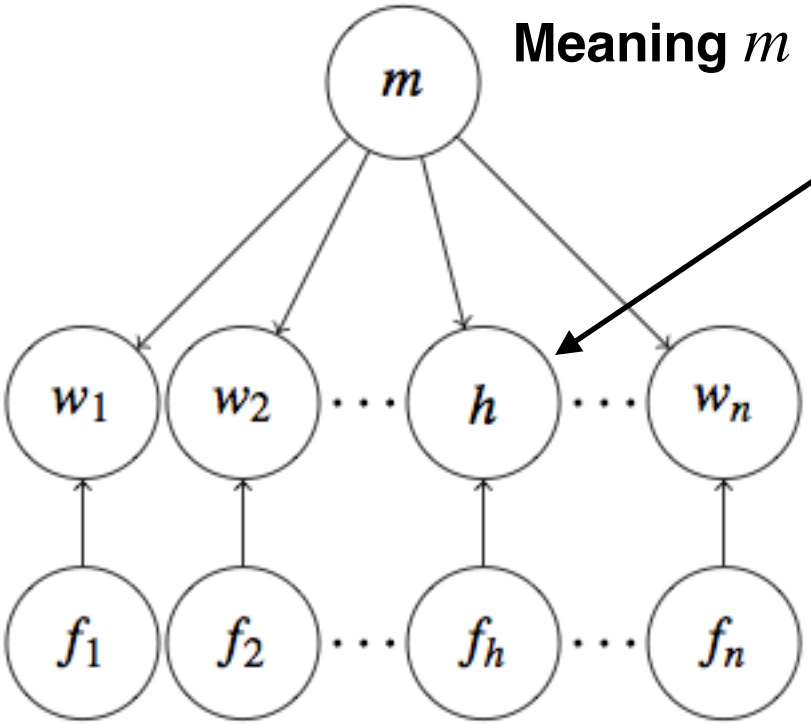
- One general idea in the humor literature: *incongruity*-- perceiving a situation from different viewpoints and finding the resulting interpretations incompatible
- We formalized this for *homophone* puns:

The magician got so mad he pulled his hare out.

- Two desiderata for incongruity:
  - The two interpretations of the target homophone should have *balanced support* from the context (**Ambiguity**)
  - The sources of support for the two interpretations should be *distinct from one another* (**Distinctiveness**)

(Kao, Levy, & Goodman, 2015)

Computational model of puns



Meaning m

Homophone h **(e.g.,** ***hare*****)**

Sentence $\vec{w}$

Semantic relevance $\vec{f}$

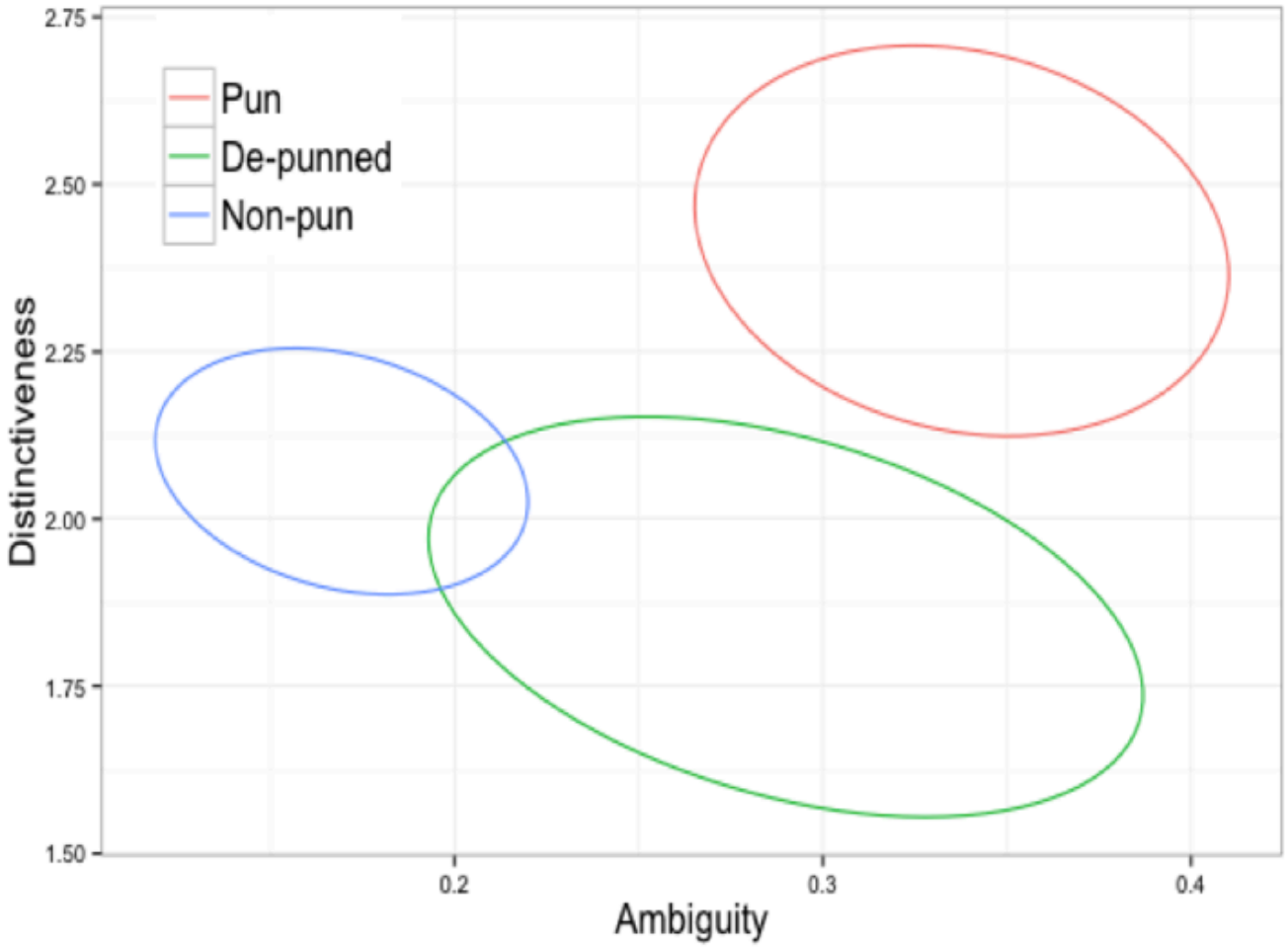
$$P(m, \vec{f} \mid \vec{w}) = \underline{P(m \mid \vec{w})} \underline{P(\vec{f} \mid m, \vec{w})}$$

Ambiguity
(should have high entropy)

Distinctiveness
(should look different for m_1 than for m_2)

Type	Example
Pun	The magician got so mad he pulled his hare out.
De-pun	The professor got so mad he pulled his hare out.
Non-pun	The hare ran rapidly across the field.
Non-pun	Some people have lots of hair on their heads.

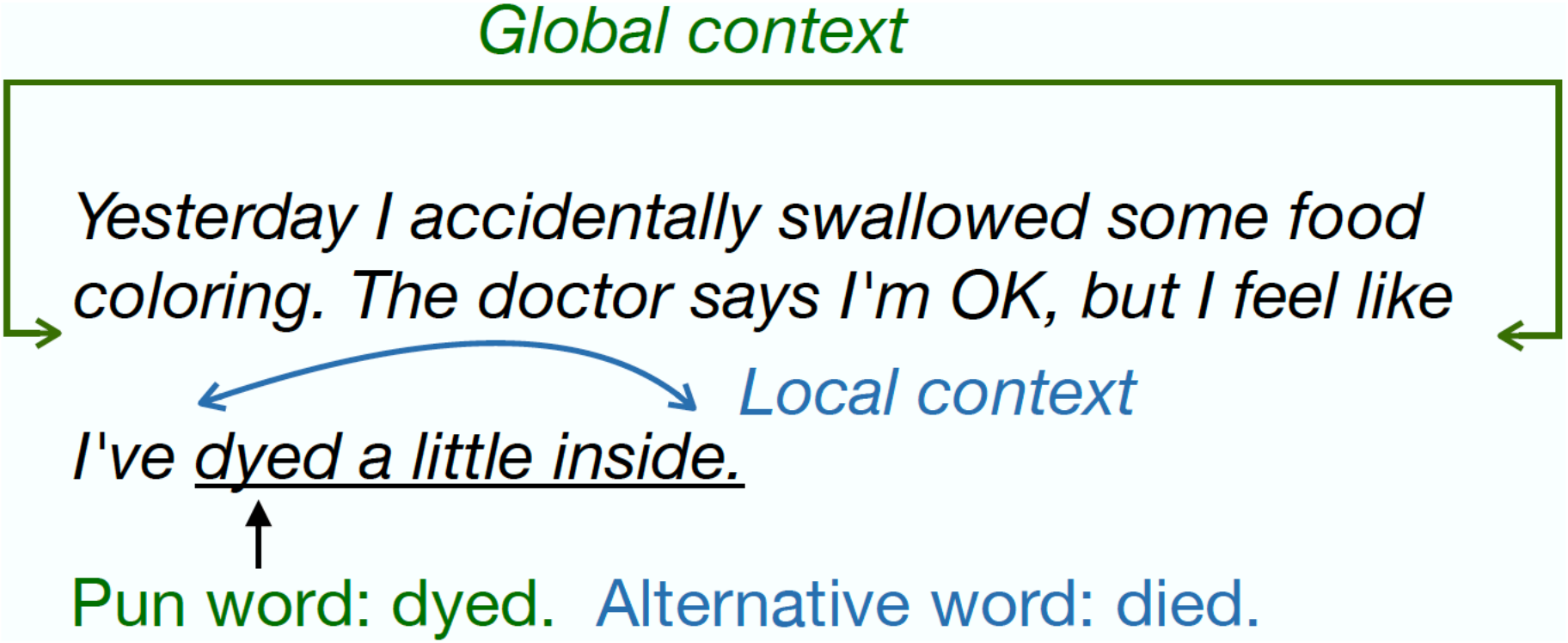
Results


Pun
De-punned
Non-pun
Distinctiveness
2.75
2.50
2.25
2.00
1.75
1.50
0.2
0.3
0.4
Ambiguity

Results

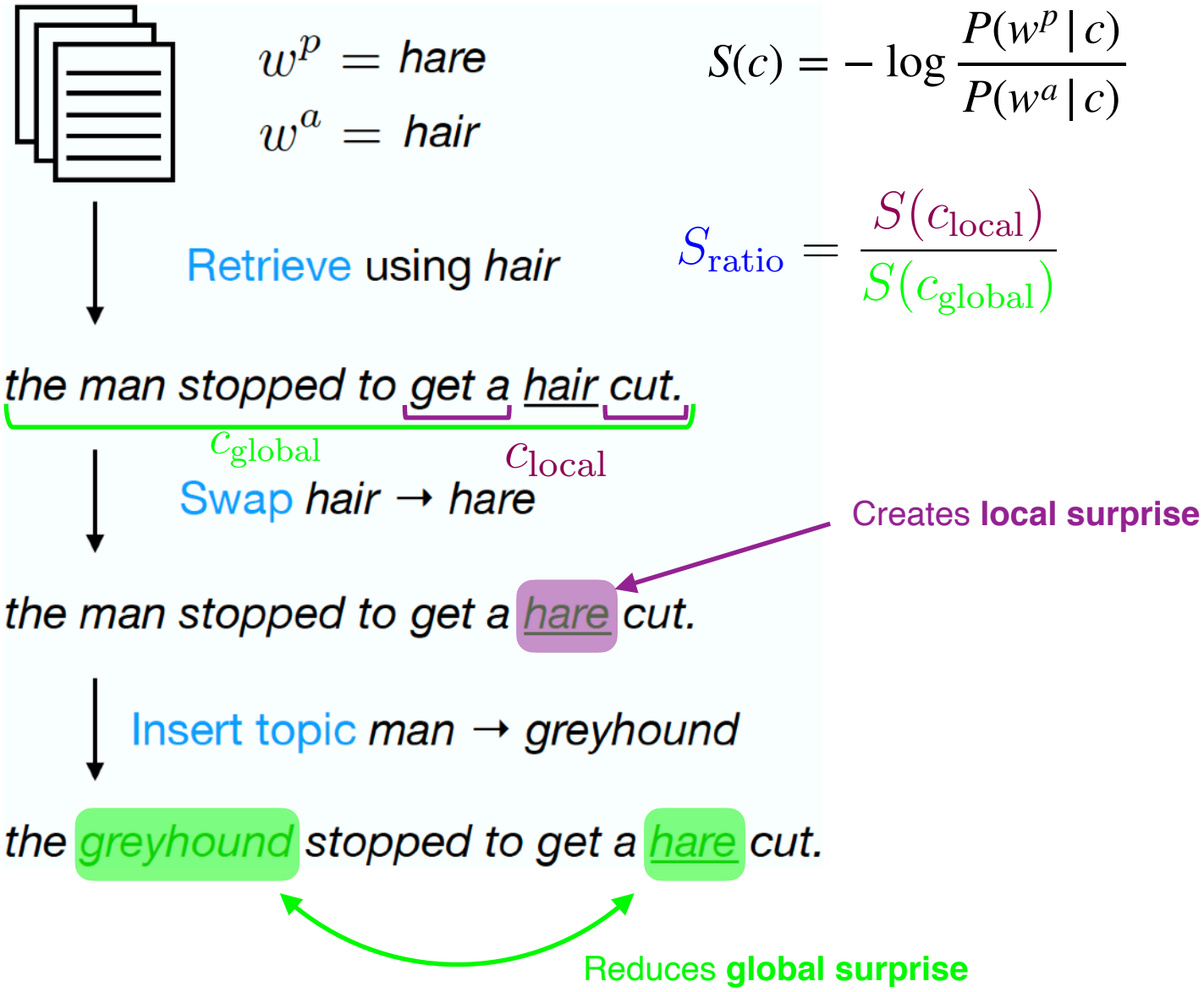
*m*1	*m*2	Type	Sentence and Semantic Focus Sets	Amb.	Disj.	Funniness
hare	hair	Pun	The **magician** got so mad he **pulled** his **hare** out.	0.378	2.291	1.714
		De-pun	The professor got so mad he **pulled** his **hare** out.	0.048	1.832	−0.103
		Non-pun	The **hare ran rapidly** through the **fields**.	0.447	1.677	−0.400
		Non-pun	Most **people** have **lots** of **hair** on their **heads**.	0.0004	2.807	−0.343
tiers	tears	Pun	It was an **emotional wedding**. Even the **cake** was in **tiers.**	0.612	2.311	1.541
		De-pun	It was an **emotional wedding**. Even the **mother-in-law** was in **tiers**.	0.189	1.802	0.057
		Non-pun	**Boxes** are **stacked** in **tiers** in the warehouse.	0.194	2.089	−0.560
		Non-pun	**Tears** ran down her **cheeks** as she watched a **sad movie**.	0.0003	3.283	−0.569

More recent work on pun generation



(He et al., 2019, NAACL)

More recent work on pun generation



(He et al., 2019, NAACL)

Automatically vs. human-generated puns

(Yu et al., 2018)

(He et al., 2019)

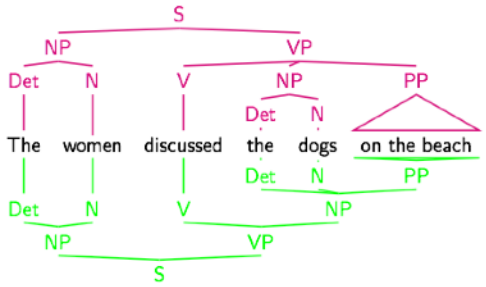
Method	Example	Rating
1. Pun/alternative word pair: butter – better		
NJD	He is going to come up with the butter a 'very good' approach to the world's economic crisis, the world's biggest economic climate.	1
SURGEN	Well, gourmet did it, he thought, it'd butter be right.	2
Human	Why did the dairy churn? The less said, the butter...	1.5
2. Pun/alternative word pair: peace – piece		
NJD	Further, he said, at the end of the peace, it's not clear that it will be a good example.	1
SURGEN	That's because negotiator got my car back to me in one peace.	1.5
Human	Life is a puzzle; look here for the missing peace.	3
3. Pun/alternative word pair: flour – flower		
NJD	Go, and if you are going on the flour.	1
SURGEN	Butter want to know who these two girls are, the new members of the holy flour.	1.5
Human	Betty crocker was a flour child.	4.5
4. Pun/alternative word pair: wait – weight		
NJD	Gordon Brown, Georgia's prime minister, said he did not have to wait, but he was not sure whether he had been killed.	0
SURGEN	Even from the outside, I could tell that he'd already lost some wait.	2
Human	Patience is a virtue heavy in wait.	3

(He et al., 2019, NAACL)

Evaluation on existing pun datasets

Metric	Pun and non-pun				Pun and swap-pun		Pun			
	SEMEVAL		KAO		SEMEVAL		SEMEVAL		KAO	
Surprisal (S_{ratio})	**0.46**	p=0.00	**0.58**	p=0.00	**0.48**	p=0.00	0.26	p=0.15	0.08	p=0.37
Ambiguity	**0.40**	p=0.00	**0.59**	p=0.00	0.18	p=0.15	0.00	p=0.98	0.00	p=0.95
Distinctiveness	-0.17	p=0.10	**0.29**	p=0.00	0.15	p=0.24	**0.41**	p=0.02	**0.27**	p=0.00
Unusualness	**0.37**	p=0.00	**0.36**	p=0.00	0.19	p=0.12	0.20	p=0.27	0.11	p=0.18

Summary & Conclusions



- Probabilistic models over richly specified linguistic structures allow us to construct and test ever-richer theories of human language knowledge

$$\text{Value}(\text{Models} + \text{Behavioral data}) > \text{Value}(\text{Models}) + \text{Value}(\text{Behavioral data})$$

- View of comprehender as rational inference-drawer and actor takes us very far

If you'd like to learn more...

- 9.19: Computational Psycholinguistics, taught by me!
  - Current syllabus:

    https://canvas.mit.edu/courses/7745
- Other relevant courses:
  - 9.59J/24.905J: Laboratory in Psycholinguistics (Professor Edward Gibson, BCS)
  - 9.66: Computational Cognitive Science (Professor Joshua Tenenbaum, BCS & CSAIL)
  - 24.904: Language Acquisition (Professor Athulya Aravind, Linguistics)

Thank you for listening!

http://www.mit.edu/~rplevy