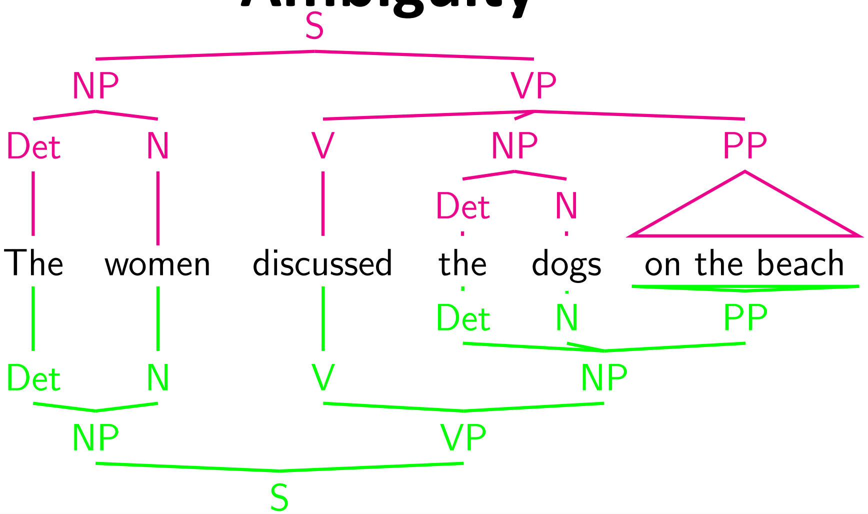
Language learning and processing in people and machines

Aida Nematzadeh   Richard Futrell   Roger Levy

DeepMind   UC Irvine   MIT

- How do humans communicate so well with language?

Ambiguity



Environmental noise

Memory Limitations

Incomplete knowledge of one's interlocutors

- How do we acquire the knowledge that enables this?
- And how can we get machines to do the same?

Overview of tutorial topics

- Human language acquisition (Aida)
  - Learning mechanisms
  - Word learning: theory & data
  - Structure learning: theory & data
- Human language comprehension (Roger)
  - Doing cognitive science through rational analysis
  - Revealing cognitive state with psycholinguistic experiments
  - Theory of human language comprehension
- Cognitive evaluation of NLP systems (Richard)
- Language evolution and emergence (Richard)

Some things to keep in mind today

- NLP and cognitive science offer each other a great deal
- NLP→cognitive science: formal theory-building for understanding human language learning & use
- Cognitive science→NLP: desiderata for human-like language processing systems
- We've seen impressive science & engineering progress, but many major open questions & problems remain
- ***There are great opportunities for everyone here!!!***

How Do Children Learn Language?

Aida Nematzadeh
nematzadeh@google.com
DeepMind

Language Acquisition in Children

Children **effortlessly learn** their language from a noisy and ambiguous input.

Language Acquisition in Machines

Understanding language acquisition might help us build AI systems that understand and produce natural languages.

Is Language Learned? How?

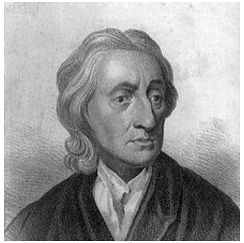
Is Language Learning Effortless?

Learning Mechanisms

Learning about Words

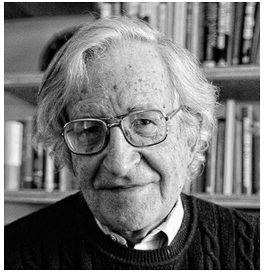
Learning the Structure

Nurture vs Nature



empiricism

nativism



Knowledge and reason come from experience.

Language: outcome of how children are **nurtured** (like table manner).

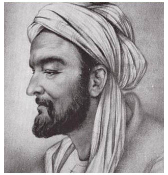
Mind has preexisting structure to interpret experience.

Language: outcome of **nature** -- an innate endowment (like upright posture).

Empiricism vs Nativism

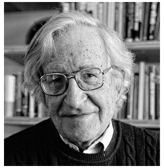
"The human intellect at birth is rather like a **tabula rasa**, a pure potentiality that is actualized through education and comes to know. Knowledge is attained through empirical familiarity with objects in this world from which one abstracts universal concepts."

Avicenna (980-1037 AD)



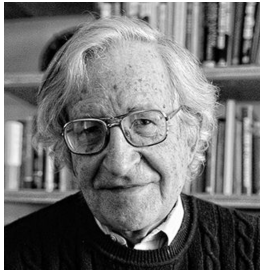
"**Language learning** is not really something that the child does; it is something that happens to the child placed in an appropriate environment, **much as the child's body grows and matures** in a predetermined way when provided with appropriate nutrition and environmental stimulation."

Chomsky (1928-)



Cognitive Revolution

behaviorism ⟷ cognitivism



Can explain behavior in terms of things external to mind.

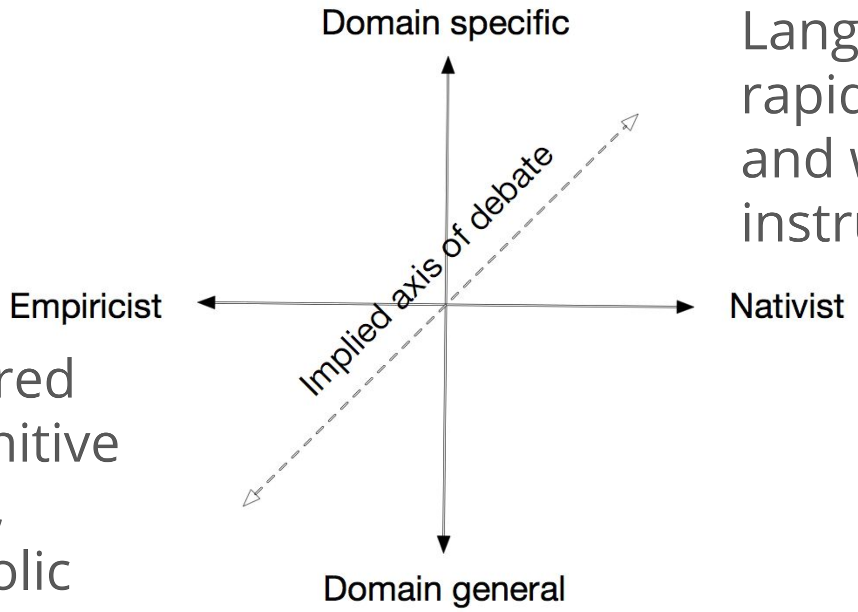
Language ~ verbal behavior

Explaining behavior requires understanding the mind.

Language ~ mental process

Domain-General vs Domain-Specific Learning

Language is acquired rapidly, effortlessly, and without direct instruction.

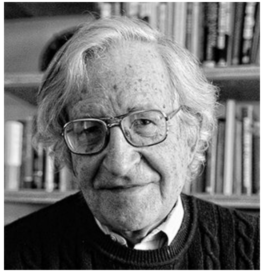


[Frank *et al*, 2019]

Language is acquired using general cognitive skills like memory, capacity for symbolic representation, and statistical learning.

Language for Communication

functionalism ⟷ formalism



Language is shaped by its communicative functions.

Language is acquired through communication (not passive observation).

Language form is independent of its function.

Acquisition of language is not affected by the fact that we use it to communicate.

Takeaways: Development vs Learnability

Modeling language development to shed light on its underlying mechanism.

Can we learn language (certain linguistic phenomena) from data?

Nature of Nature

Investigate the innateness/learnability of

- knowledge -- inborn linguistic knowledge?
- computational procedure -- domain-general or domain-specific learning mechanism?

Is Language Learned? How?

Is Language Learning Effortless?

Learning Mechanisms

Learning about Words

Learning the Structure

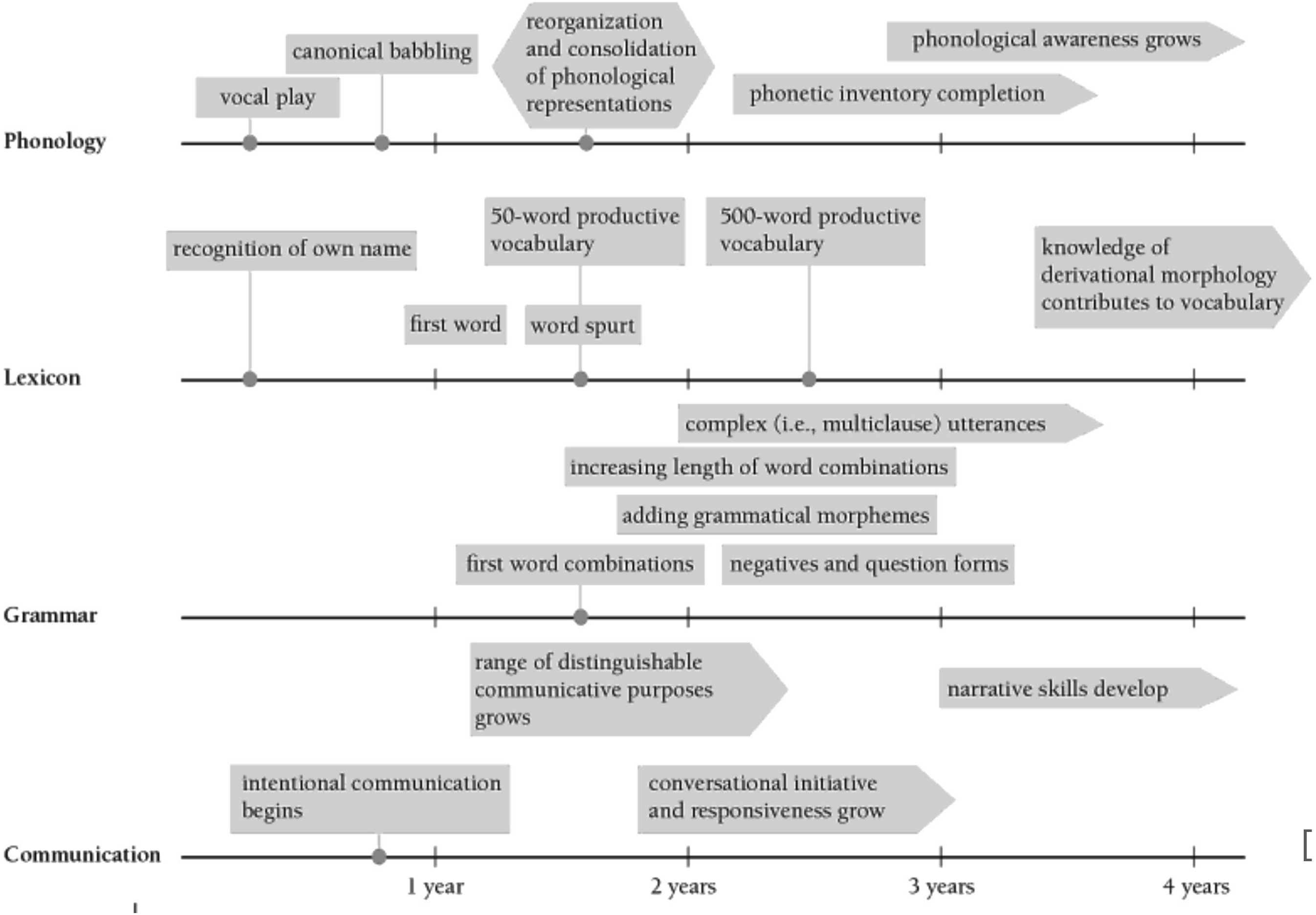
0-12m	12-24m	18-30m	24-48m
prelinguistic communication	single words	telegraphic speech	grammatical development
"bobo"	“mummy” “doggy”	“daddy sleep” “orange juice”	“I want some eggs” “Put it table”

Takes children 5 years (14,600h, 8h/day).

Would take adults 56 years (2920 weeks, 5h/week).

0-12m	12-24m	18-30m	24-48m
prelinguistic communication	single words	telegraphic speech	grammatical development
“bobo”	“mummy”	**“daddy sleep”**	“I want some eggs”
	“doggy”	“orange juice”	**“Put it table”**

Children make errors but learn to correct them.


Phonology
vocal play
canonical babbling
reorganization and consolidation of phonological representations
phonetic inventory completion
phonological awareness grows
Lexicon
recognition of own name
first word
50-word productive vocabulary
word spurt
500-word productive vocabulary
knowledge of derivational morphology contributes to vocabulary
Grammar
first word combinations
adding grammatical morphemes
increasing length of word combinations
negatives and question forms
complex (i.e., multiclause) utterances
Communication
intentional communication begins
range of distinguishable communicative purposes grows
conversational initiative and responsiveness grow
narrative skills develop
1 year
2 years
3 years
4 years

[Hoff, 2004]

Takeaways

Should AI models make the same mistakes as children?

Should we model all the domains at the same time?

Is Language Learned? How?
Is Language Learning Effortless?
Learning Mechanisms
Learning about Words
Learning the Structure

Babies as Statistical Learners [Saffran *et al*, Science 1996]

8-month-old infants learn within- and between-word transitional probabilities from novel speech.

- bidakupadotigolabutupiropadotibidaku

Statistical learning in other domains: phonology, syntax, & words. [Gomez *et al*, 2000; Mintz *et al*, 2002; Smith & Yu, 2008; Romberg & Saffran, 2010]

Statistical learning is domain- & species- general.

Babies as Rule Learners [Marcus *et al*, Science 1999]

Seven-month-old infants can learn simple "algebra-like" rules.

- "ga ti ti" "li la la" (ABB) or "li la li" "ga la ga" (ABA)

Rule learning is statistical learning? [Christiansen & Curtin, 1999; Seidenberg & Elman, 1999; McClelland & Plaut, 1999]

Babies as Social Learners

Sharing joint attention.

Understanding and sharing intention. [Tomasello *et al*, 2005]

Infants learn about phonetics by listening to native speakers but not their audio/video. [Kuhl *et al*, 2003]

Takeaways

What type of learning does each linguistic domain require?

What modeling frameworks are suitable for each?

Is Language Learned? How?
Is Language Learning Effortless?
Learning Mechanisms
Learning about Words
Learning the Structure

Word Learning Stages

Segmenting speech to words.

Mapping a meaning to words.

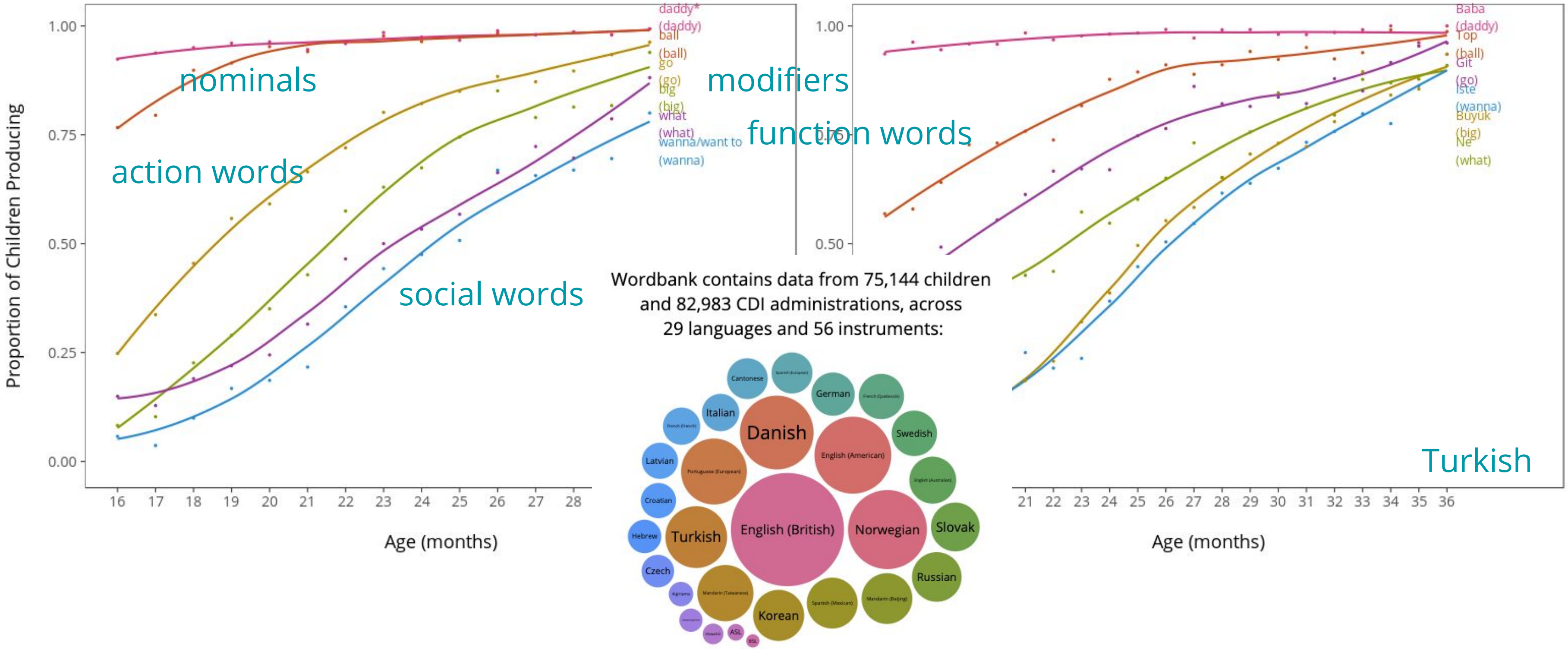
Context-bound Words

Used only in one context: saying “duck” **only** when hitting the toy to the bathtub. [Barrett, 1986]

Are parts of language games.

Function-specific understanding -- different from adults’ mental representations of words.

Early Words



http://wordbank.stanford.edu/

Word Learning Errors

Underextension: using words in a more restricted fashion; “dog” to refer to spaniels.

Overextension: using words more broadly; all four-legged animals as “doggie”.

- “cat”: cat, cat’s usual location on the top of TV when absent. [Rescorla, 1980]

Cross-situational Learning

People (as young as 12-month-old infants) are sensitive to the statistical regularities across situations. [Pinker 1989; Yu & Smith 2007; Smith & Yu, 2008]

A zant

Look at the zant!

Biases that Guide Word Learning

The input is noisy and ambiguous: many possible mappings/hypotheses for word meanings.

People learn word meanings from a few exposures.

Learned/innate biases might facilitate learning.

Biases that Guide Word Learning

mutual exclusivity bias
[Markman & Wachtel, 1988]
taxonomic bias
[Markman & Hutchinson, 1984; Markman, 1989]
basic-level bias
[Rosch *et al*, 1976; Markman, 1991]

whole-object bias [Markman, 1991]
shape bias [Smith & Jones, 1988]

attention
[Samuelson & smith, 1998;
Yu *et al*, 2017]

social-pragmatic biases
communicative intentions
[Bloom, 2000; Tomasello, 2001]
following eye gaze
[Baldwin, 1993]

syntax
[Brown, 1957;
Gelman & Markman, 1985]
noun bias
[Gentner, 1982]

The Whole-Object Bias [Markman, 1991]

What is dax?

Learn word labels for the whole object.

The Mutual Exclusivity Bias [Markman & Wachtel, 1988]

What is dax?

18-month children exhibit the bias. [Markman *et al*, 2003]

familiar object

unfamiliar object ✓

Limit the number of possible word labels for a familiar object.

The Basic-Level Bias

Zant → Golden Retriever?

Zant → dog (any dog breed)?

Zant → animal?

Cross-situational statistics are **consistent** with all.

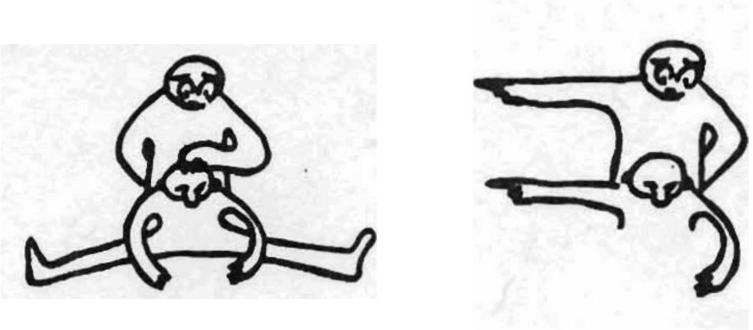
Why dog? A bias that focuses generalization to the **basic-level** (cognitively natural) categories.

Syntactic Bootstrapping

Language structure supports learning new verbs.

[Gleitman, 1990; Fisher et al, 1994]

[Naigles, 1990]



“The rabbit is gorping the duck.” or

“The rabbit and the duck are gorping.”

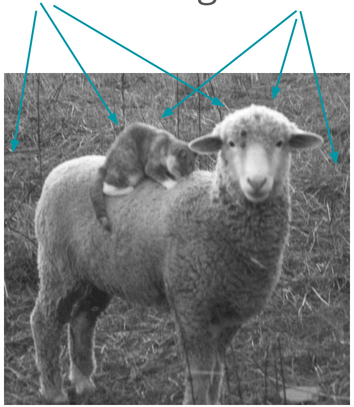
“where is gorping now?”

Modeling Word Learning

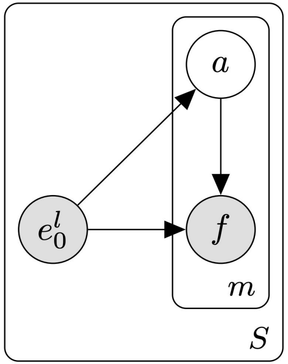
Solving the translation problem: mapping words to observations.

[Siskind, 1996; Yu & Ballard, 2007; Frank *et al*, 2009; Fazly *et al*, 2010; Nematzadeh *et al*, 2015]

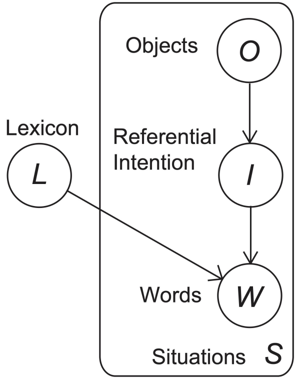
“the cat is sitting on the sheep”



objects



words



[Frank *et al*, 2009]

Is Language Learned? How?
Is Language Learning Effortless?
Learning Mechanisms
Learning about Words
Learning the Structure

Language is Productive

We have the capacity to produce and understand an infinite number of new sentences.

Two productive systems:

- Syntax: sentence structure; ordering of words.
- Morphology: structure of words & word parts.

Syntax: Level of Abstraction

"Rita drinks milk."

- Sentence → Rita + drinks + milk (not productive)
- Sentence → agent of action + action + theme

"Rita resembles Ray."

- Sentence → noun + verb + noun

What is origin of the variables and the rules?

Syntax: Type of Structure

Sentences have hierarchical structure.

- *"The (clever) cat cried (a river)."*
- S $\rightarrow$ NP + VP, NP $\rightarrow$ (det) + (adj) + N, VP $\rightarrow$ V + NP

Is human language use hierarchical? [Frank *et al*, 2012]

Morphology

Adds grammatical information to words.

- Plural s in English

Children learn morphology earlier when language is morphologically rich. [Peters, 1995]

Easy morphemes to learn: frequent, fixed form and relative position to stem, clear function.

Do Children Know Grammatical Rules?

Early word combinations are systematic.

- “my teddy” (possessor + possessed)
- “daddy sit” (actor + action)

Overgeneralization errors:

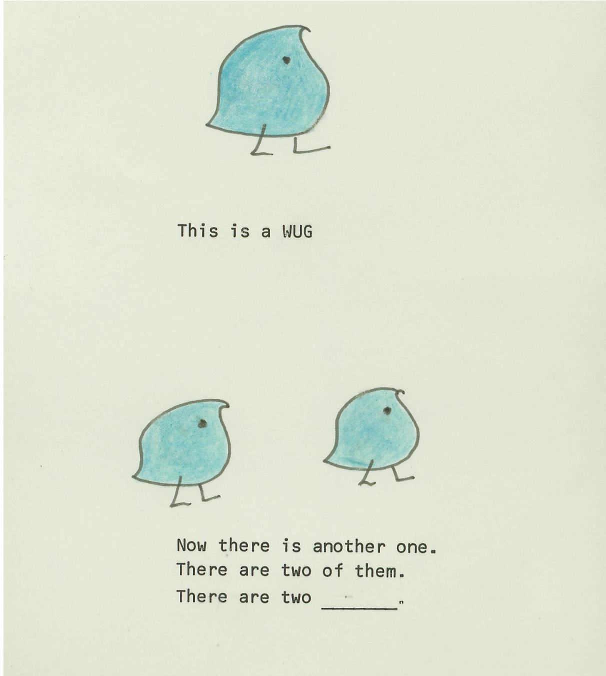
- “I am a good boy, amn’t I” (syntax)
- “toothes”; “breaked” (morphology)

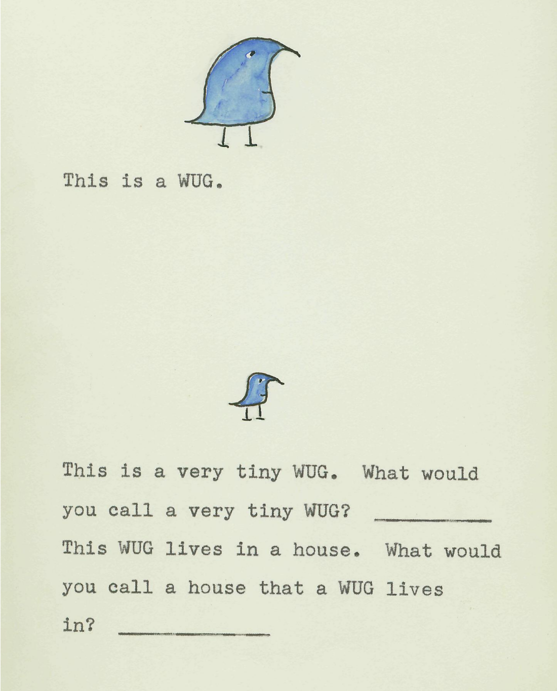
Do Children Know Syntactic Rules?

4-year old children can use novel verbs heard in one sentence structure in others. [Pinker *et al*,1987; Gropen *et al*, 1991]

“The pig is pilking the horse” → “The horse is being pilked by the pig”

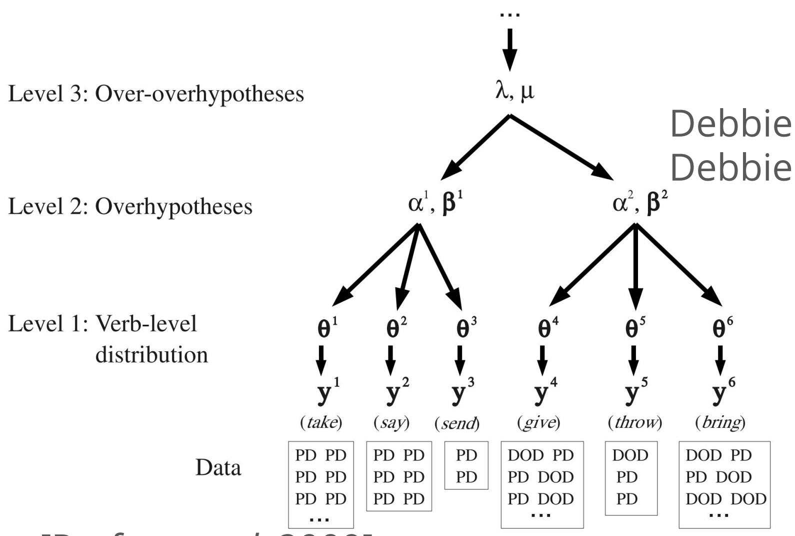
Do Children Know Morphological Rules? [Berko, 1958]





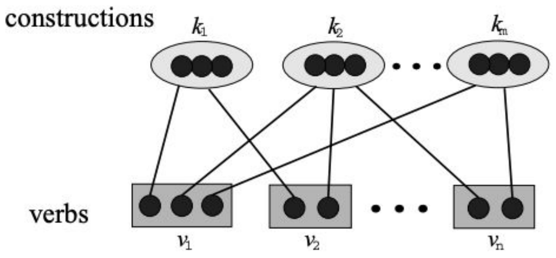
Modeling Structure

Learning abstractions through hierarchical representations. [Alishahi & Stevenson, 2008; Perfors *et al*, 2009; Barak *et al*, 2013]



[Perfors *et al*, 2009]

Debbie gave a pretzel to Dean (PD)
Debbie gave Dean a pretzel (DOD)



[Alishahi & Stevenson, 2008]

Generalization to Test Linguistic Knowledge

Children's knowledge of language is examined by generalization tasks:

- Mapping novel words to new/familiar objects.
- Using a new verb in "unheard" structures.
- Applying morphological rules to new words.

Can AI models pass these generalization tasks?

Nature of Nature

Thanks!

Abstract knowledge (priors/inductive biases/constraints) guides our generalization.

What are the origins of our abstract knowledge?
Can it be learned from experience?

Language learning and processing in people and machines

Part II: Human language processing

Aida Nematzadeh, Richard Futrell, and Roger Levy

Goals of part II of tutorial

- Overview of human language processing
  - Theoretically deep questions about language and mind
  - Helps establish long-term benchmarks for human-like AI systems for language
- Main points:
  - How we can study human language processing
  - First-cut theory
  - Limitations for first-cut theory:
    - Memory considerations
    - Character of input representations
  - More advanced theory
  - Open frontiers

Structure and surprise

Structure and surprise

The

Structure and surprise

The woman

Structure and surprise

The woman brought

Structure and surprise

The woman brought the

Structure and surprise

The woman brought the sandwich

Structure and surprise

The woman brought the sandwich from

Structure and surprise

The woman brought the sandwich from the

Structure and surprise

The woman brought the sandwich from the kitchen

Structure and surprise

The woman brought the sandwich from the kitchen tripped.

Structure and surprise

Structure and surprise

The woman who was given the sandwich from the kitchen tripped.

Structure and surprise

The woman given the sandwich from the kitchen tripped.

who was

Structure and surprise

The woman given the sandwich from the kitchen tripped.

The woman given the sandwich from the kitchen tripped.

who was

Structure and surprise

The woman brought the sandwich from the kitchen tripped.

who was

The woman given the sandwich from the kitchen tripped.

The woman given the sandwich from the kitchen tripped.

who was

Structure and surprise

The woman brought the sandwich from the kitchen tripped.

The woman brought the sandwich from the kitchen tripped.

who was

The woman given the sandwich from the kitchen tripped.

The woman given the sandwich from the kitchen tripped.

who was

Structure and surprise

The woman brought the sandwich from the kitchen tripped.

The woman brought the sandwich from the kitchen tripped.

who was

The woman given the sandwich from the kitchen tripped.

The woman given the sandwich from the kitchen tripped.

who was

	Simple past	Past participle
bring	***brought***	***brought***
give	***gave***	***given***

Structure and surprise

The woman brought the sandwich from the kitchen tripped.

The woman brought the sandwich from the kitchen tripped.

who was

The woman given the sandwich from the kitchen tripped.

The woman given the sandwich from the kitchen tripped.

who was

	Simple past	Past participle
bring	***brought***	***brought***
give	***gave***	***given***

Meaning can help us avoid surprise, too:

The evidence examined by the lawyer from the firm was unreliable.

Anatomy of ye olde garden path sentence

The woman brought the sandwich from the kitchen tripped.

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension

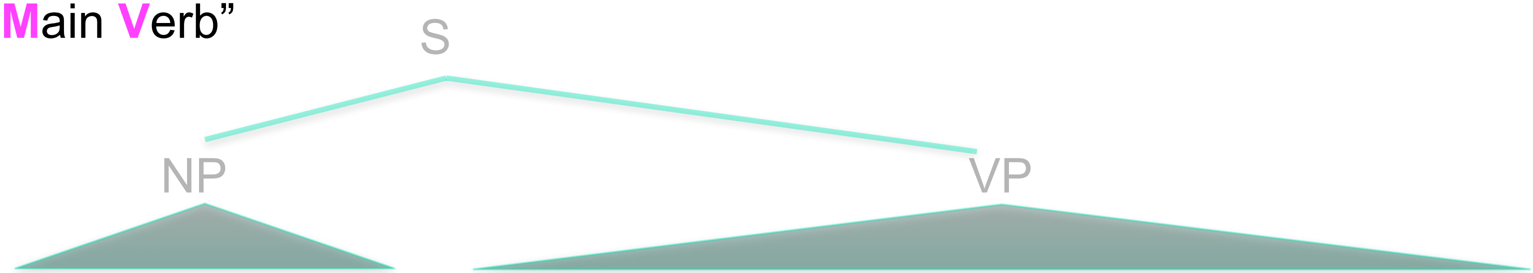
The woman brought the sandwich from the kitchen tripped.

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension

"Main Verb"



The woman brought the sandwich from the kitchen tripped.

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension

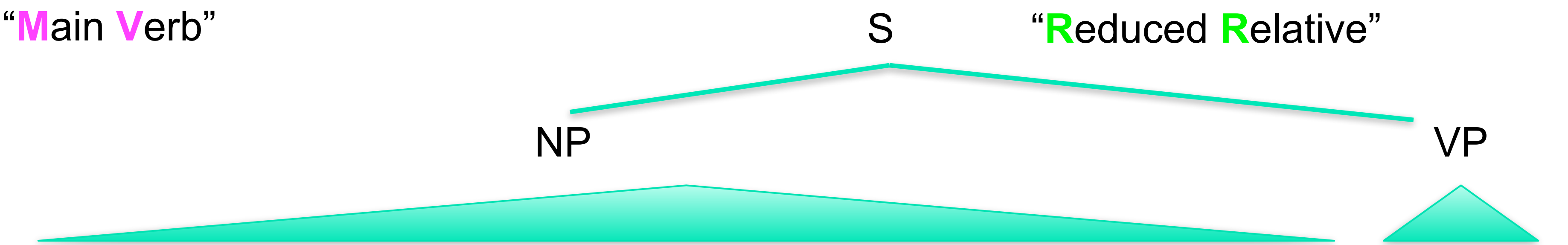
"**M**ain **V**erb"

The woman brought the sandwich from the kitchen tripped.

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension

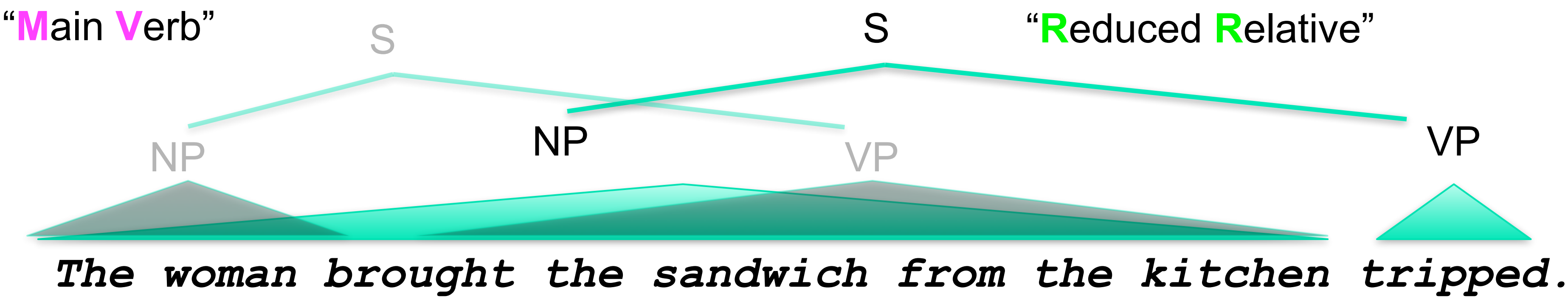


The woman brought the sandwich from the kitchen tripped.

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

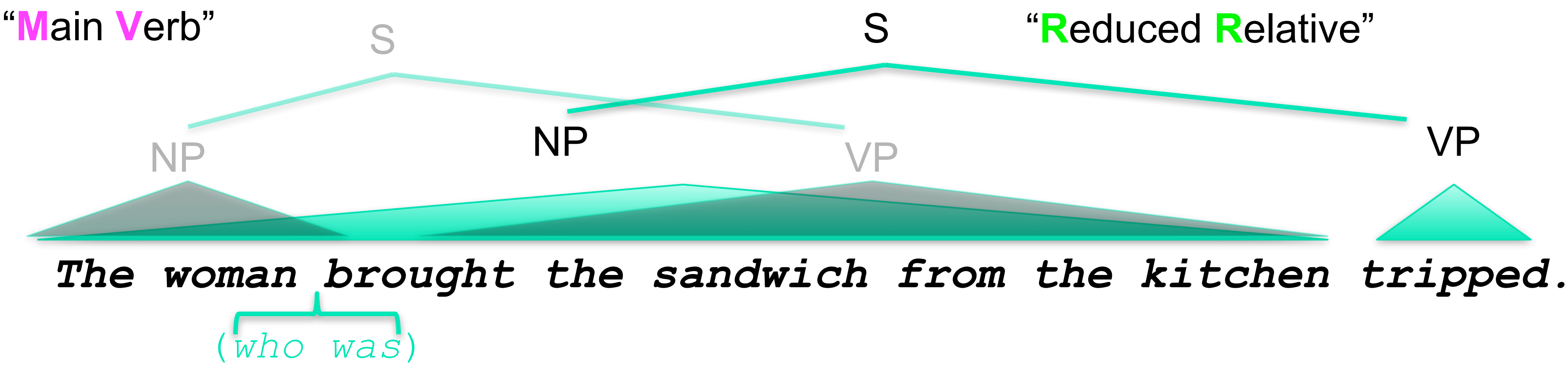
- Classic example of incrementality in comprehension



(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

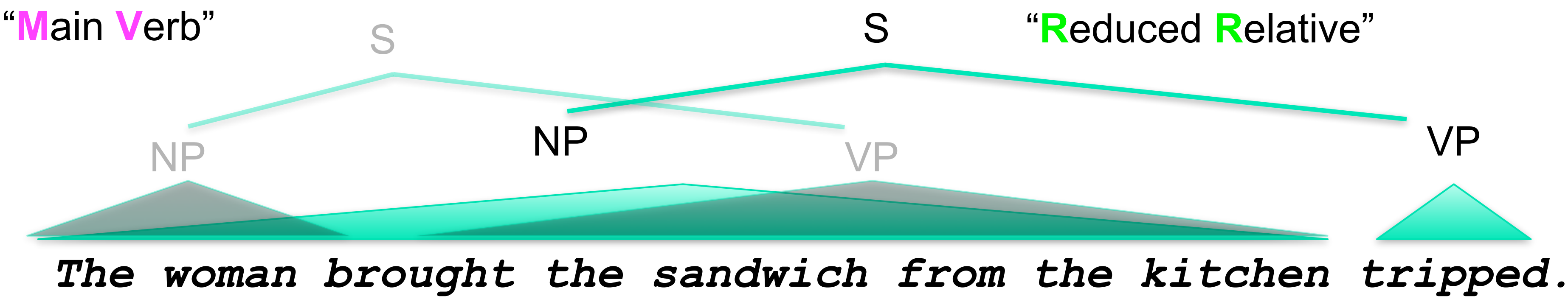
- Classic example of incrementality in comprehension



(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

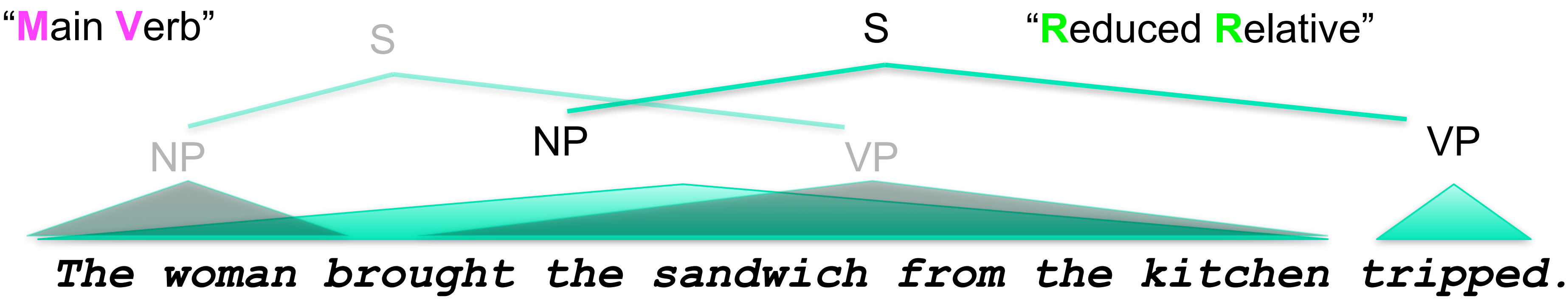
- Classic example of incrementality in comprehension



(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension

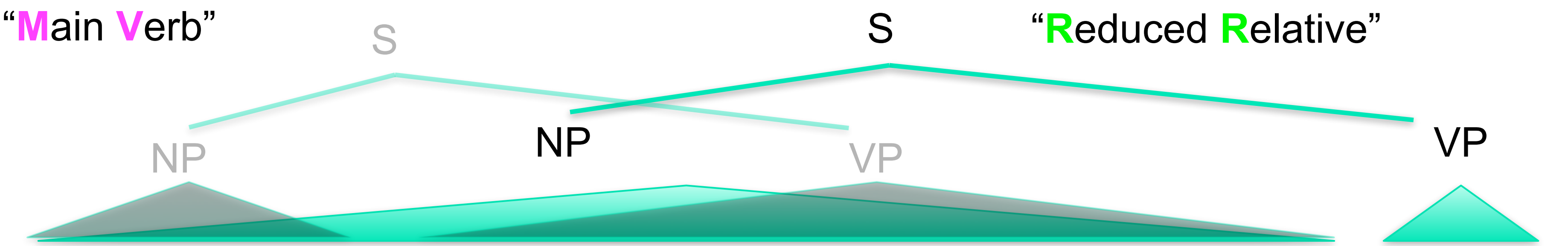


- People fail to understand it most of the time

(c.f. *The horse raced past the barn fell*; Bever, 1970)

Anatomy of ye olde garden path sentence

- Classic example of incrementality in comprehension



The woman brought the sandwich from the kitchen tripped.

- People fail to understand it most of the time
- People are likely to *misunderstand* it—e.g.,
  - The woman *who* brought the sandwich from the kitchen tripped
  - The woman brought the sandwich from the kitchen *and* tripped
  - "What's a kitchen tripped?"

(c.f. *The horse raced past the barn fell*; Bever, 1970)

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)

Measuring human incremental processing state

- Eye movements in the visual world
- Word-by-word reading times
  - Self-paced reading
  - Eye movements during natural reading

Behavioral

- Recordings of brain activity
  - Electrophysiological (EEG/ERP)
  - Magneto-encephalography (MEG)
  - functional Magnetic Resonance Imaging (fMRI)
  - Electrocorticography (ECoG)

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

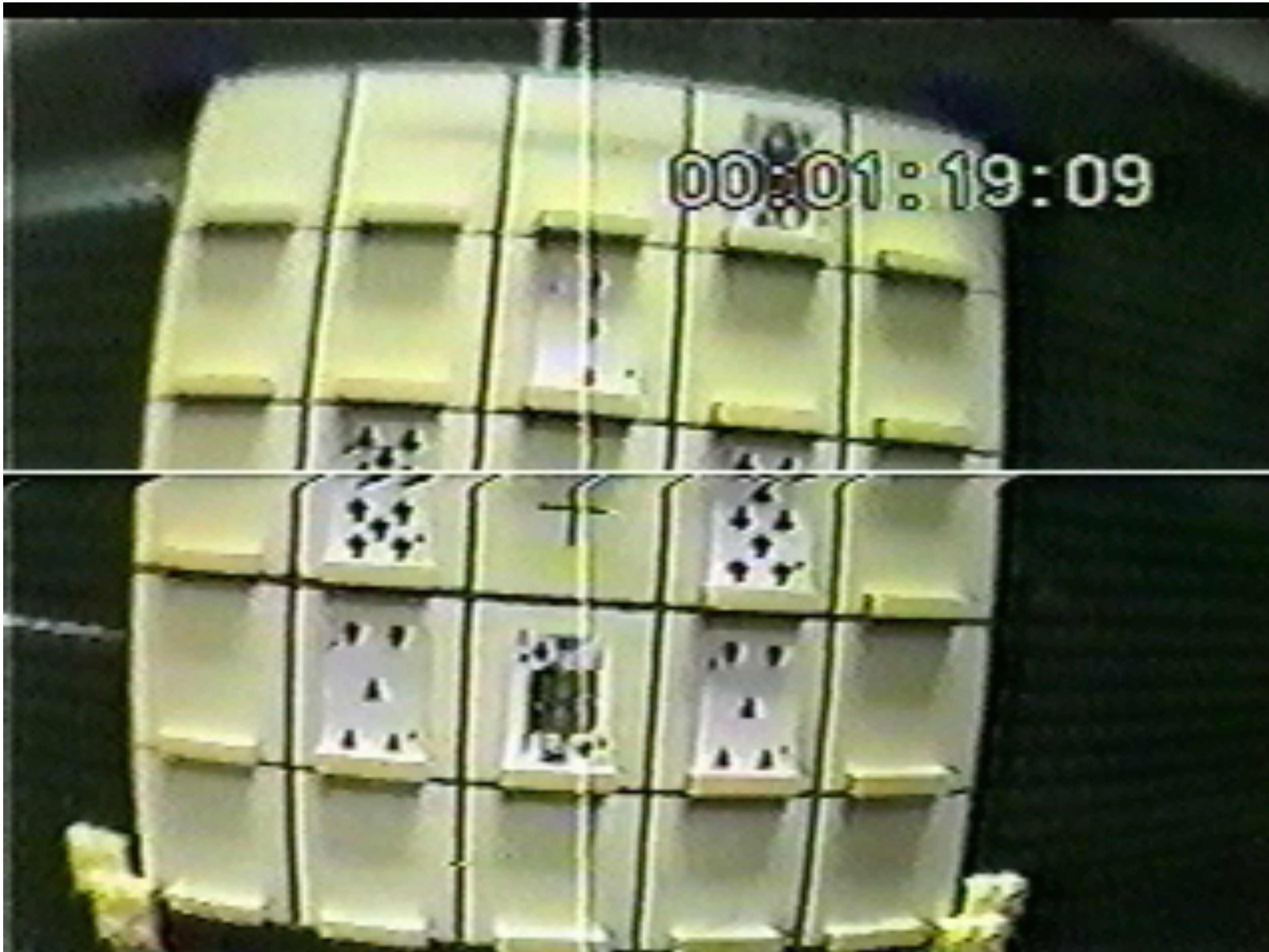
Eye movements in the visual world

(*Video courtesy of Mike Tanenhaus*)

Eye movements in the visual world

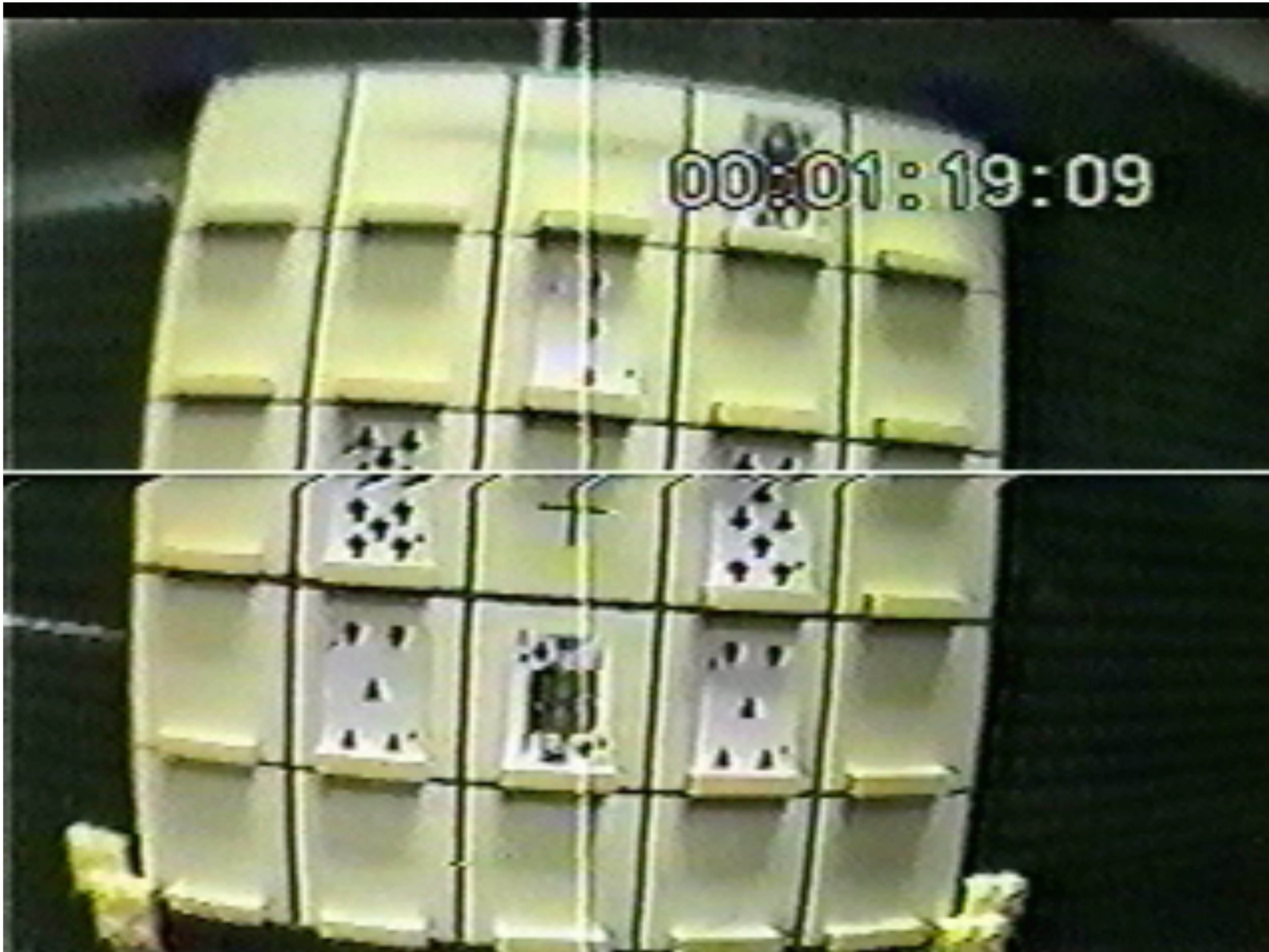
(Video courtesy of Mike Tanenhaus)

Eye movements in the visual world



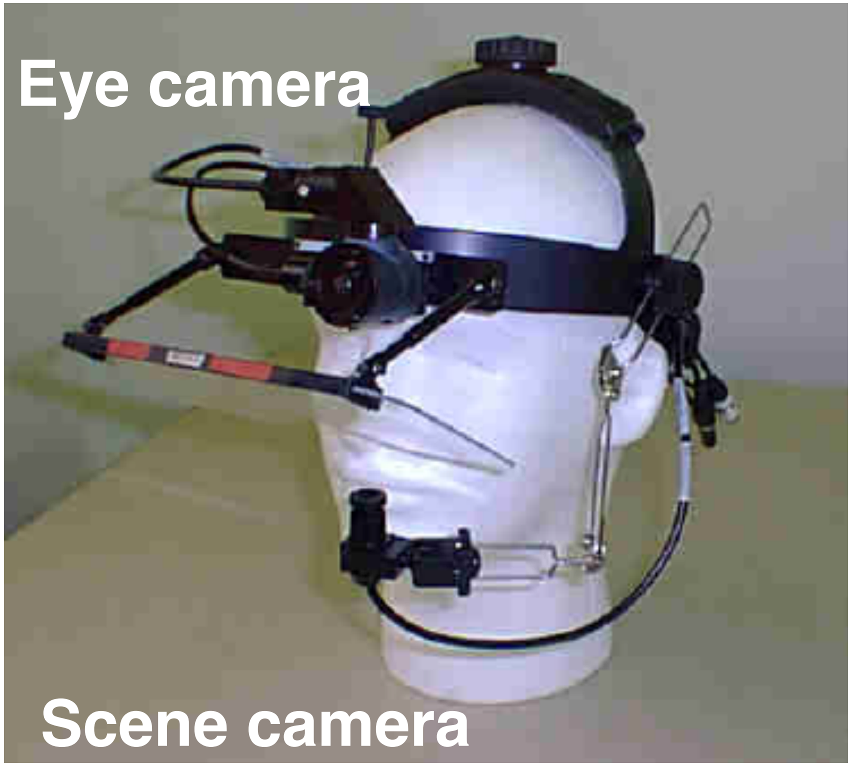
(Video courtesy of Mike Tanenhaus)

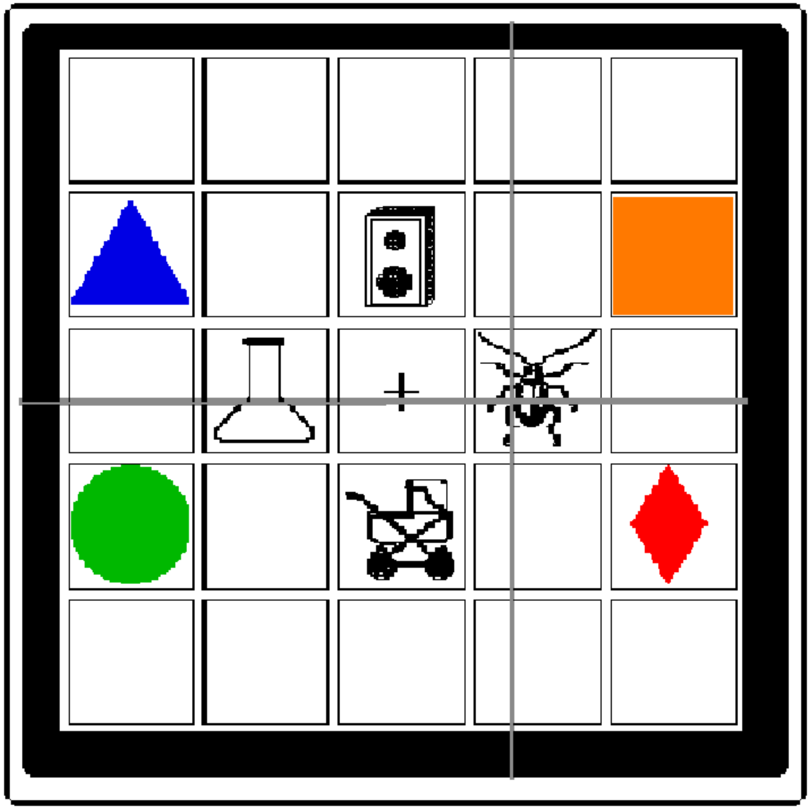
Eye movements in the visual world



(Video courtesy of Mike Tanenhaus)

A visual world experiment

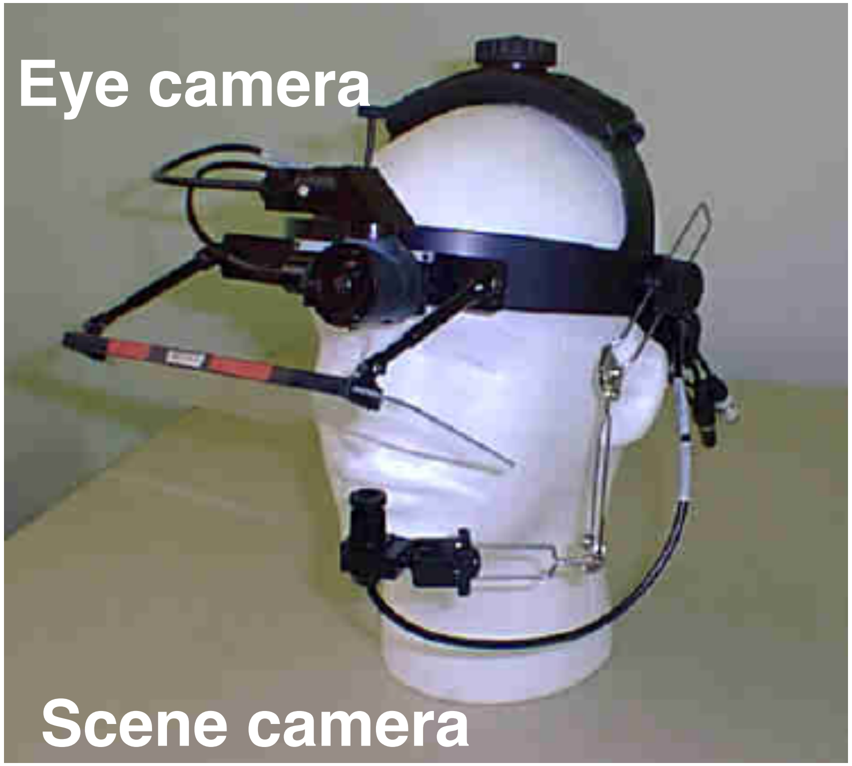


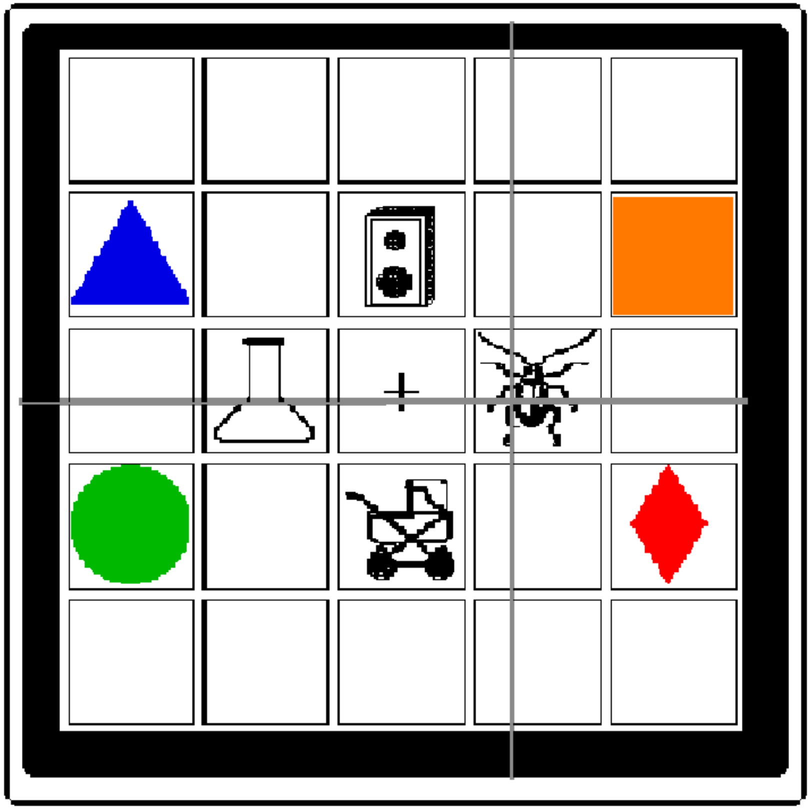


(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)

A visual world experiment



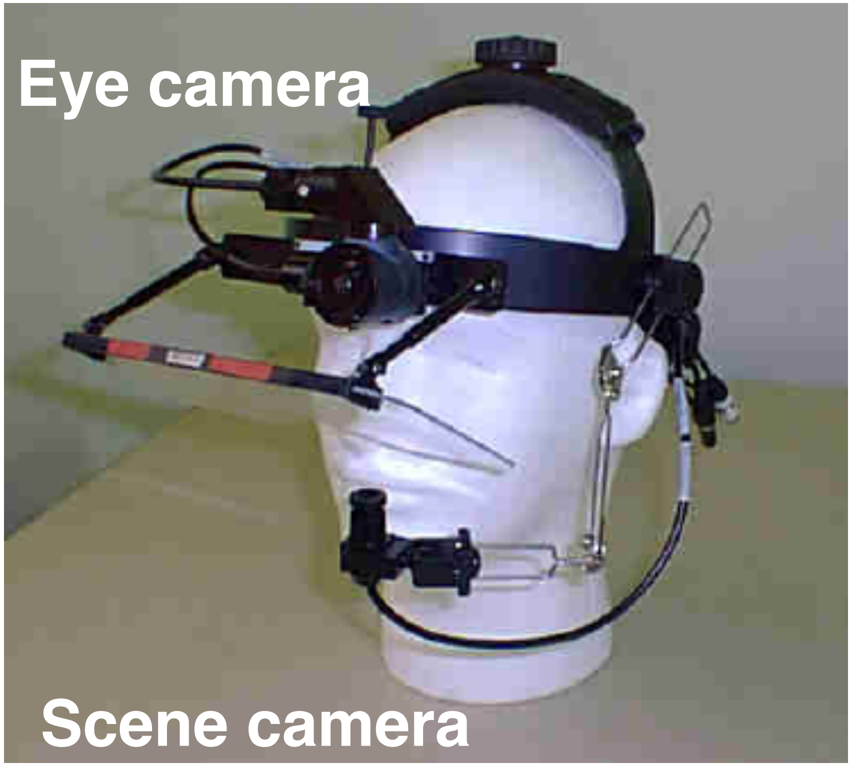


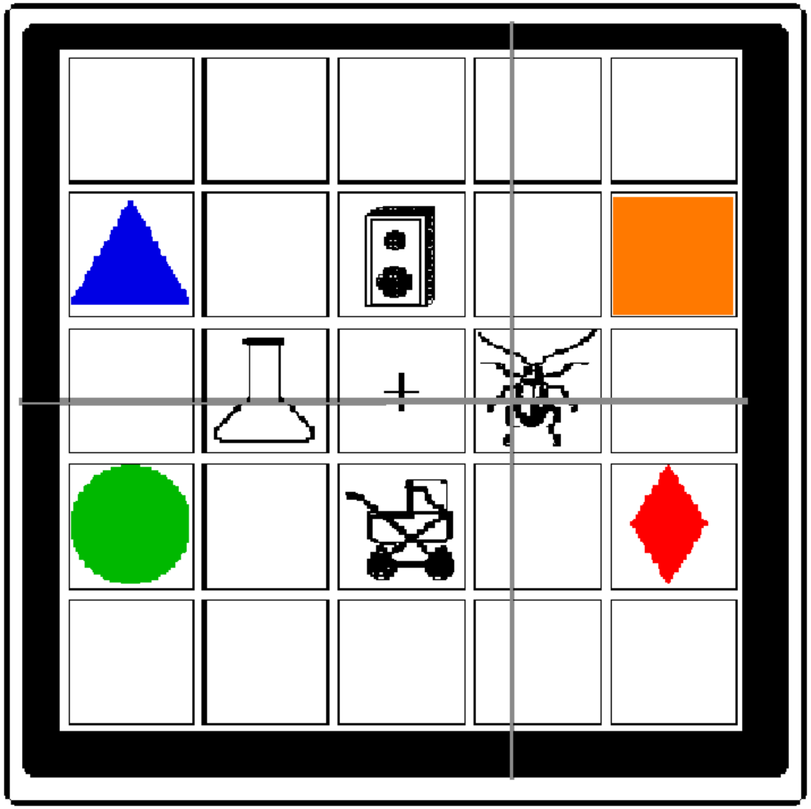
Instruction to experimental participant:

(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)

A visual world experiment





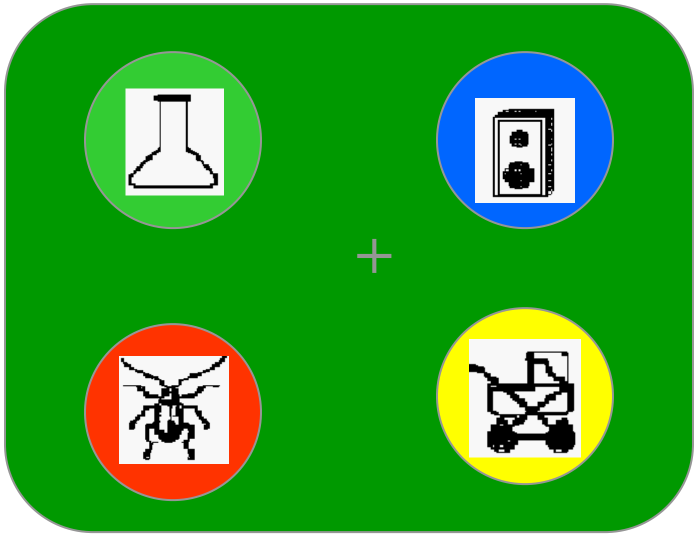
Instruction to experimental participant:

"Pick up the beaker"

(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)

Data from human eye movements



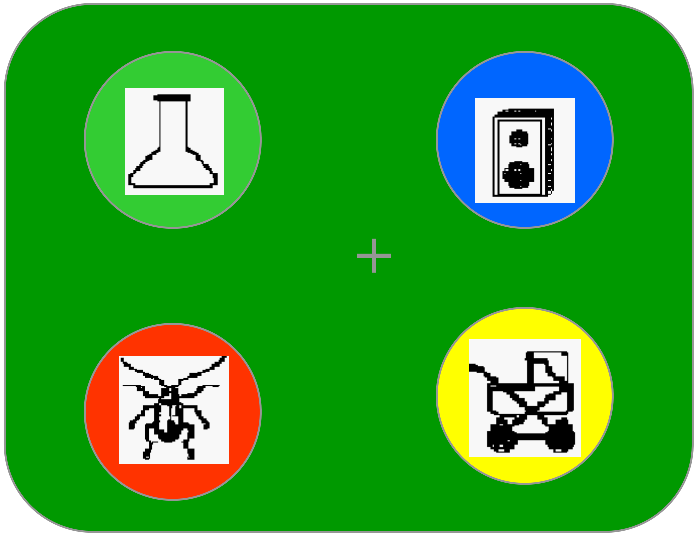
Target = beaker

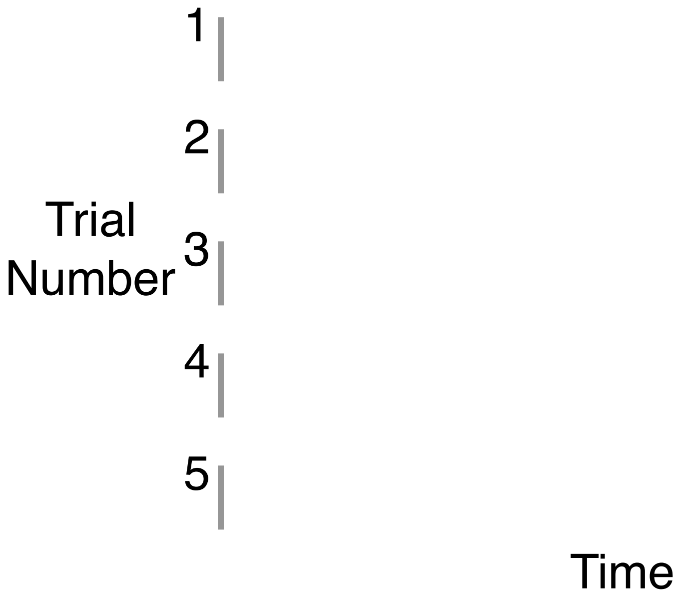
Cohort = beetle

Unrelated = carriage

(Slide courtesy of Mike Tanenhaus)

Data from human eye movements





Target = beaker

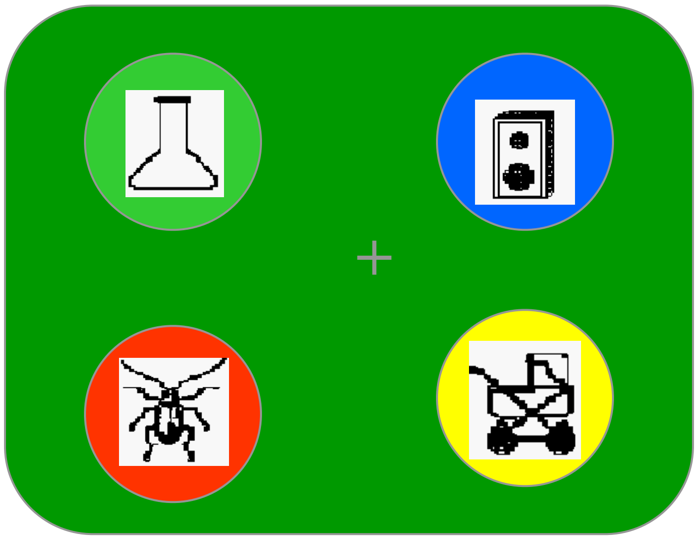
Cohort = beetle

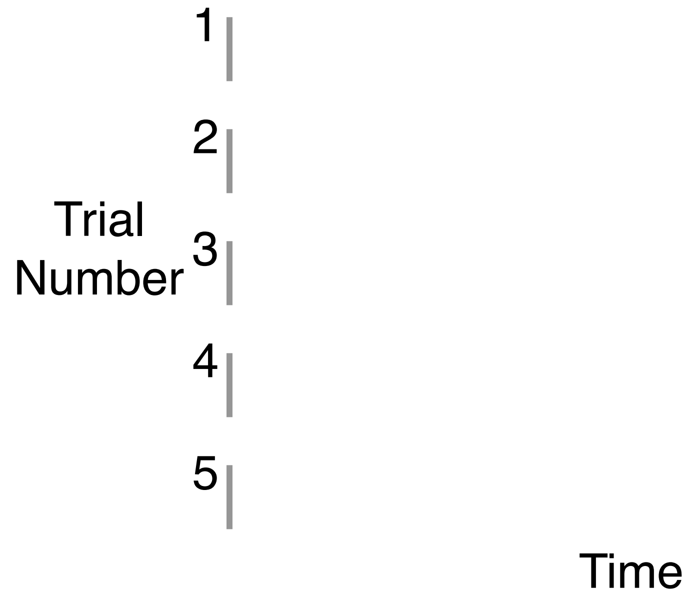
Unrelated = carriage

(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."





Target = beaker

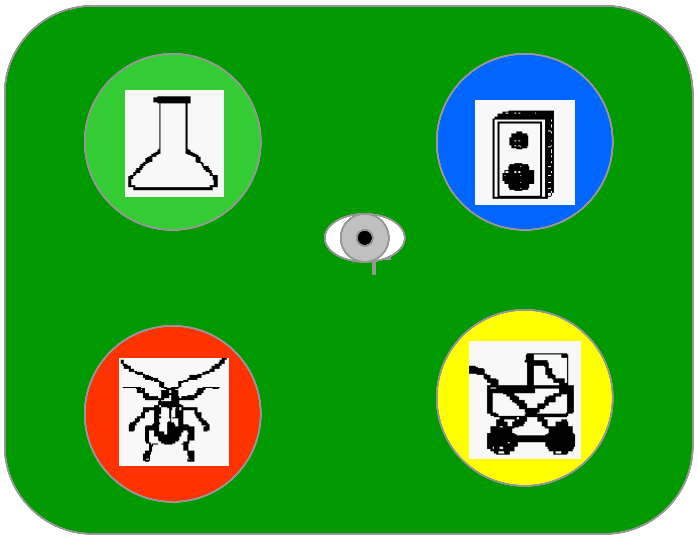
Cohort = beetle

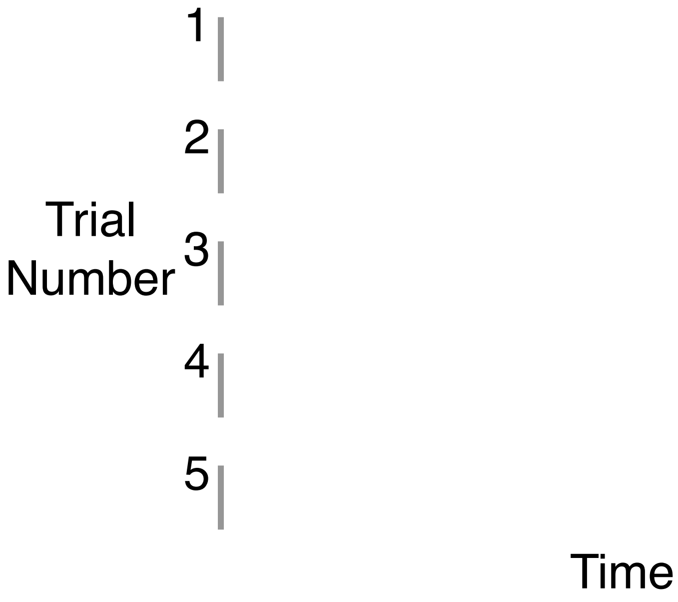
Unrelated = carriage

(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."





Target = beaker

Cohort = beetle

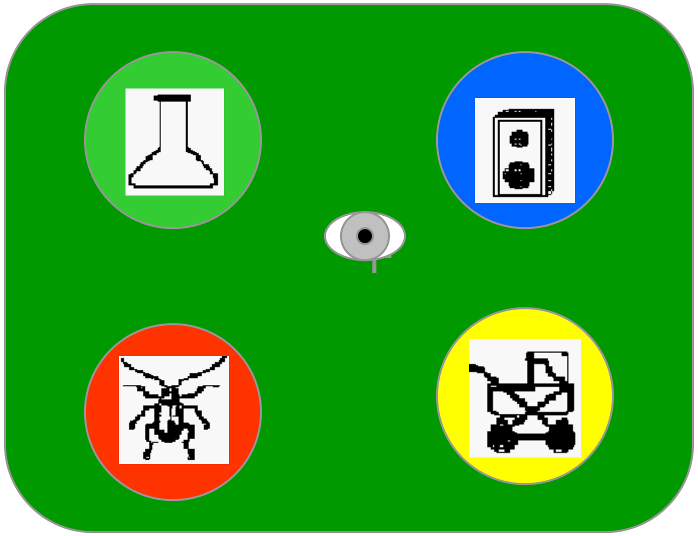
Unrelated = carriage

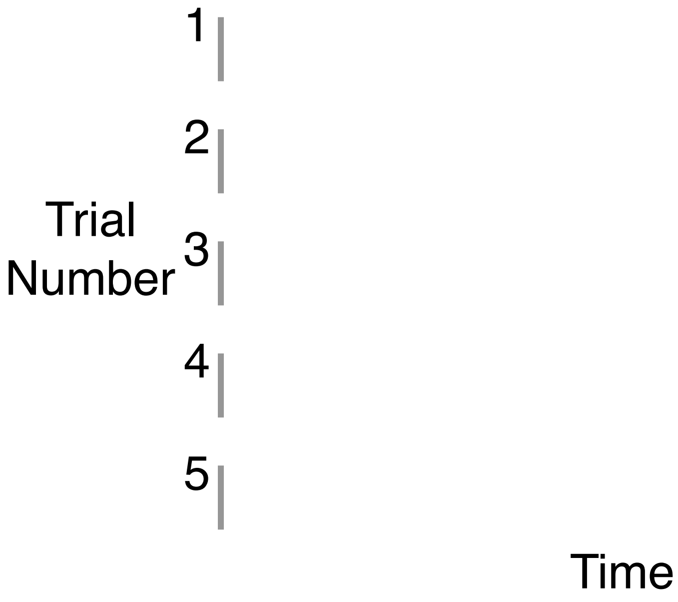
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

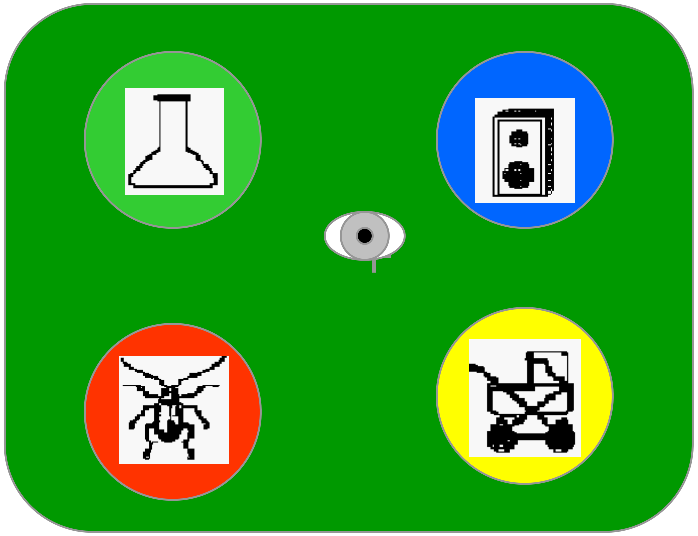
Unrelated = carriage

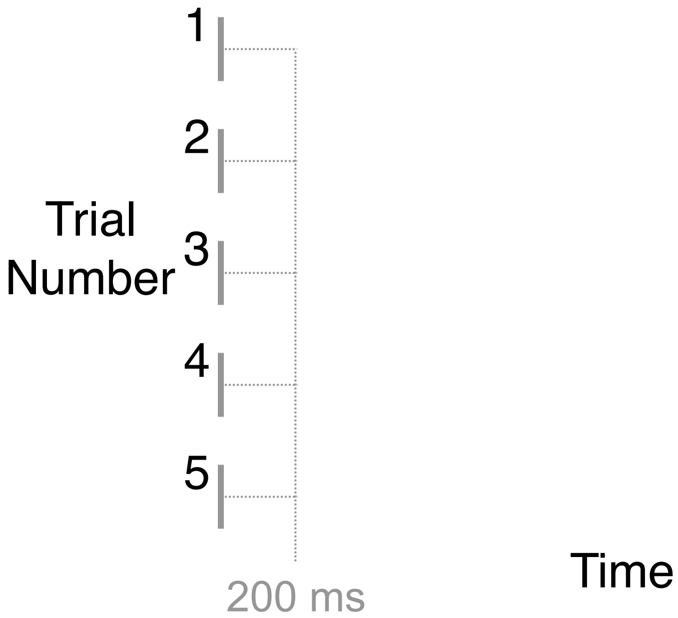
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

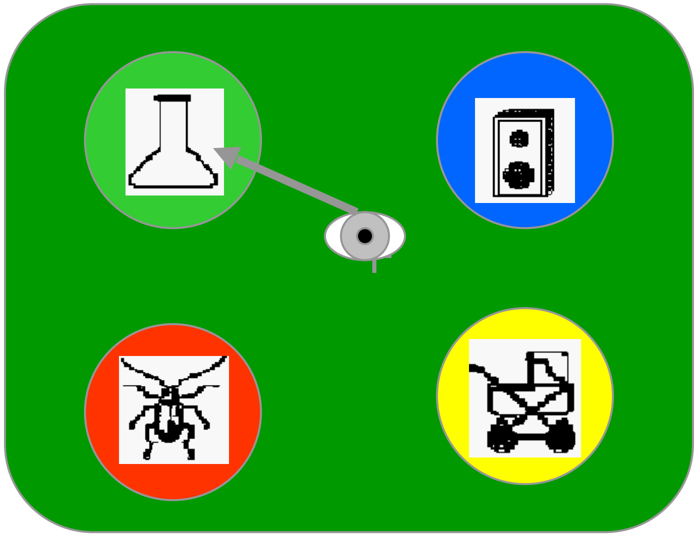
Unrelated = carriage

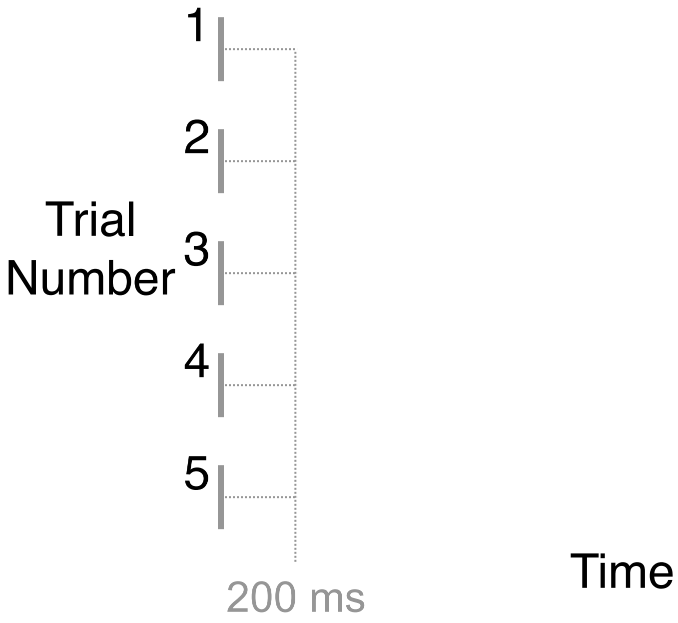
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

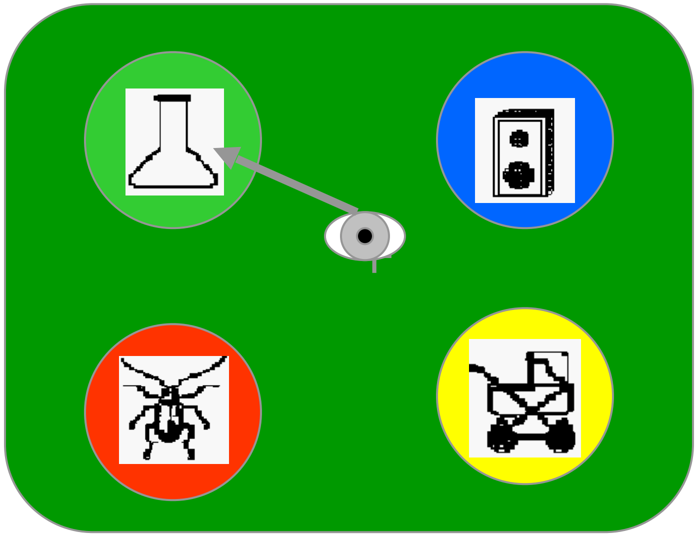
Unrelated = carriage

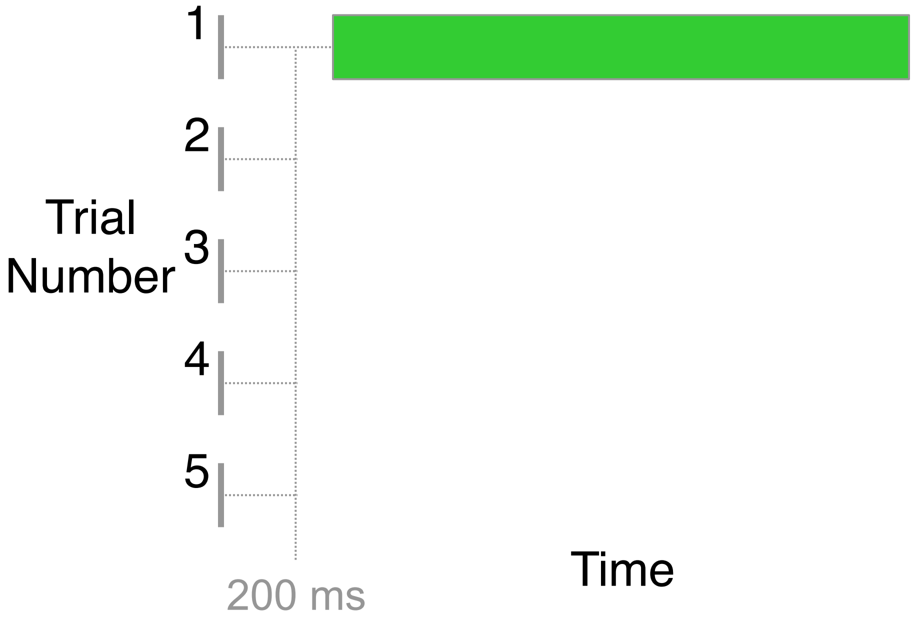
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

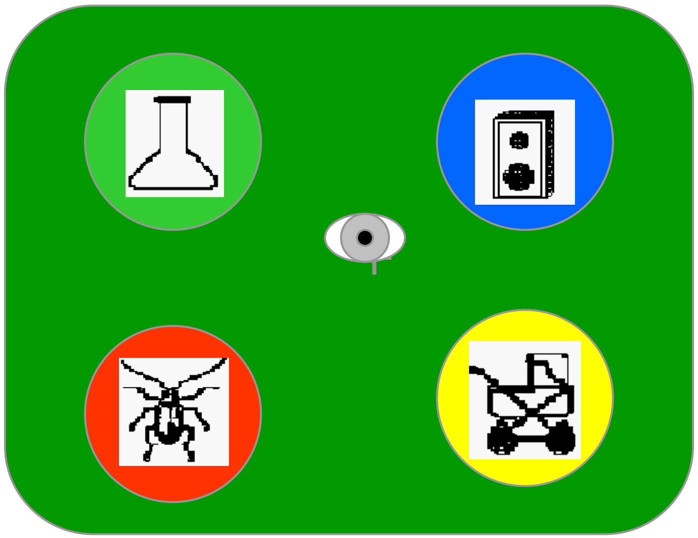
Unrelated = carriage

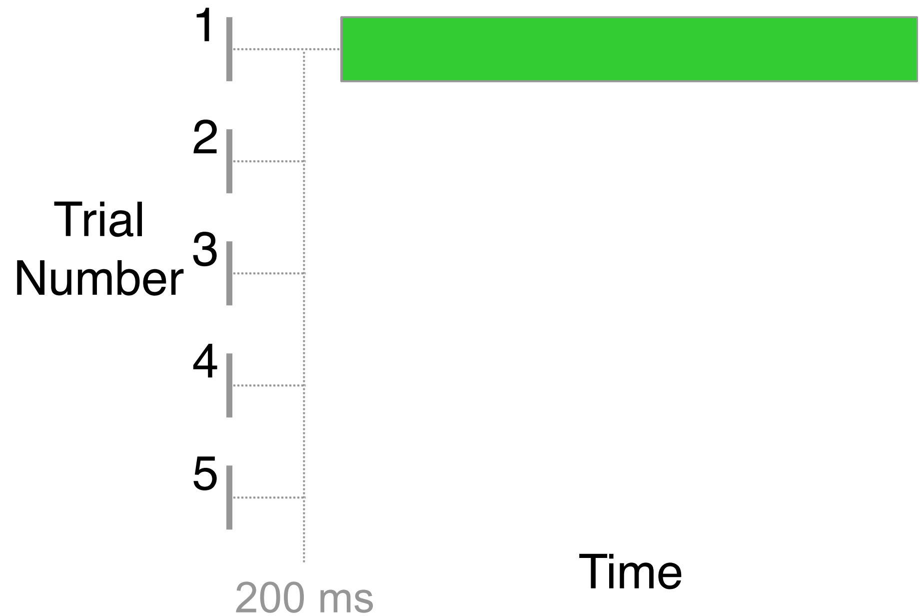
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

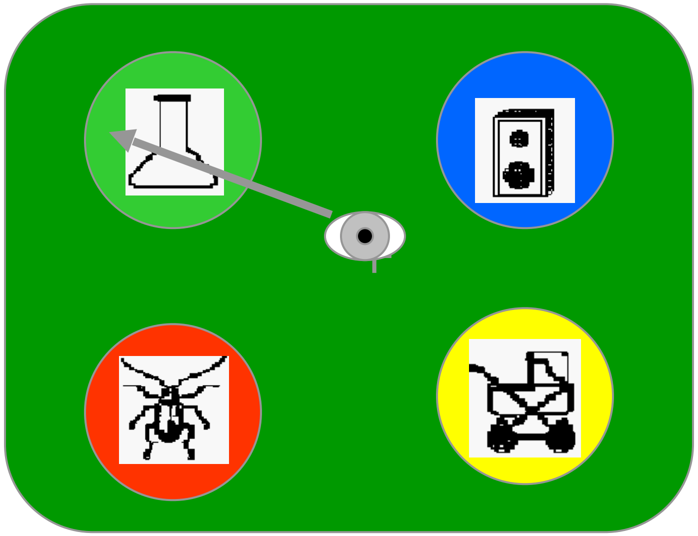
Unrelated = carriage

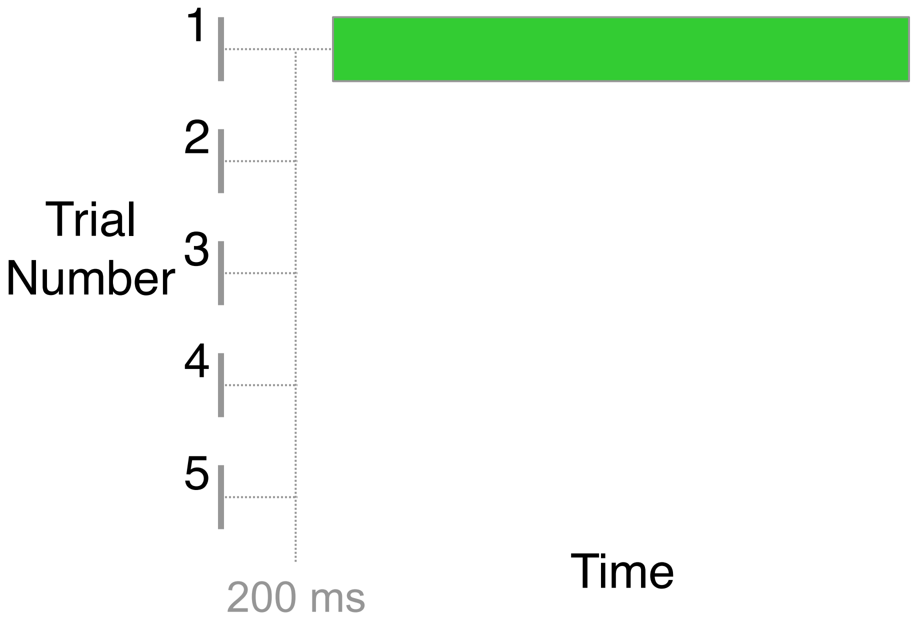
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

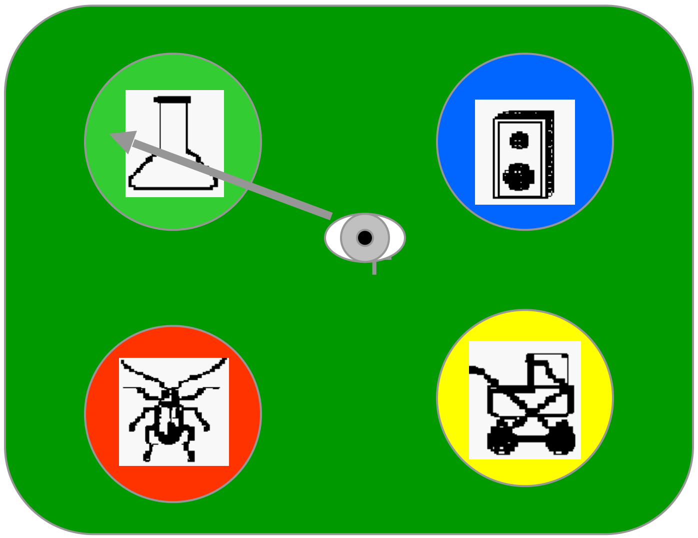
Unrelated = carriage

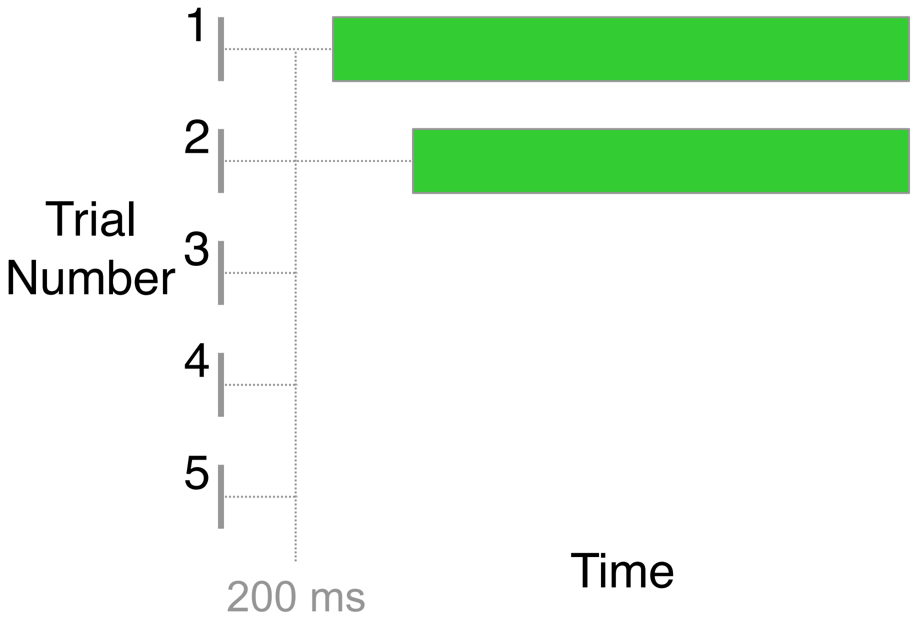
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

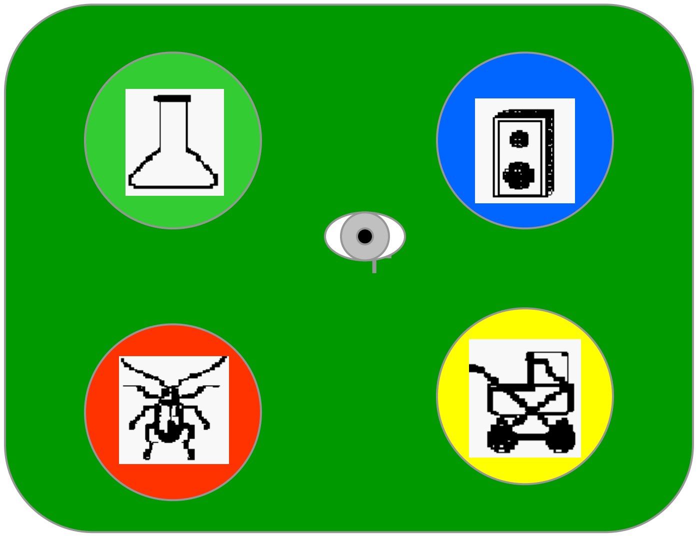
Unrelated = carriage

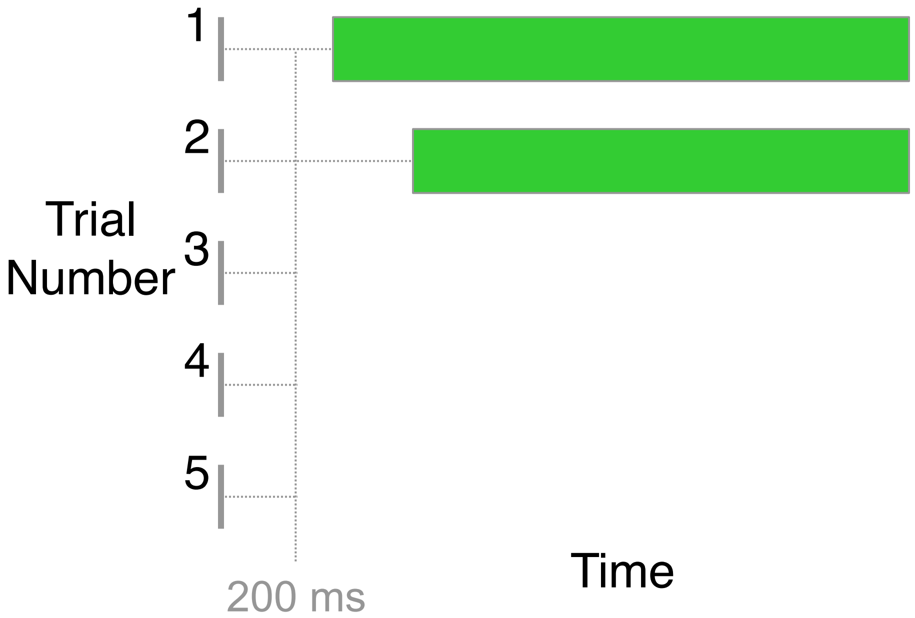
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

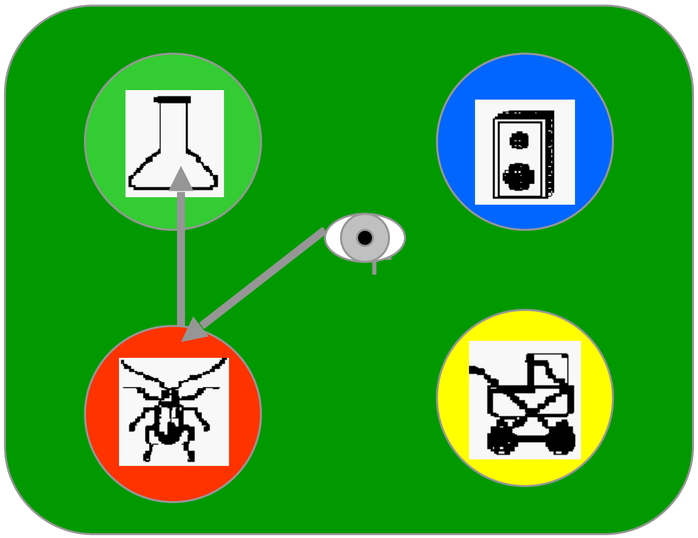
Unrelated = carriage

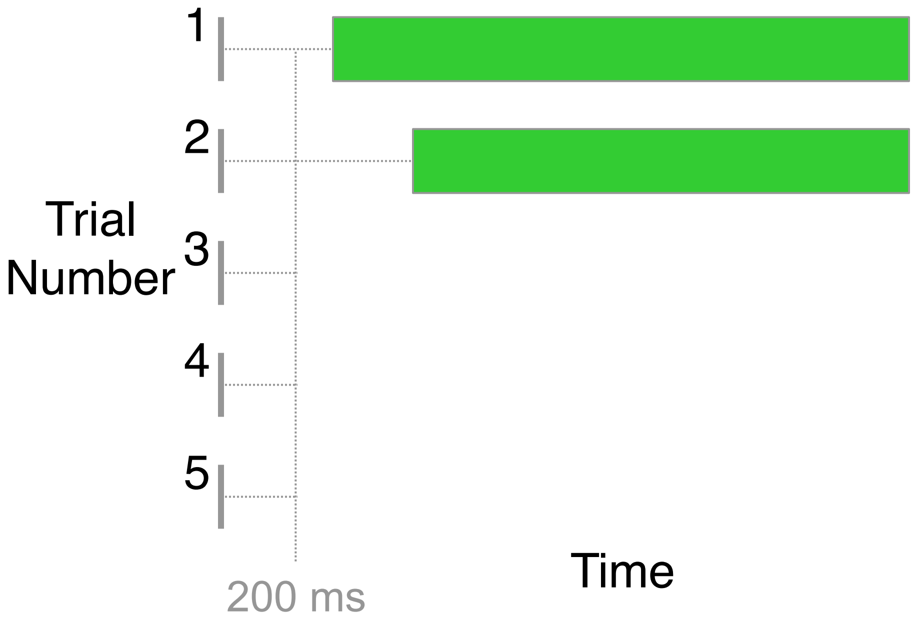
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

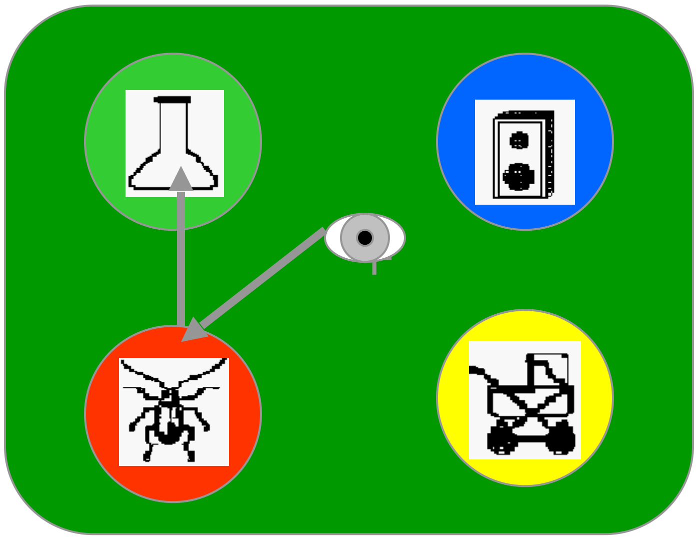
Unrelated = carriage

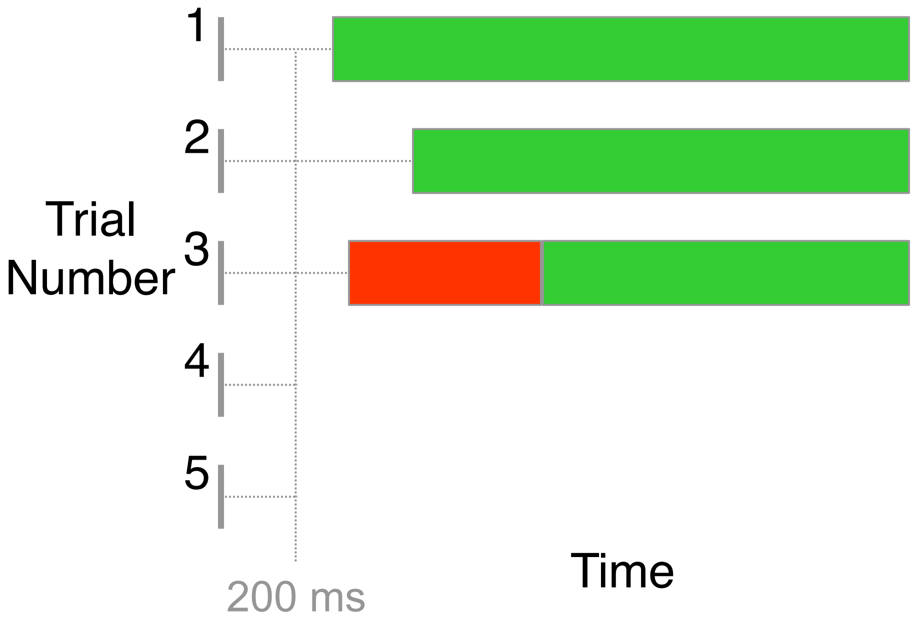
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

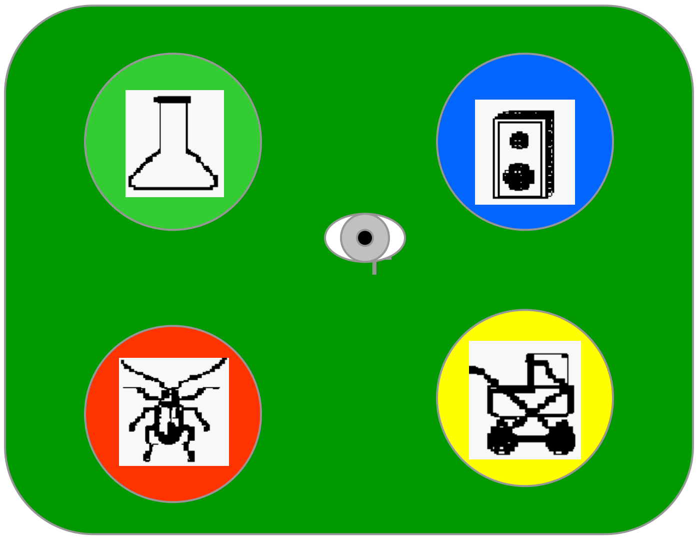
Unrelated = carriage

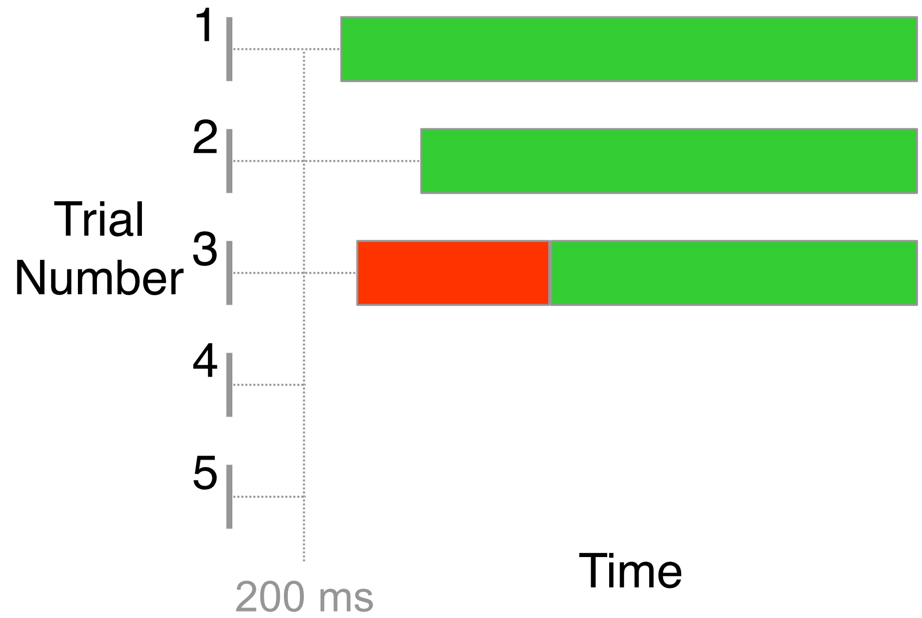
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

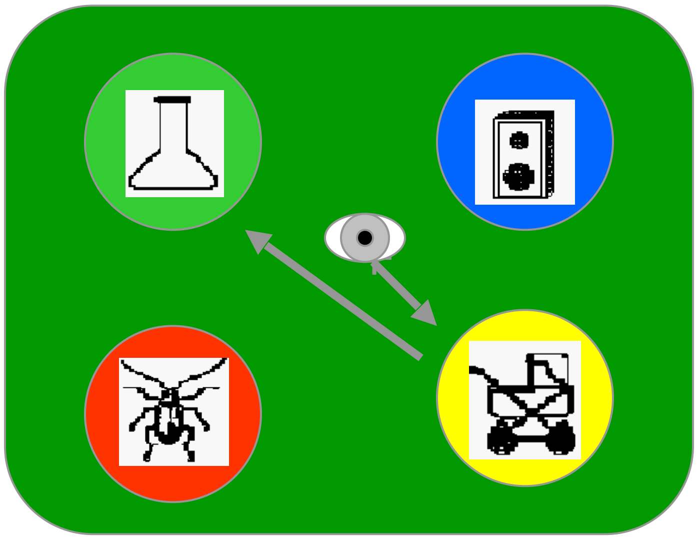
Unrelated = carriage

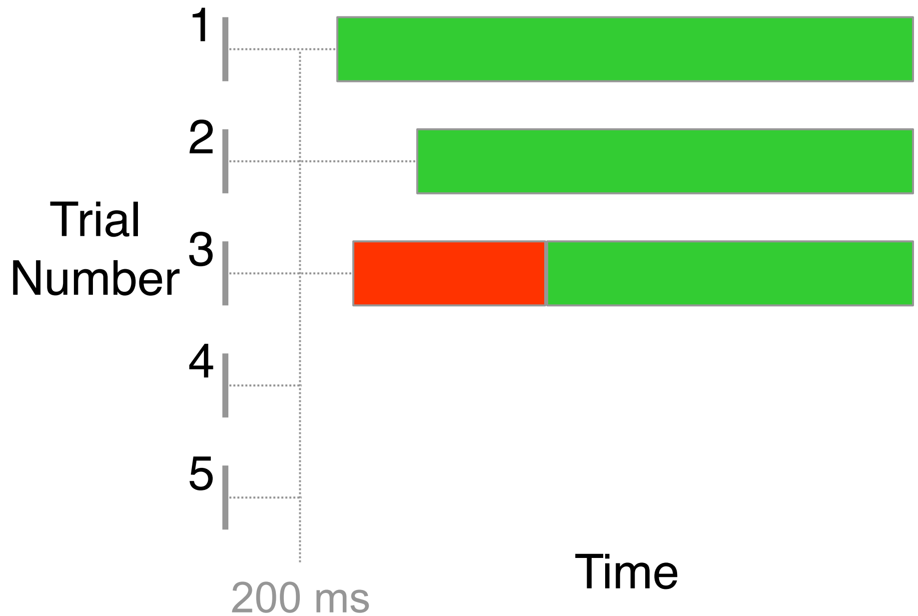
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

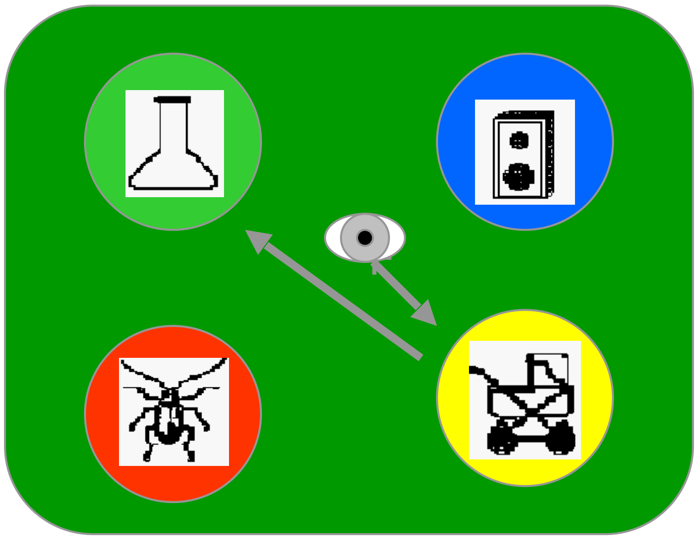
Unrelated = carriage

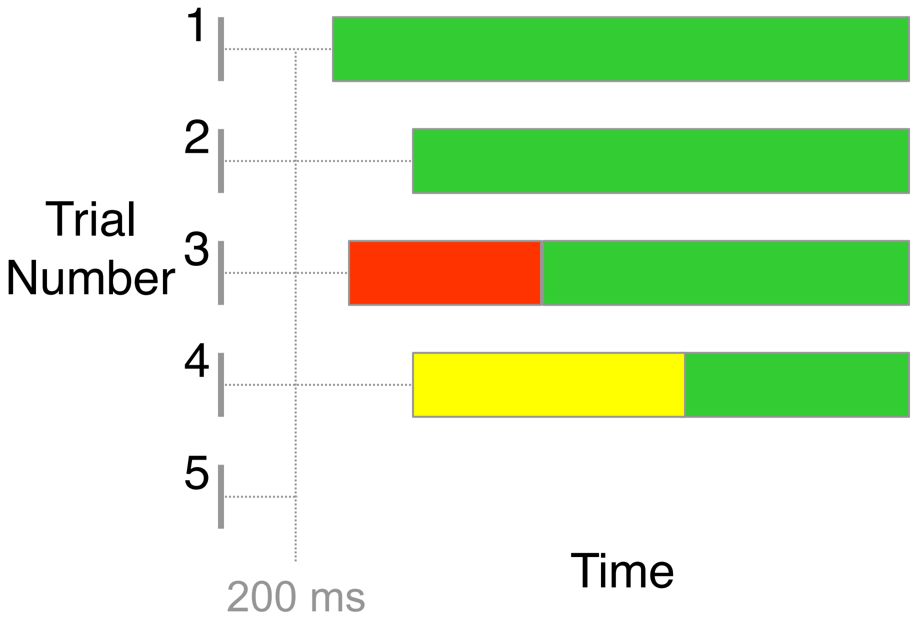
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

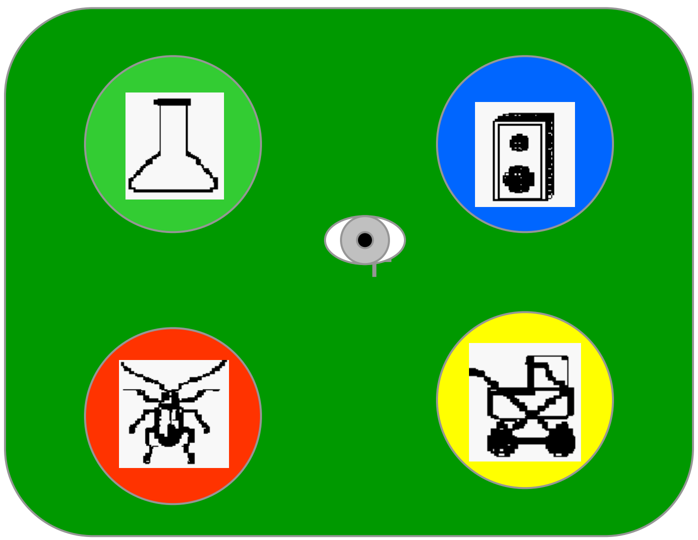
Unrelated = carriage

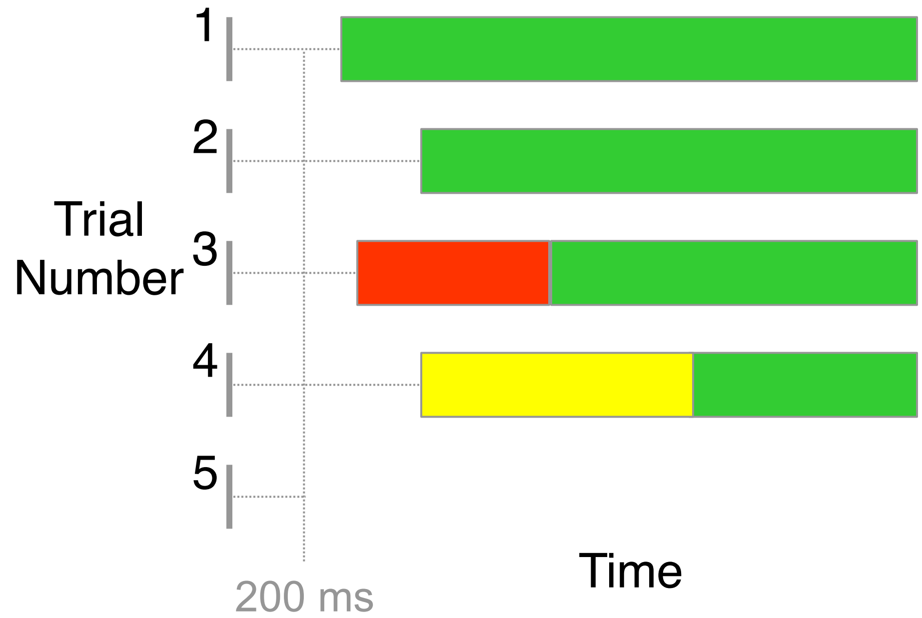
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

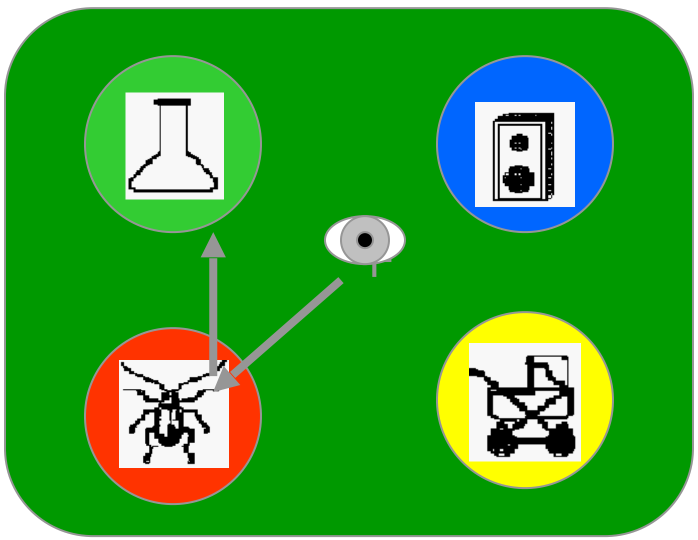
Unrelated = carriage

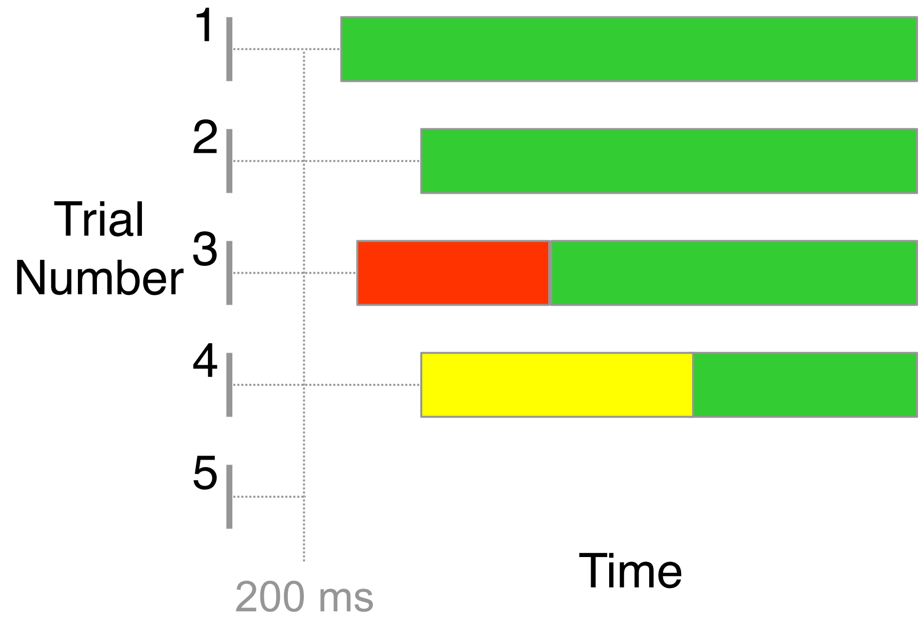
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

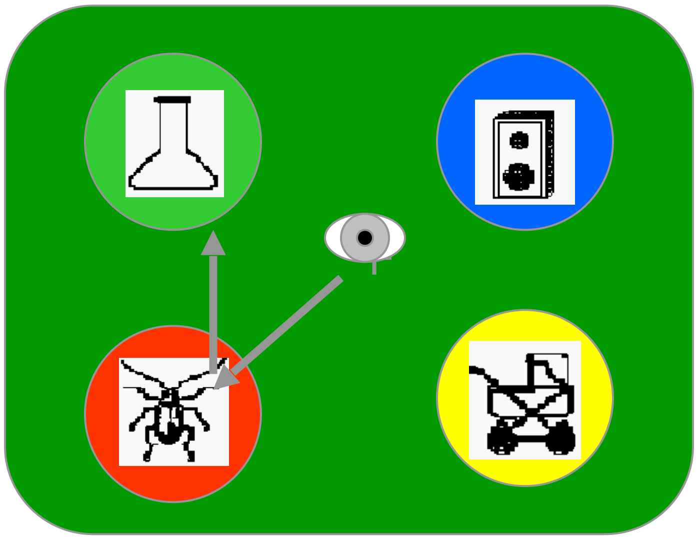
Unrelated = carriage

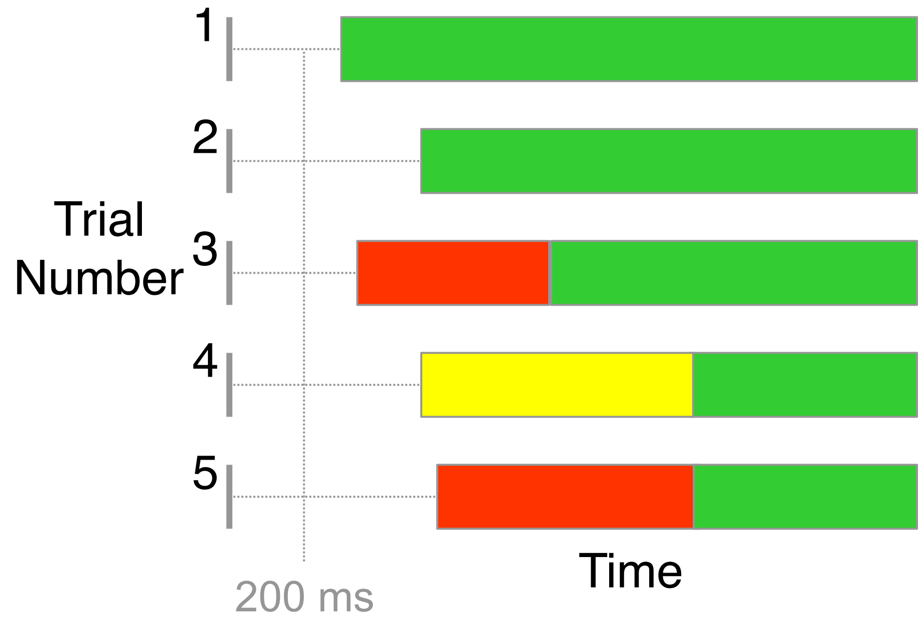
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

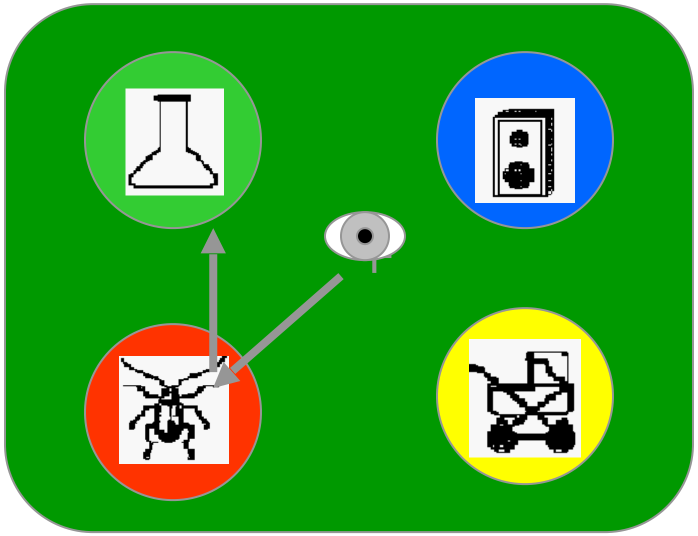
Unrelated = carriage

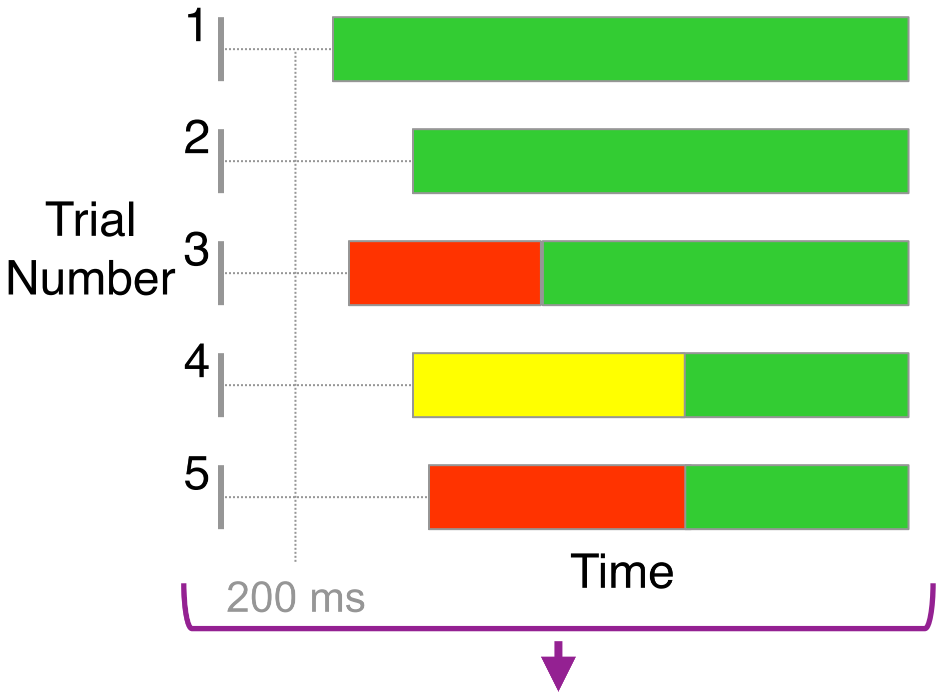
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

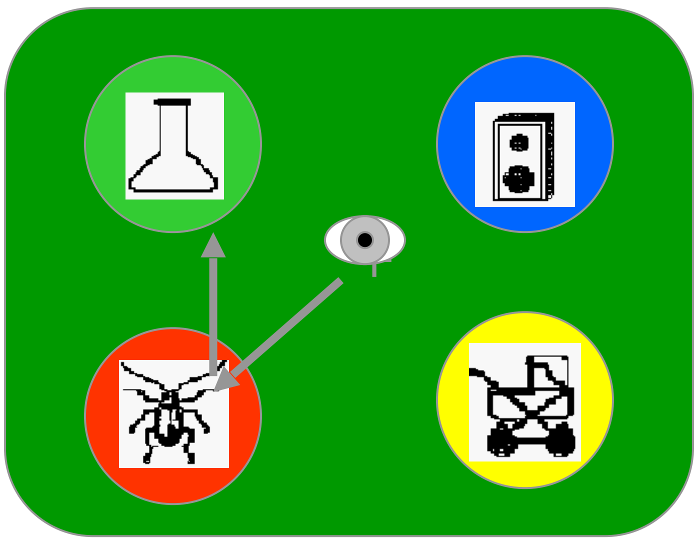
Unrelated = carriage

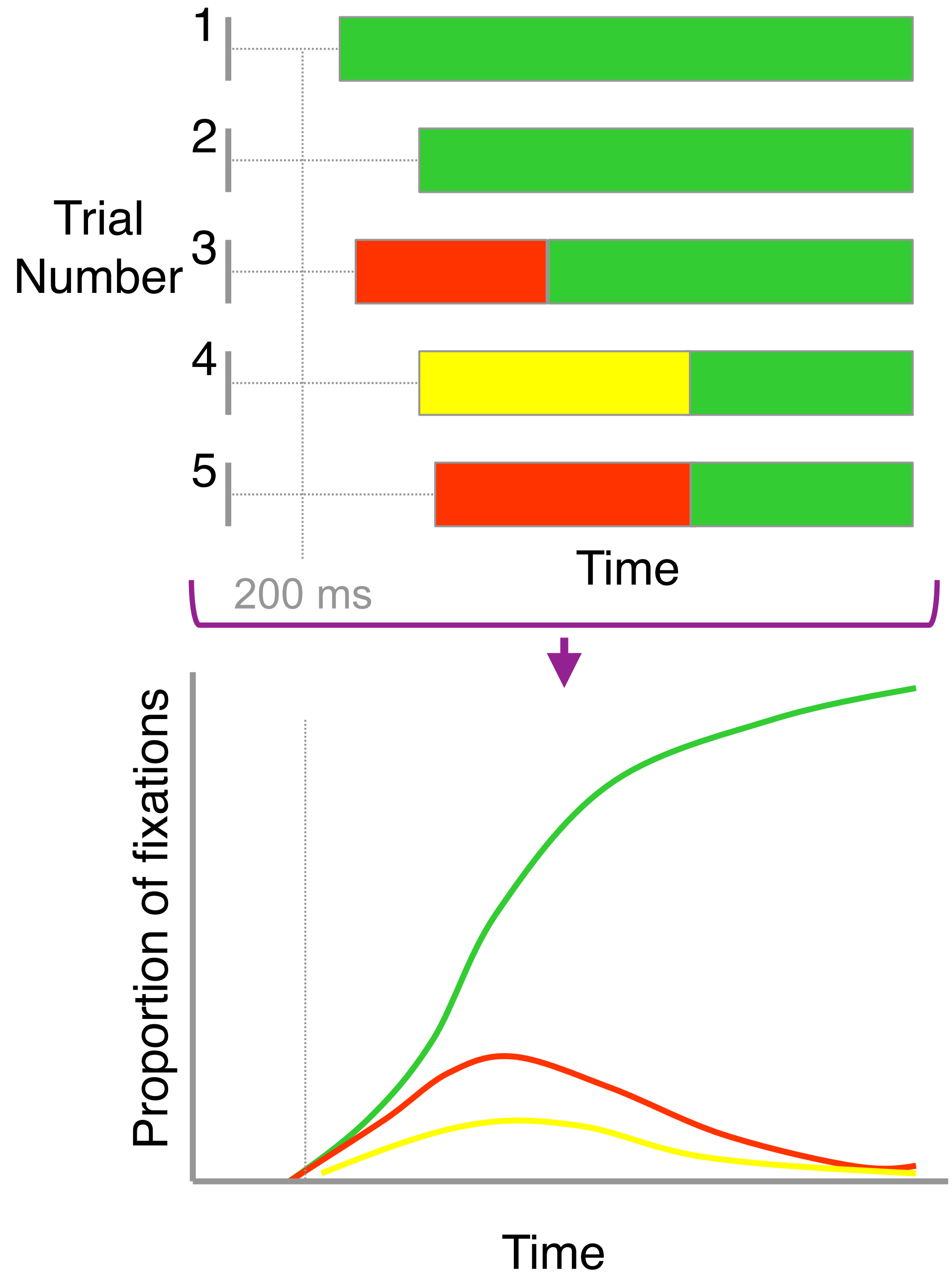
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

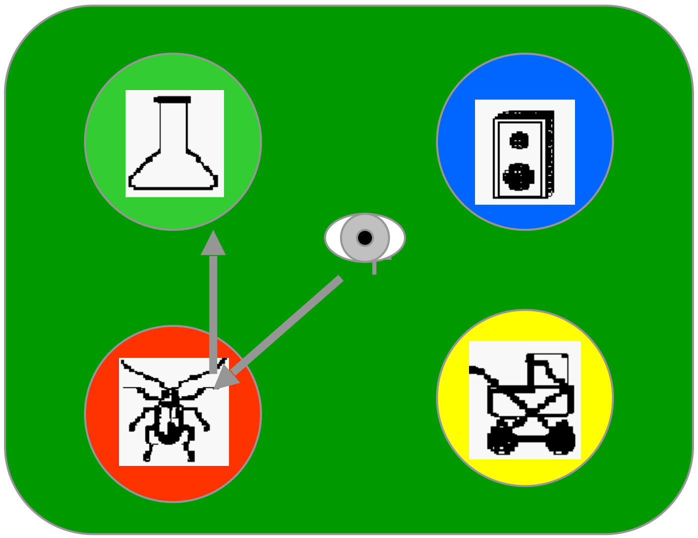
Unrelated = carriage

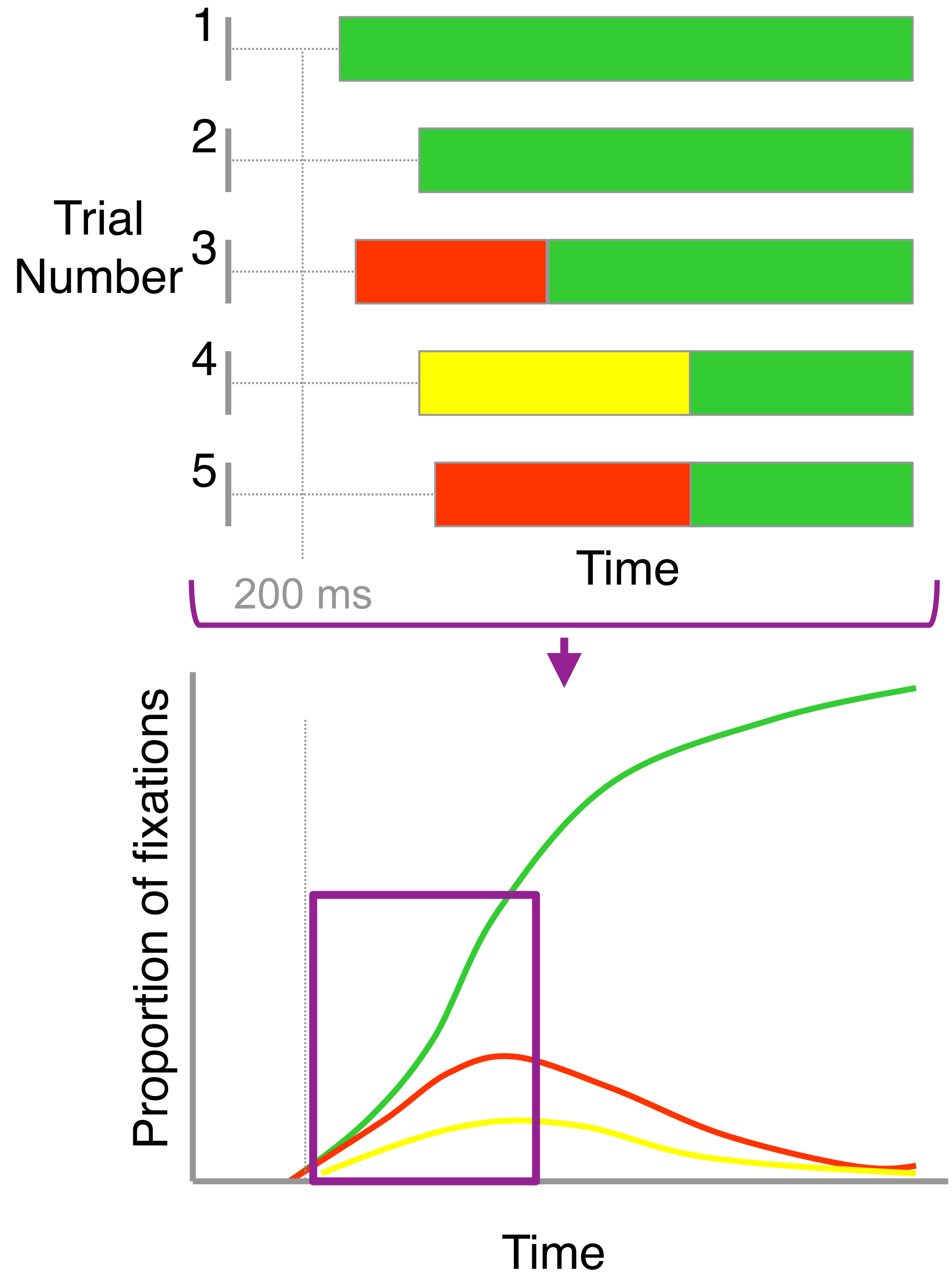
(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

"Pick up the beaker."





Target = beaker

Cohort = beetle

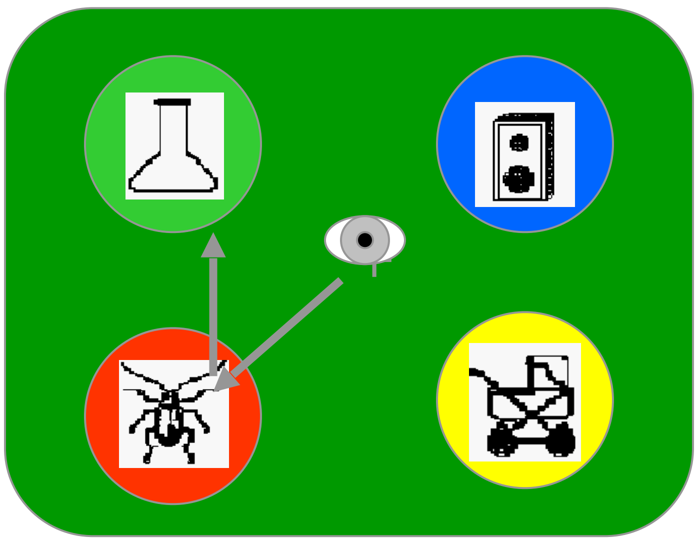
Unrelated = carriage

(Slide courtesy of Mike Tanenhaus)

Data from human eye movements

"Look at the cross."

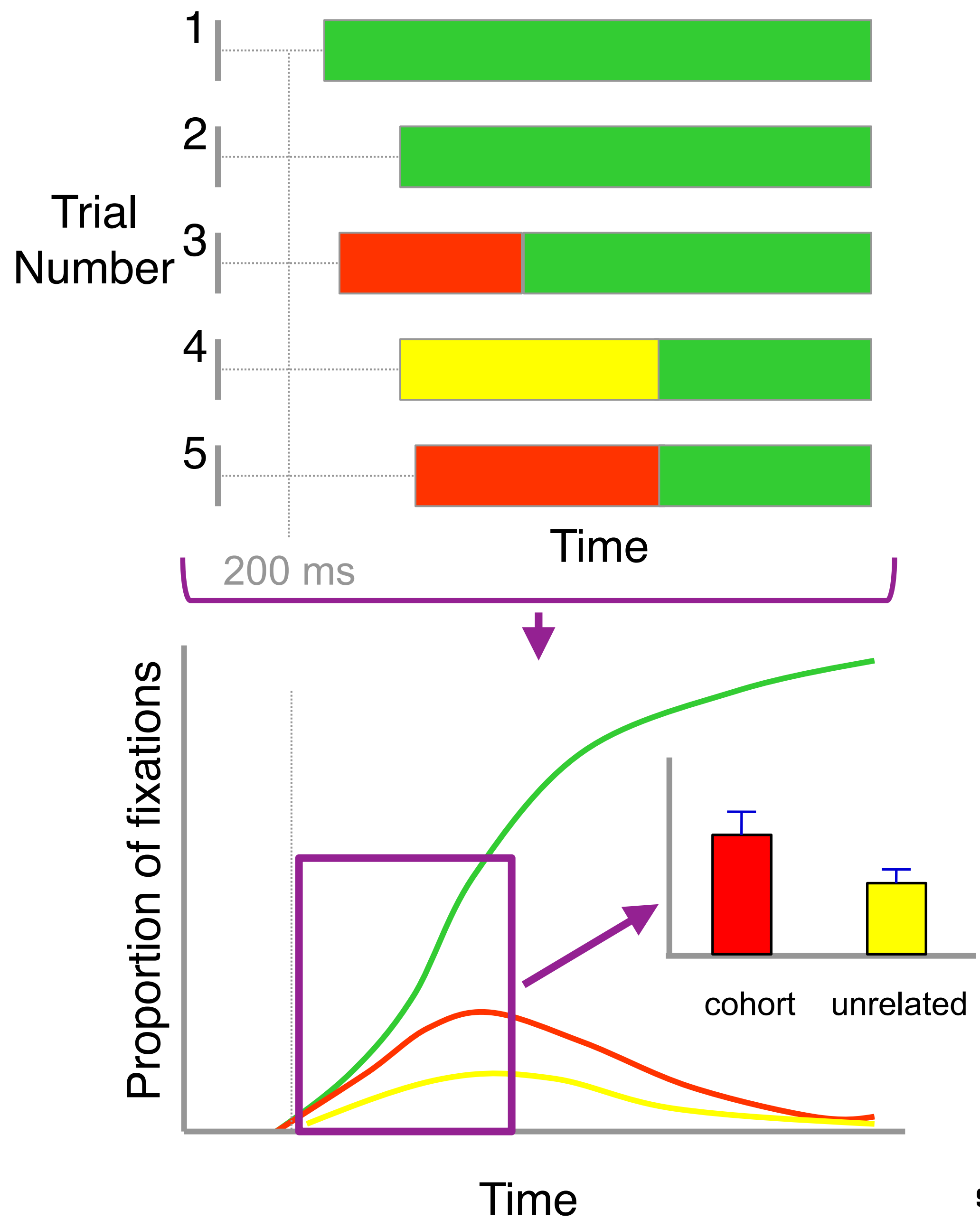
"Pick up the beaker."



Target = beaker

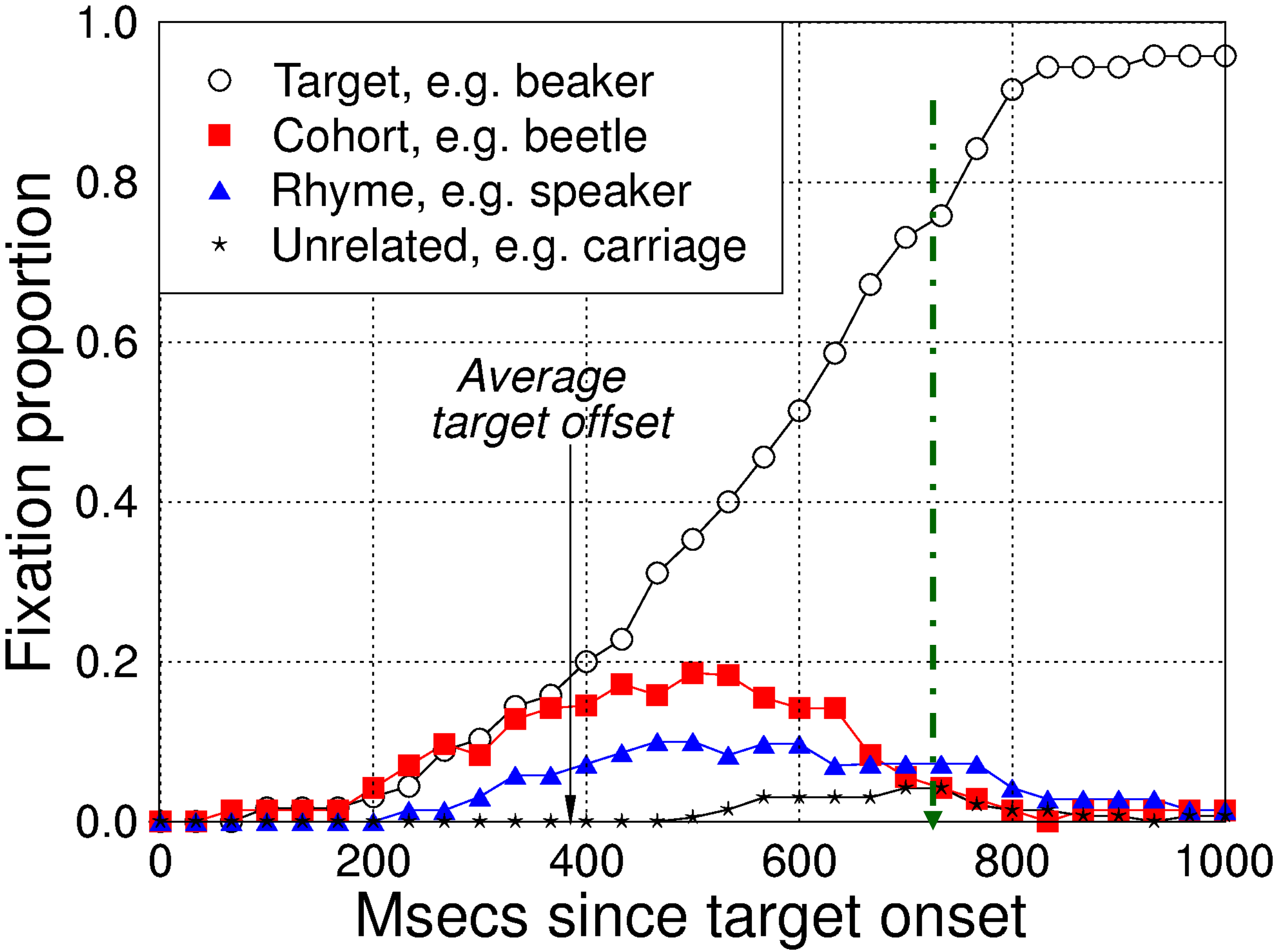
Cohort = beetle

Unrelated = carriage



(Slide courtesy of Mike Tanenhaus)

Allopenna, Magnuson & Tanenhaus (1998)



(Slide courtesy of Mike Tanenhaus)

Self-paced reading

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

-- --

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
While ----------------------------------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
----- the -------------------------------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
--------- clouds --------------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
---------------- crackled, ----------------------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
--------------------------- above ----------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
---------------------------------- the ------------------------------------
```

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
------------------------------------- glider -------------------------------
```

(*Mitchell, 1984*)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

`------------------------------------------- soared --------------------`

(*Mitchell, 1984*)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
---------------------------------------- soared ------------------
```

- Readers aren't allowed to backtrack

(Mitchell, 1984)

Self-paced reading

- Participant presses a button to reveal each successive word and mask previous words:

```
-------------------------------------- soared -------------------
```

- Readers aren't allowed to backtrack
- Duration between button presses="reading time" for each word

(Mitchell, 1984)

Language processing signal from the eyes


There are advantages and disadvantages of both electronic and hardcopy journals. Hardcopy journals are more easily browsed, more portable and, of course people are very much used to their format. Electronic journals save on paper and their format has improved considerably over the past few years, but there are still problems over managing copyright restrictions and persuading people to use electronic instead of hardcopy journals. There is also the problem of portability. More and more journals are now being published in electronic format, although some publishers will only let you subscribe to an electronic journal provided you also subscribe to the hardcopy (more money for the same thing). Some electronic journals cost over 100% more than their equivalent hardcopy. With all these factors in mind I have been discussing individual and shared-subscriptions with the Biochemistry Department, the RSL and Blackwell's. Whilst I feel that a move from hardcopy to electronic journals will be a very slow process in the ULP Library, electronic publishing is being carefully monitored and I would hope to introduce a few electronic texts into the Library alongside the journals which are already available for free over the Internet.

(movie by Piers Cornelissen)

Leaves a fine-grained trace of the real-time language comprehension record — we will put this to use later in the tutorial!

(Rayner, 1998)

Language processing signal from the eyes

(movie by Piers Cornelissen)

Leaves a fine-grained trace of the real-time language comprehension record — we will put this to use later in the tutorial!

(Rayner, 1998)

Electroencephalography (EEG/ERP)

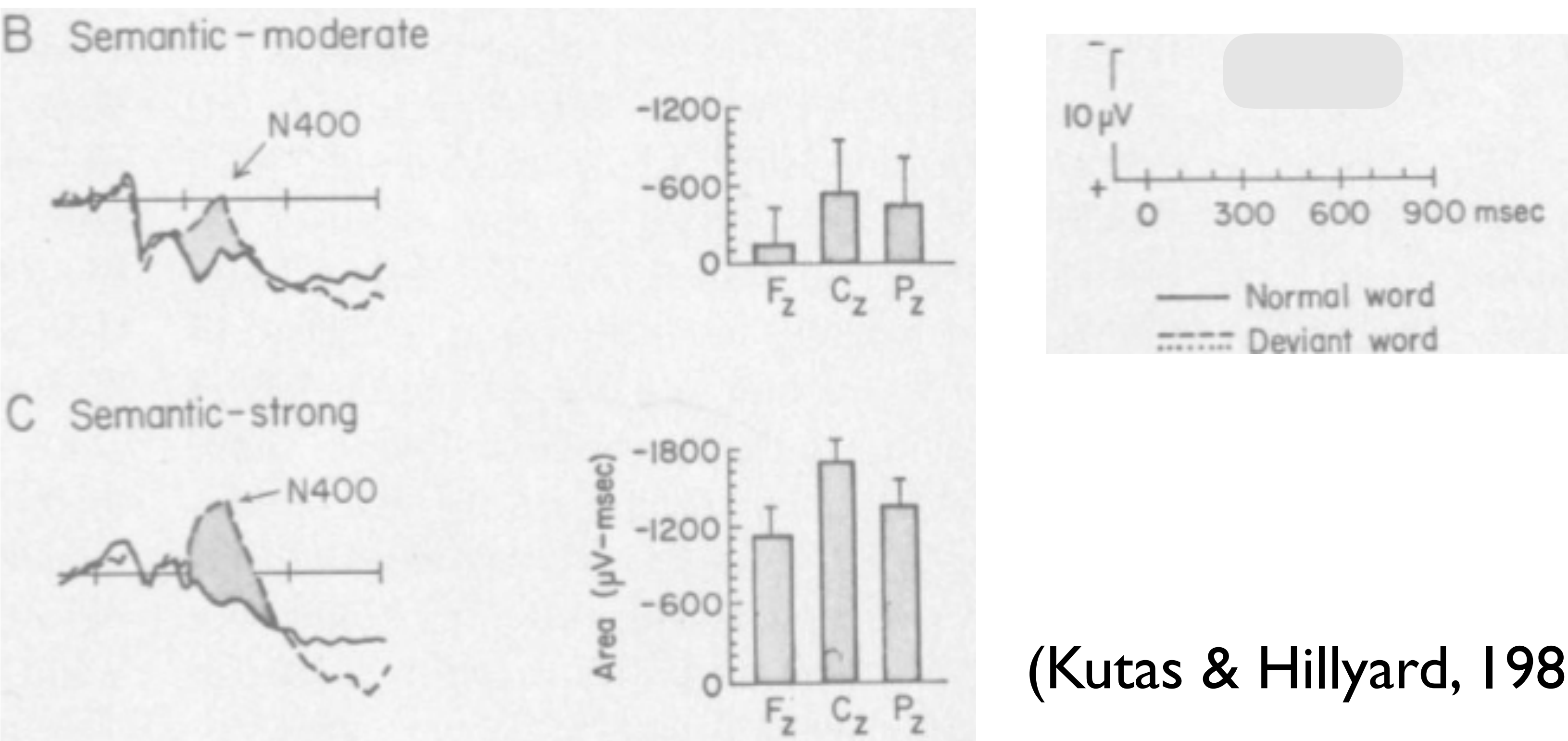
Rapid Serial Visual Presentation

*

Rapid Serial Visual Presentation

The N400 ERP component 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

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

The P600 ERP component in language comprehension

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

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

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.

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

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.

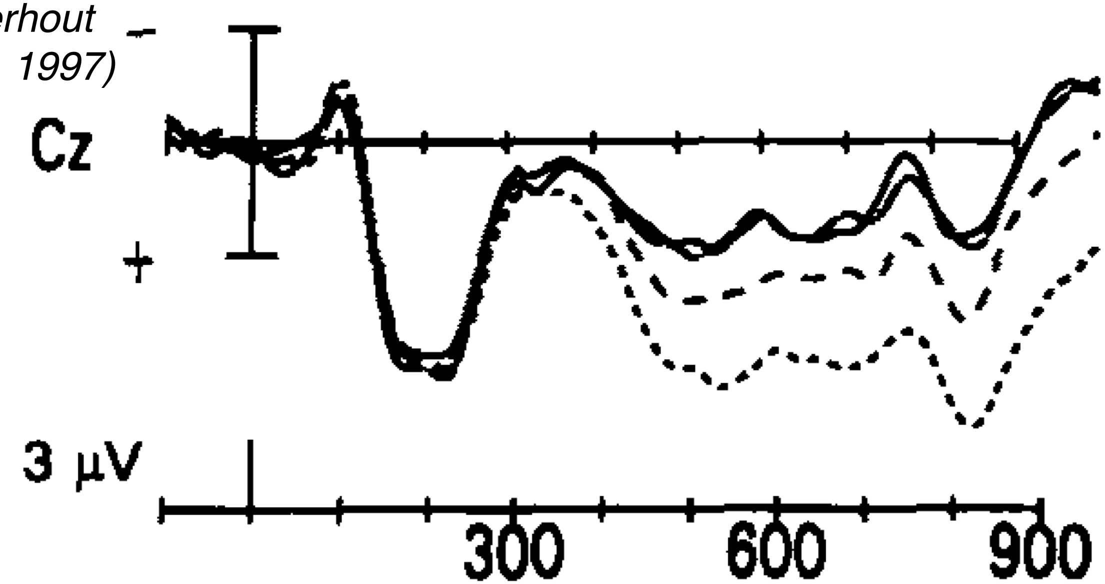
(Osterhout et al., 1997; see also reading time studies by Sturt, 2003; Duffy & Keir, 2004, inter alia)

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.

(Osterhout et al., 1997)



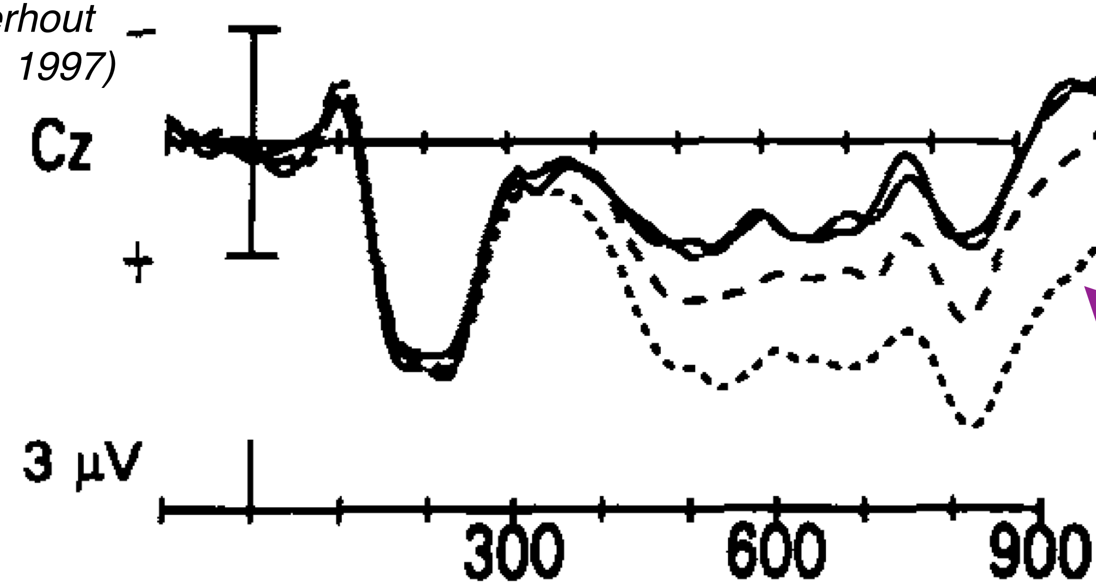
(Osterhout et al., 1997; see also reading time studies by Sturt, 2003; Duffy & Keir, 2004, inter alia)

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.

(Osterhout et al., 1997)



"Definitional" mismatch (man…herself)

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

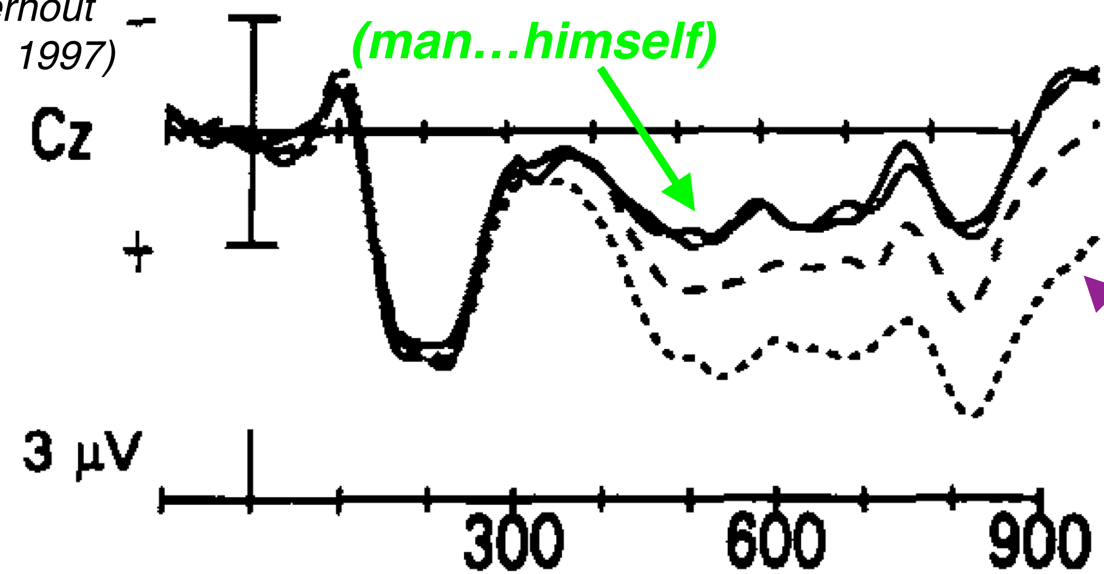
The P600 ERP component in language comprehension

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

The man prepared herself for the interview.

"Definitional" match (man…himself)

(Osterhout et al., 1997)



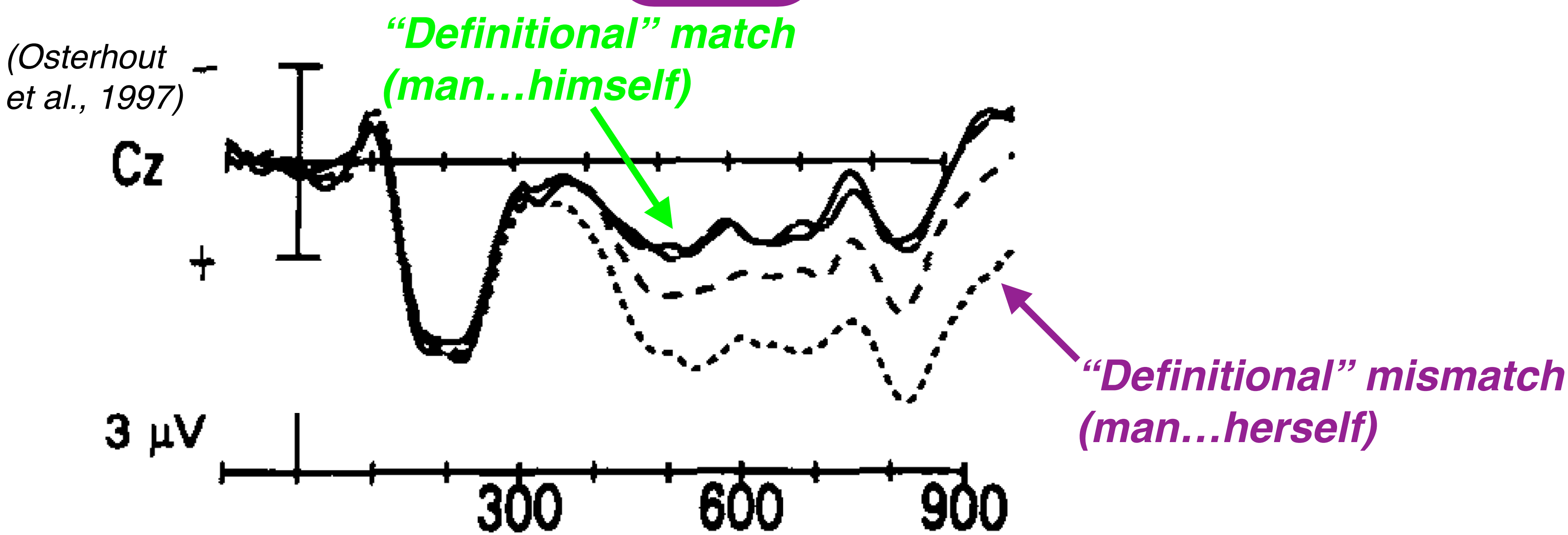
"Definitional" mismatch (man…herself)

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

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

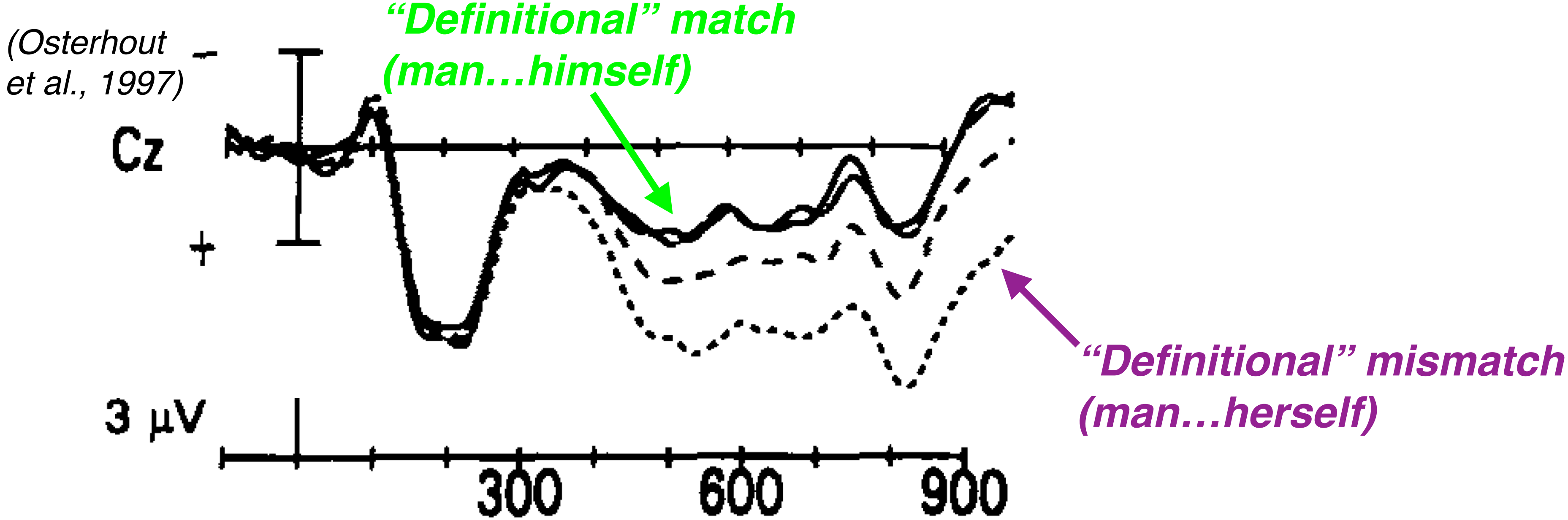
    The nurse prepared himself for the operation.

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

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

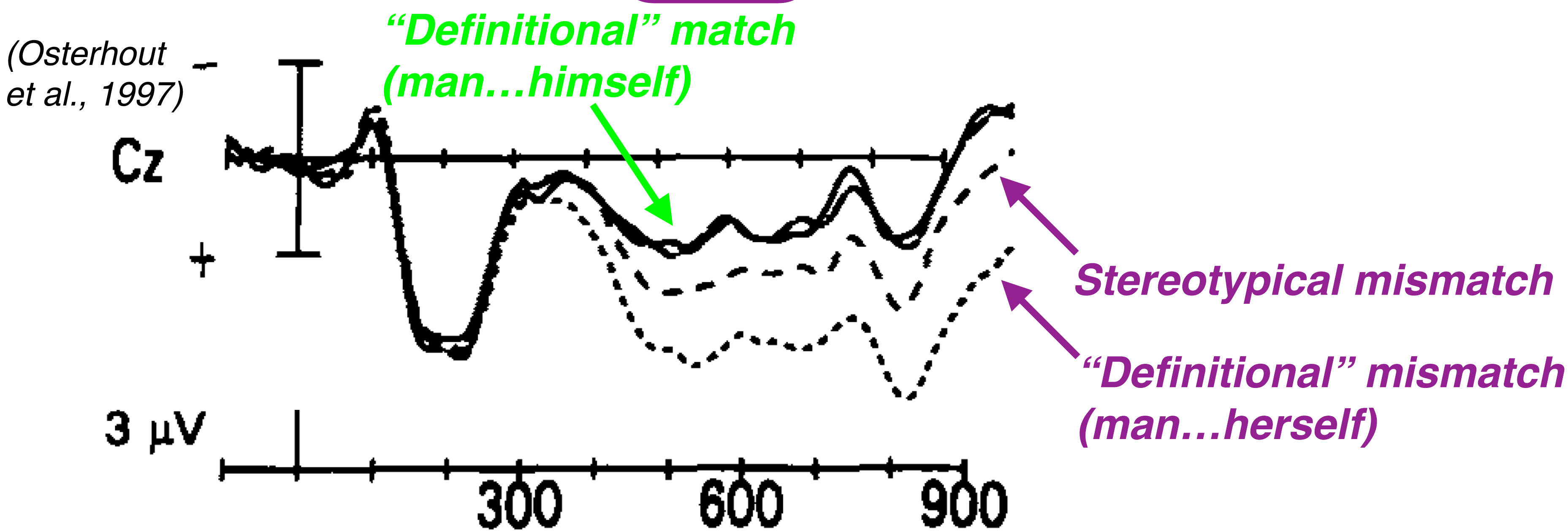
  The nurse prepared himself for the operation.

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

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.

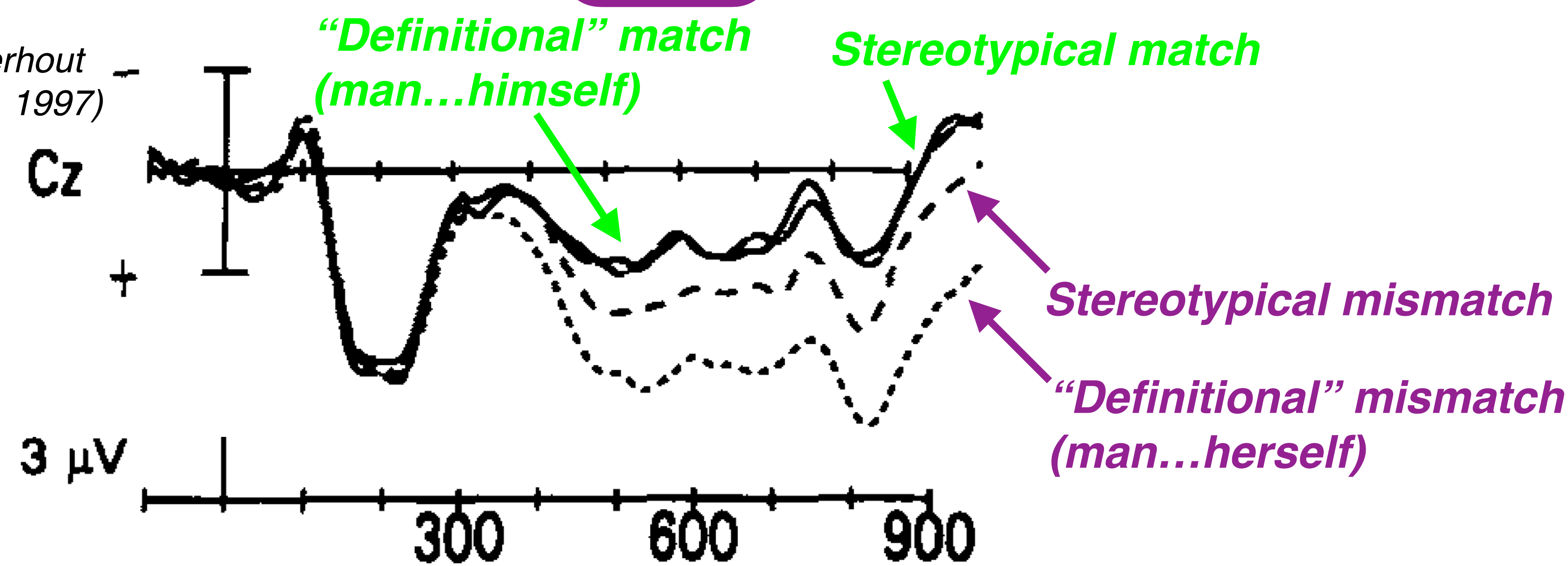
(Osterhout et al., 1997; see also reading time studies by Sturt, 2003; Duffy & Keir, 2004, inter alia)

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.

(Osterhout et al., 1997)



- 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

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	+

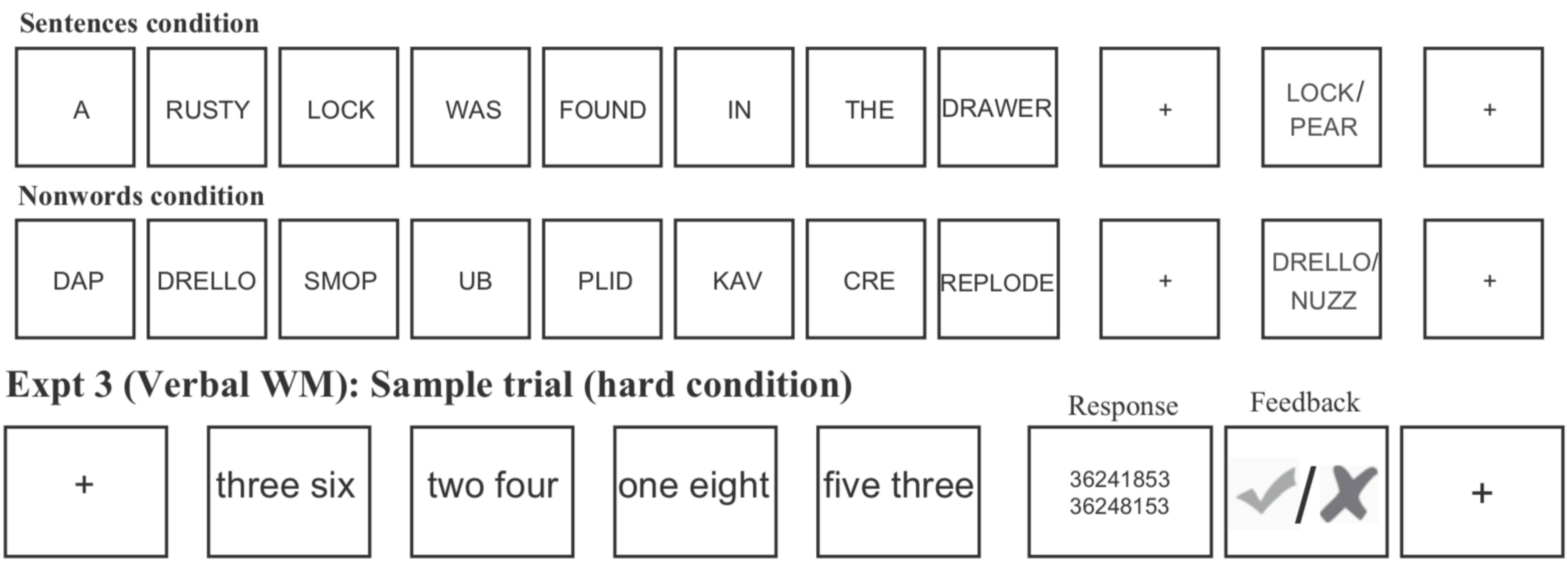
Nonwords condition

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

(Fedorenko et al., 2011)

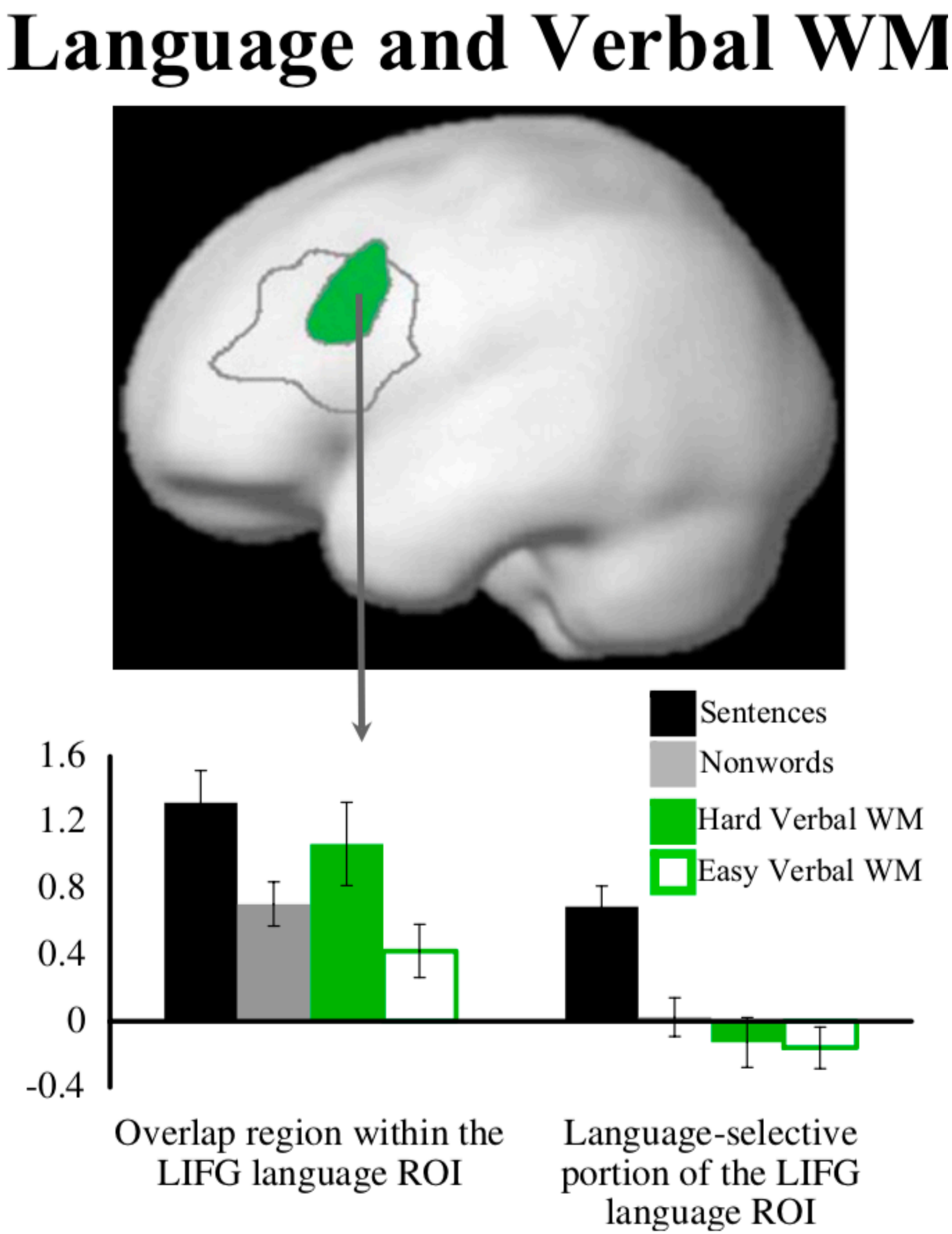
fMRI recordings during comprehension

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



(Fedorenko et al., 2011)

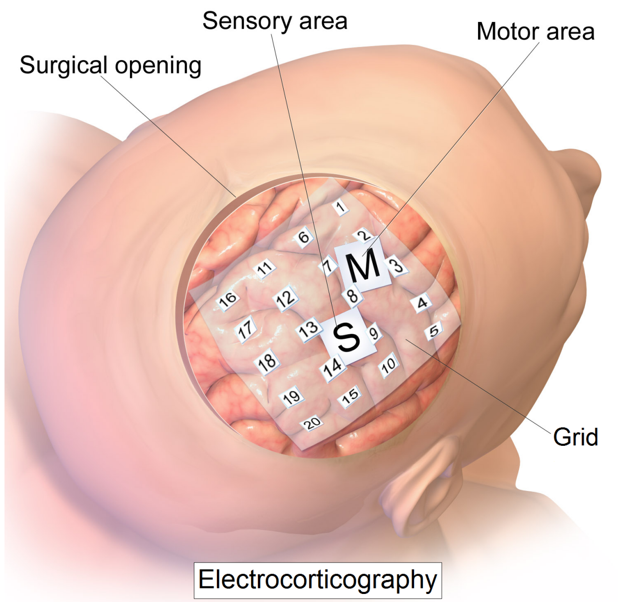
Functional brain specificity for language



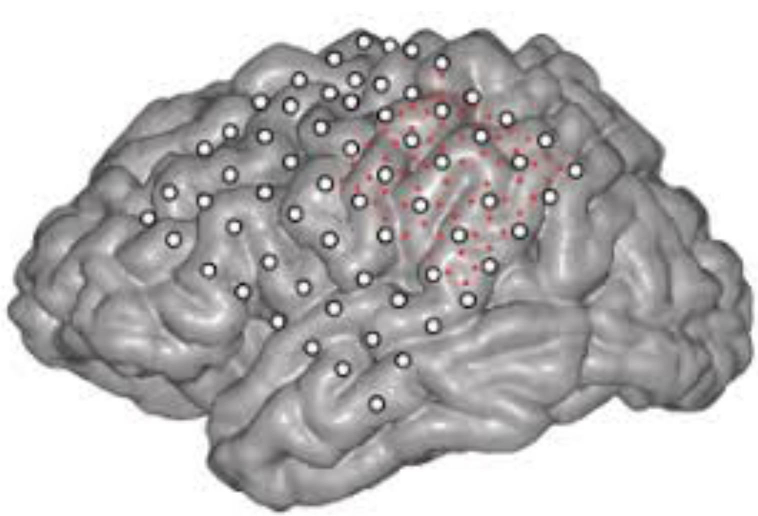
(Fedorenko et al., 2011)

Electrocorticography

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



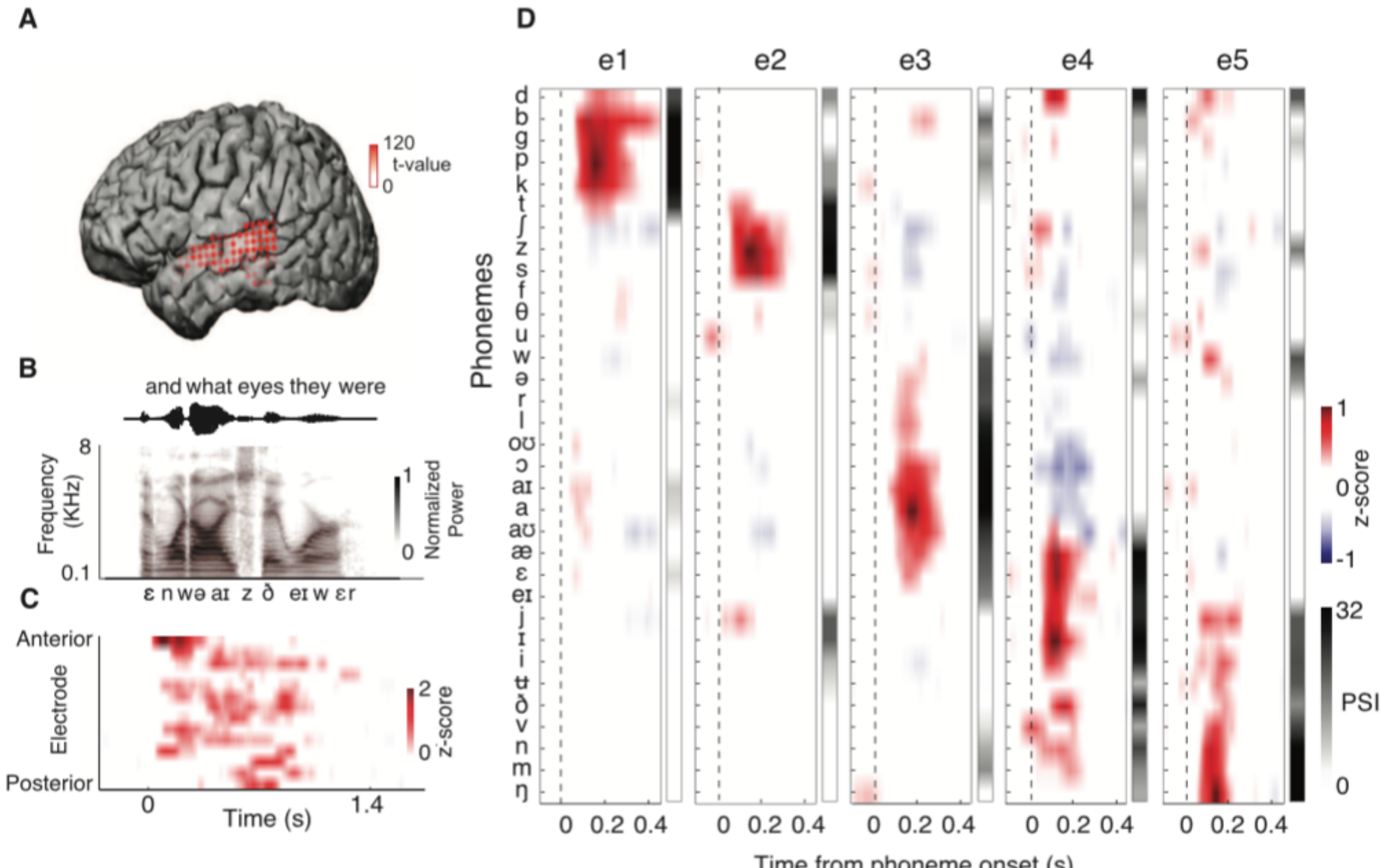
https://commons.wikimedia.org/wiki/File:Intracranial_electrode_grid_for_electrocorticography.png



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

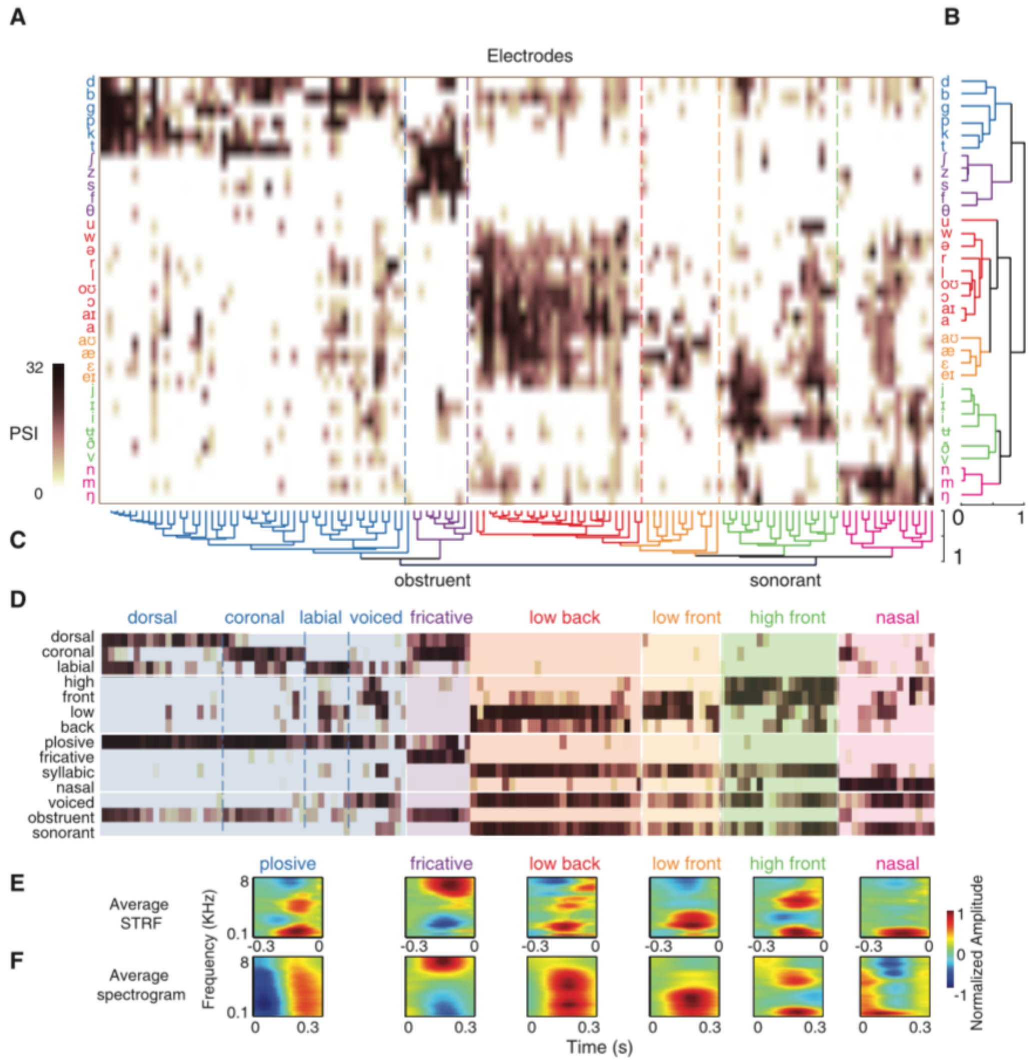
- During pre-surgical monitoring many patients generously donate their energy & attention for experiments

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

Rational analysis

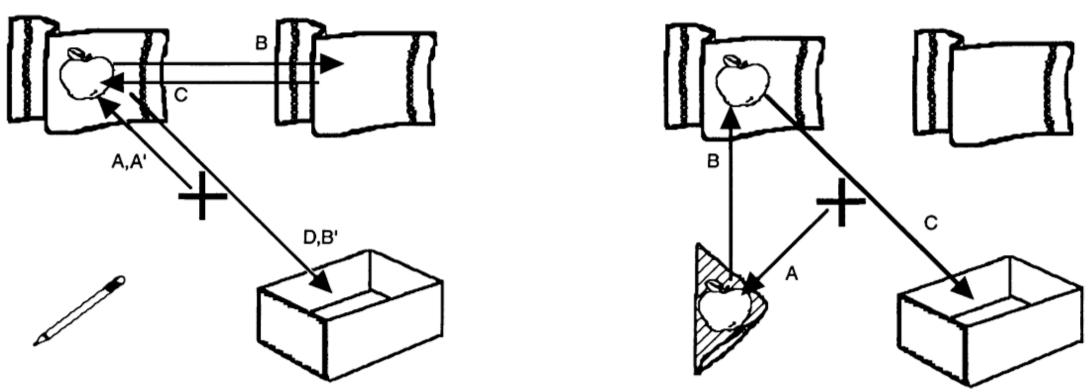
- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively

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)

Incrementality and Rationality

- Real-time language understanding 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, Science)*

- Enter probabilistic grammars from computational linguistics...

Probabilistic Context-Free Grammars

A *probabilistic* context-free grammar (PCFG) consists of a tuple (N, V, S, R, P) such that:

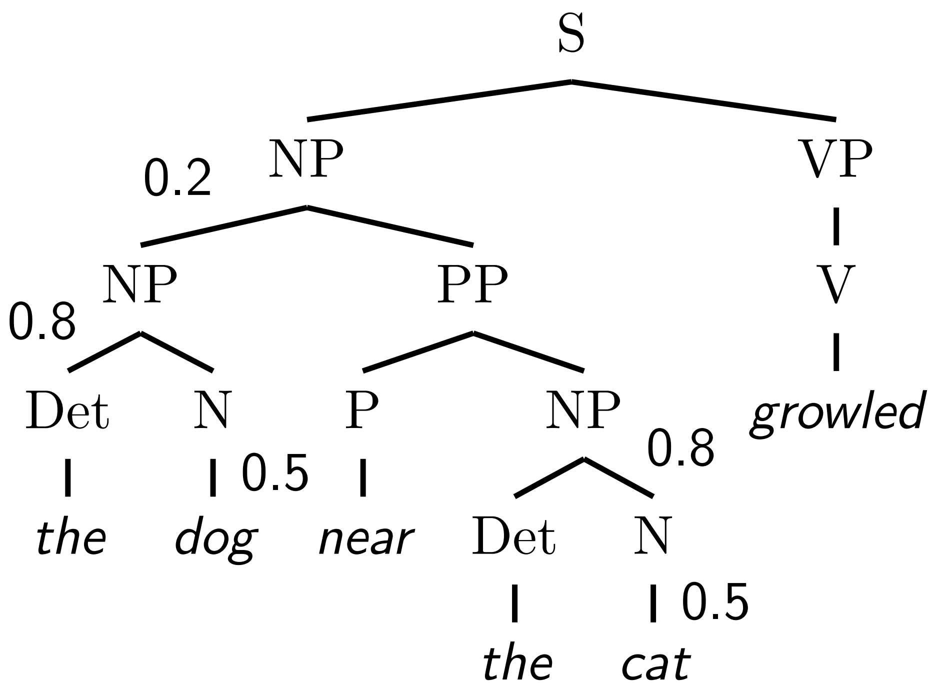
- N is a finite set of non-terminal symbols;
- V is a finite set of terminal symbols;
- S is the start symbol;
- R is a finite set of rules of the form $X \to \alpha$ where $X \in N$ and α is a sequence of symbols drawn from $N \cup V$;
- P is a mapping from R into probabilities, such that for each $X \in N$,

$$\sum_{[X \to \alpha] \in R} P(X \to \alpha) = 1$$

PCFG *derivations* and *derivation trees* are just like for CFGs. The probability $P(T)$ of a derivation tree is simply the product of the probabilities of each rule application.

Example PCFG

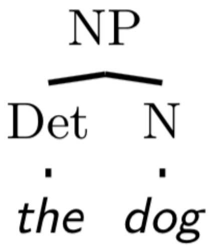
1	S $\rightarrow$ NP VP	1	Det $\rightarrow$ the
0.8	NP $\rightarrow$ Det N	0.5	N $\rightarrow$ dog
0.2	NP $\rightarrow$ NP PP	0.5	N $\rightarrow$ cat
1	PP $\rightarrow$ P NP	1	P $\rightarrow$ near
1	VP $\rightarrow$ V	1	V $\rightarrow$ growled



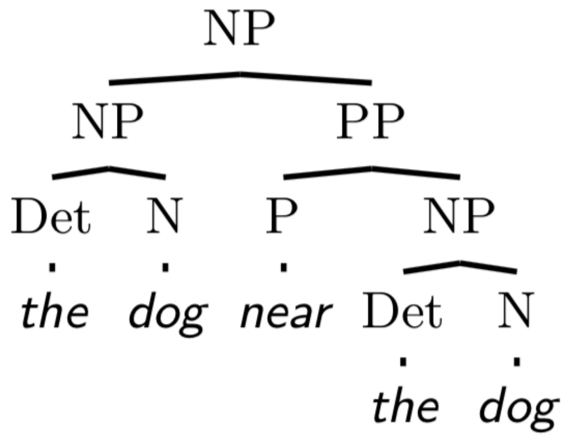
$$
\begin{aligned}
P(T) &= 1 \times 0.2 \times 0.8 \times 1 \times 0.5 \times 1 \times 1 \times 0.8 \times 1 \times 0.5 \times 1 \times 1 \\
&= 0.032
\end{aligned}
$$

		1	Det $\rightarrow$ the
$\frac{2}{3}$	NP $\rightarrow$ Det N	$\frac{2}{3}$	N $\rightarrow$ dog
$\frac{1}{3}$	NP $\rightarrow$ NP PP	$\frac{1}{3}$	N $\rightarrow$ cat
1	PP $\rightarrow$ P NP	1	P $\rightarrow$ near

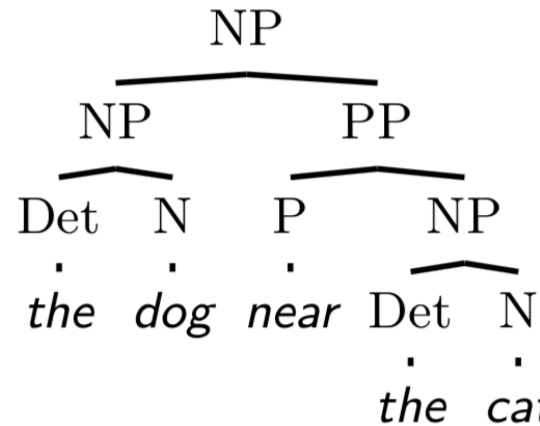
Incrementality: you can think of a *partial* tree as marginalizing over all completions of the partial tree.
It has a corresponding marginal probability in the PCFG.



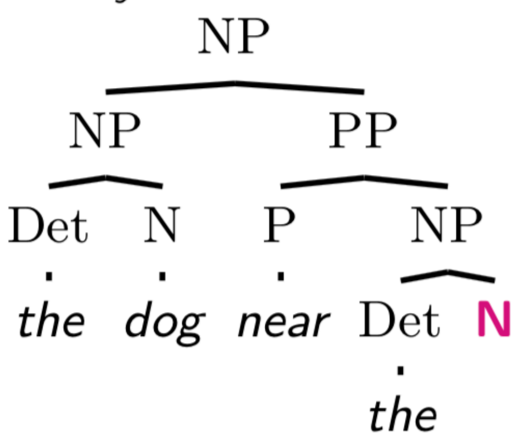
$\frac{4}{9}$



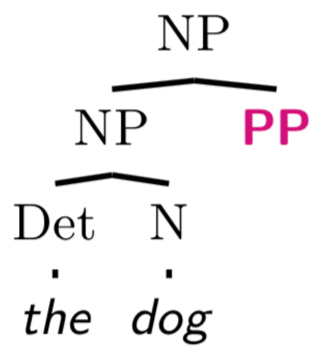
$\frac{8}{81}$



$\frac{4}{81}$



$\frac{12}{81}$



$\frac{4}{27}$

A zeroth-cut theory of incremental comprehension

- Human knowledge described by a probabilistic grammar

1	$S \rightarrow NP\ VP$		1	$Det \rightarrow$ the	
0.8	$NP \rightarrow Det\ N$		0.5	$N \rightarrow$ dog	
0.2	$NP \rightarrow NP\ PP$		0.5	$N \rightarrow$ cat	
1	$PP \rightarrow P\ NP$		1	$P \rightarrow$ near	
1	$VP \rightarrow V$		1	$V \rightarrow$ growled	
	…			…	

- Incremental input interpretation follows **Bayes Rule**:

$$P(\mathrm{T} \mid \mathrm{words}) \propto P(\mathrm{words} \mid T)P(T)$$

(Jurafsky, 1996)

Strong garden-pathing

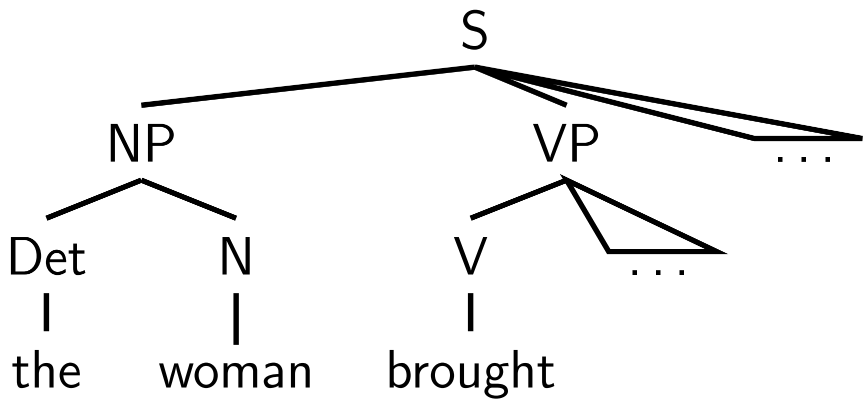
(Levy, Reali, & Griffiths, 2009)

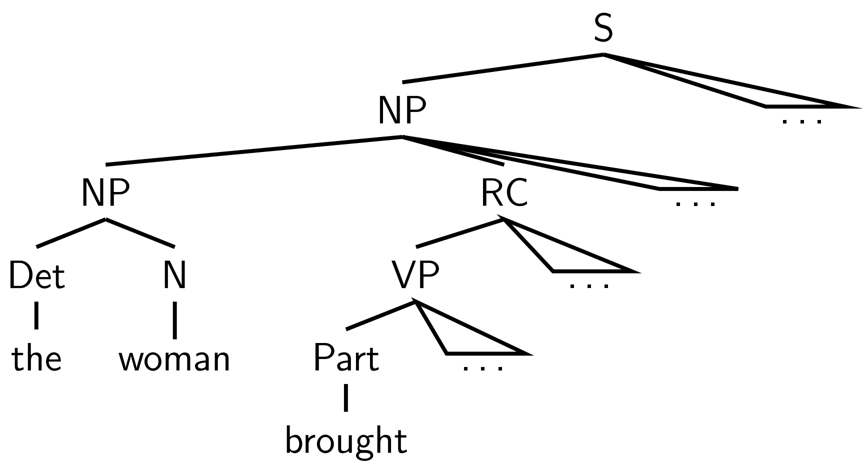
Strong garden-pathing

The woman brought

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing

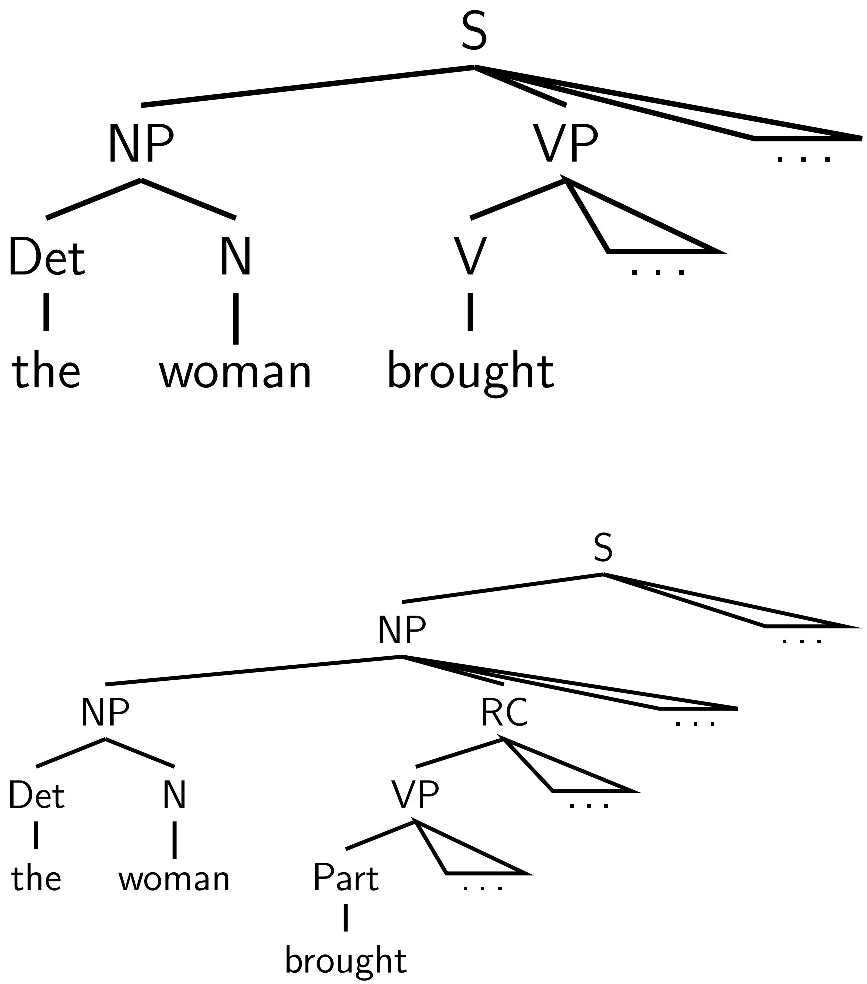




The woman brought

(*Levy, Reali, & Griffiths, 2009*)

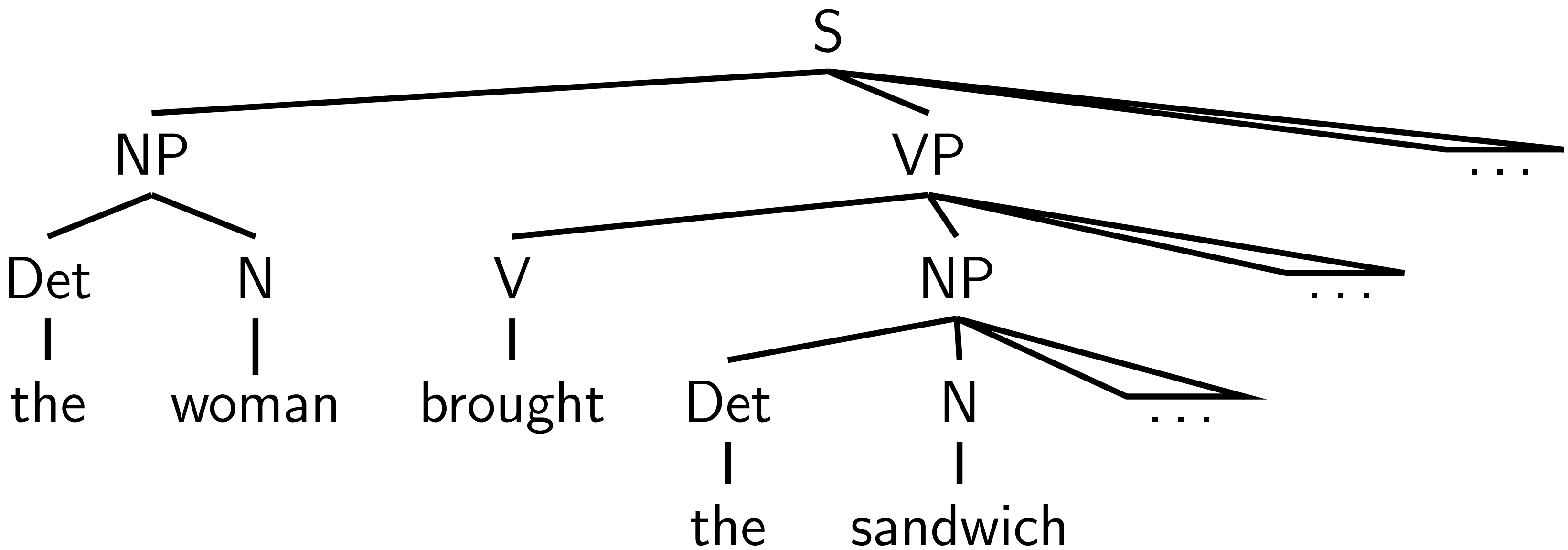
Strong garden-pathing

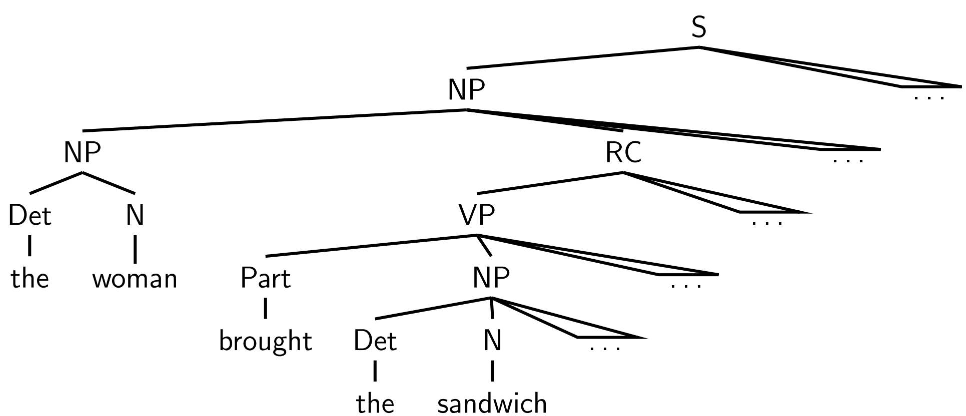


The woman brought the sandwich

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing

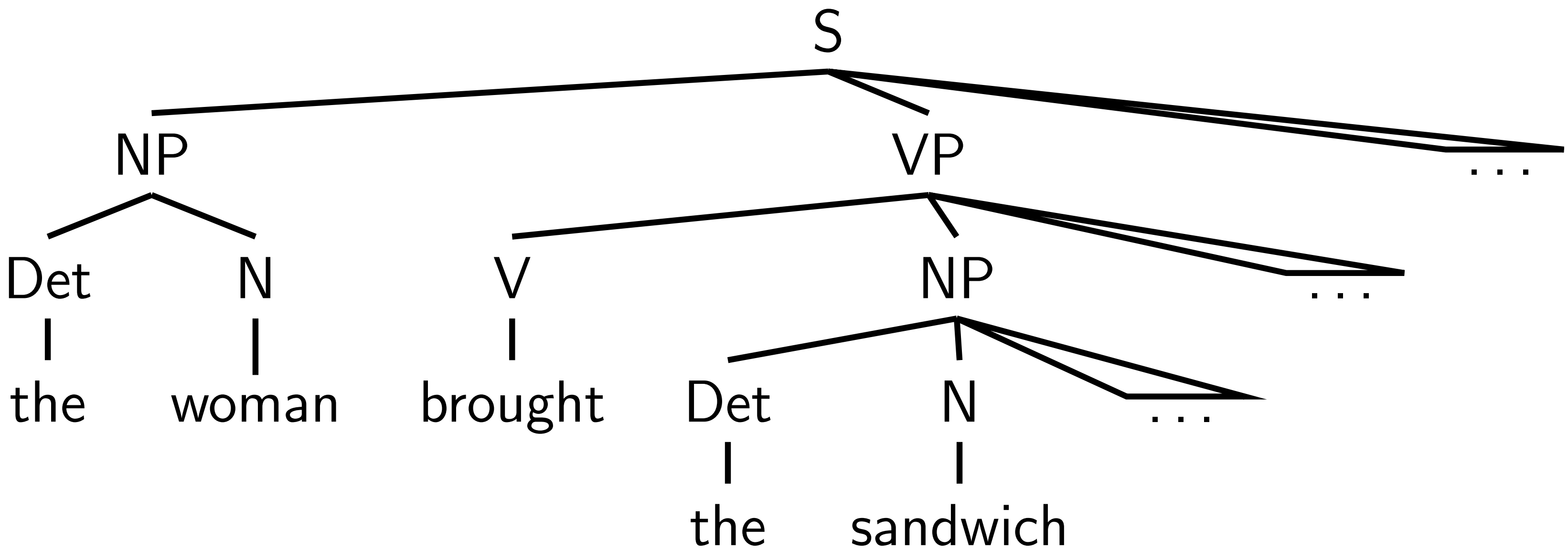


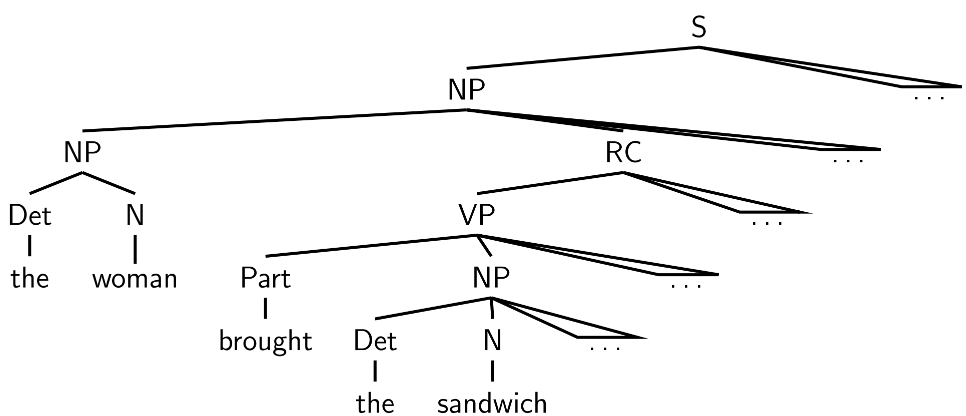


The woman brought the sandwich

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing

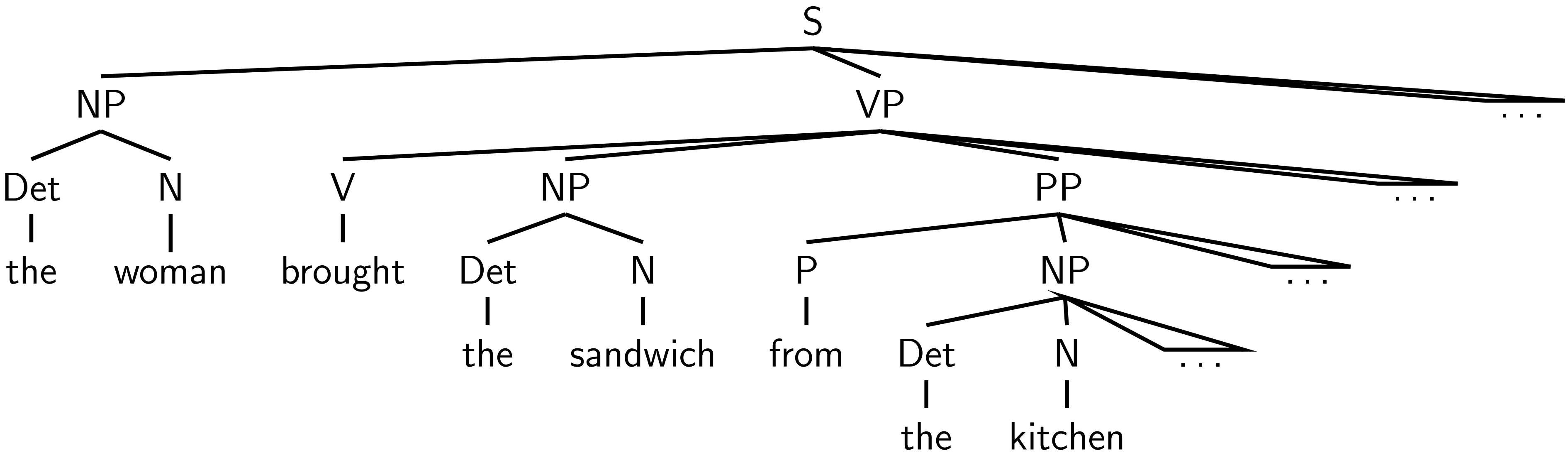


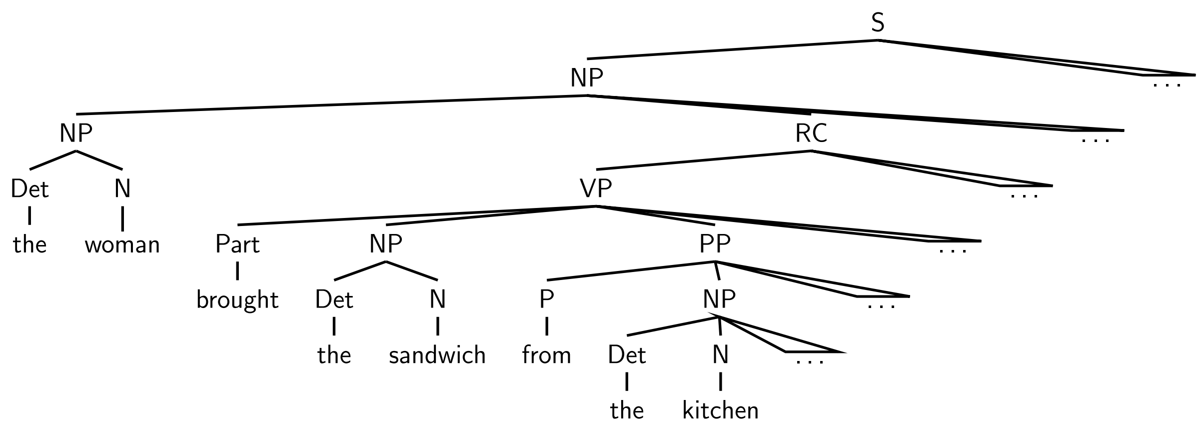


`The woman brought the sandwich from the kitchen`

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing

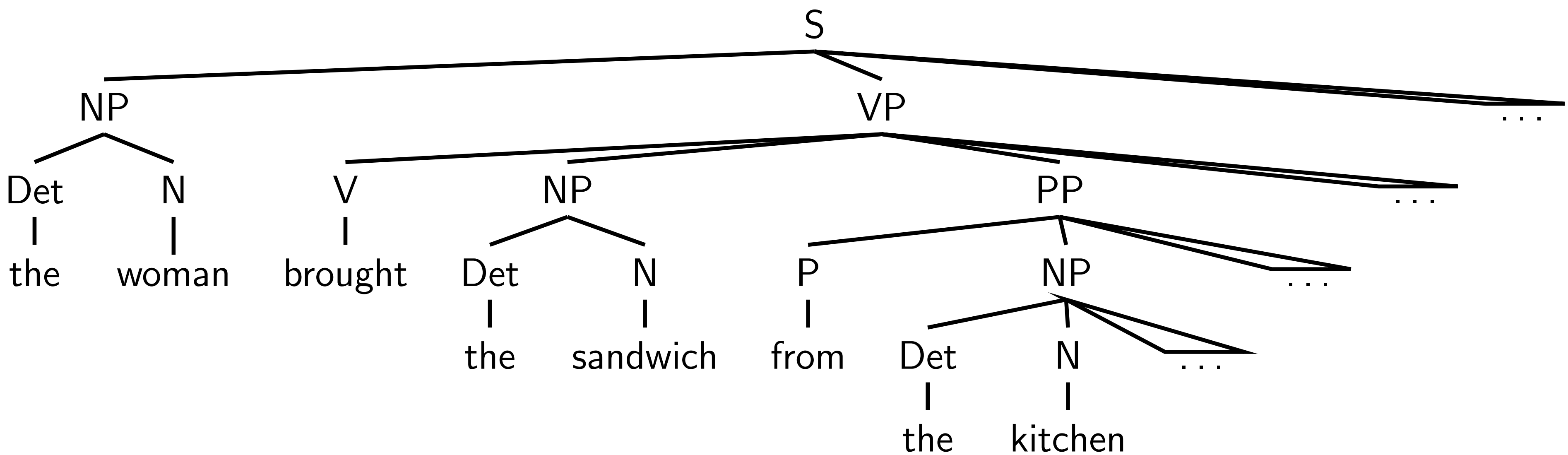


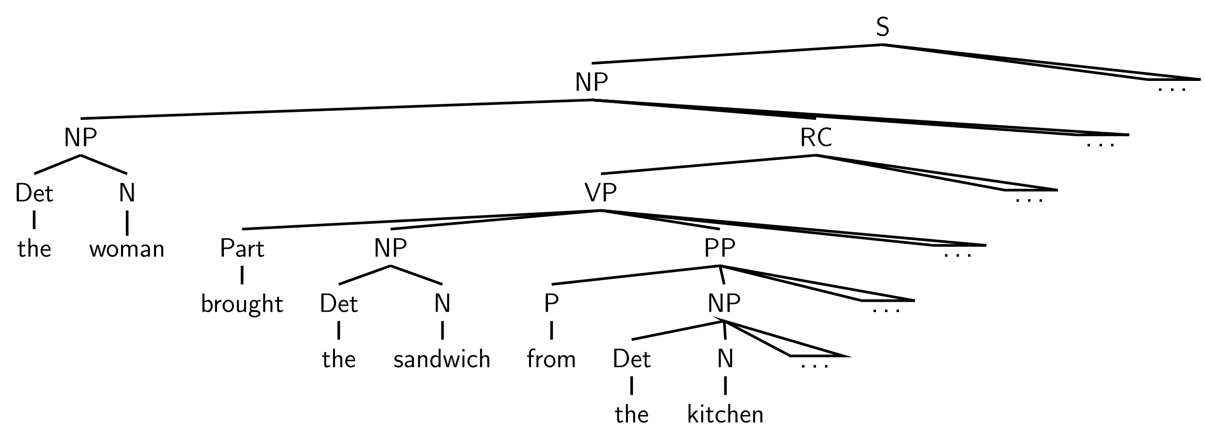


The woman brought the sandwich from the kitchen

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing

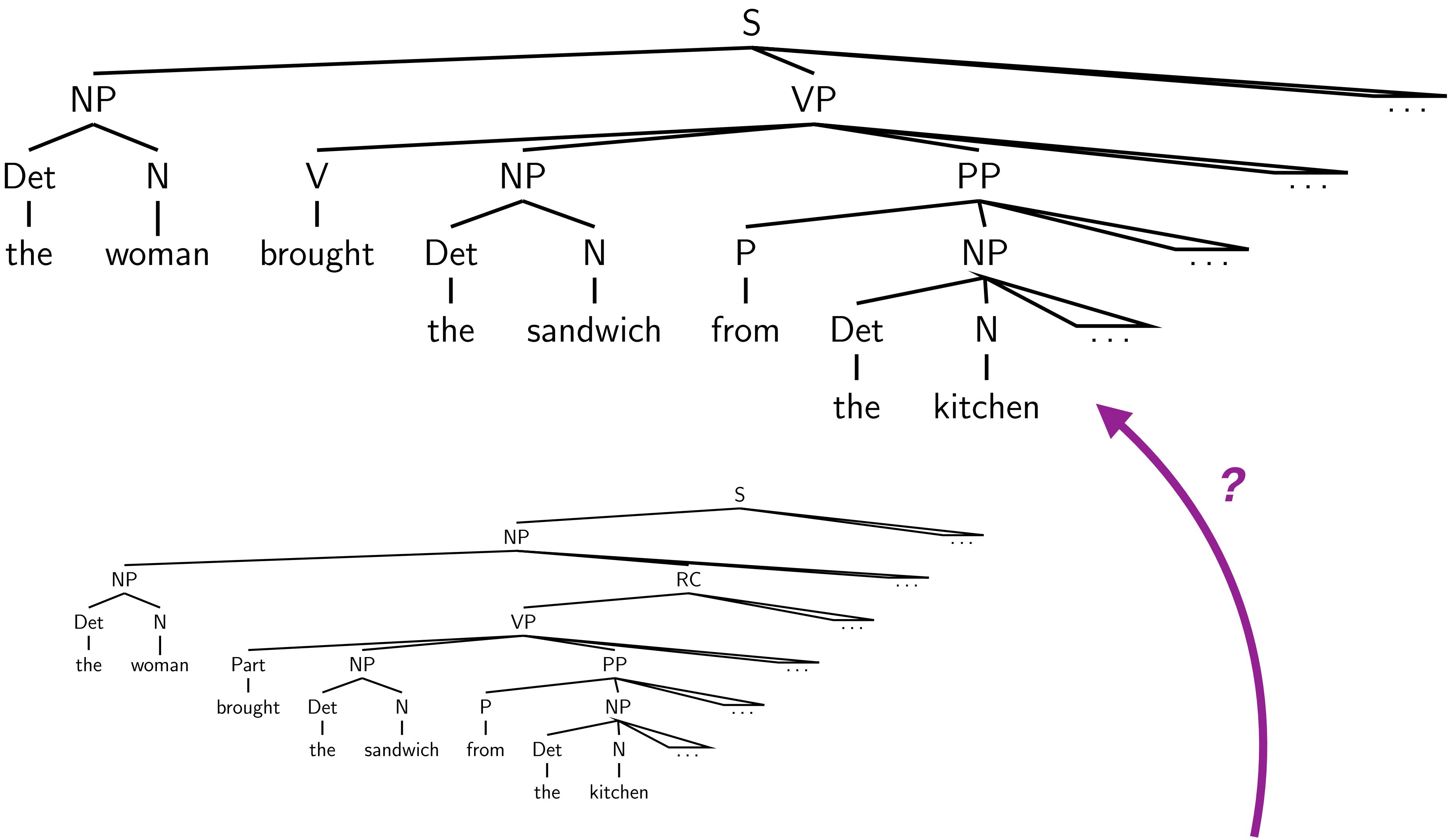




The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

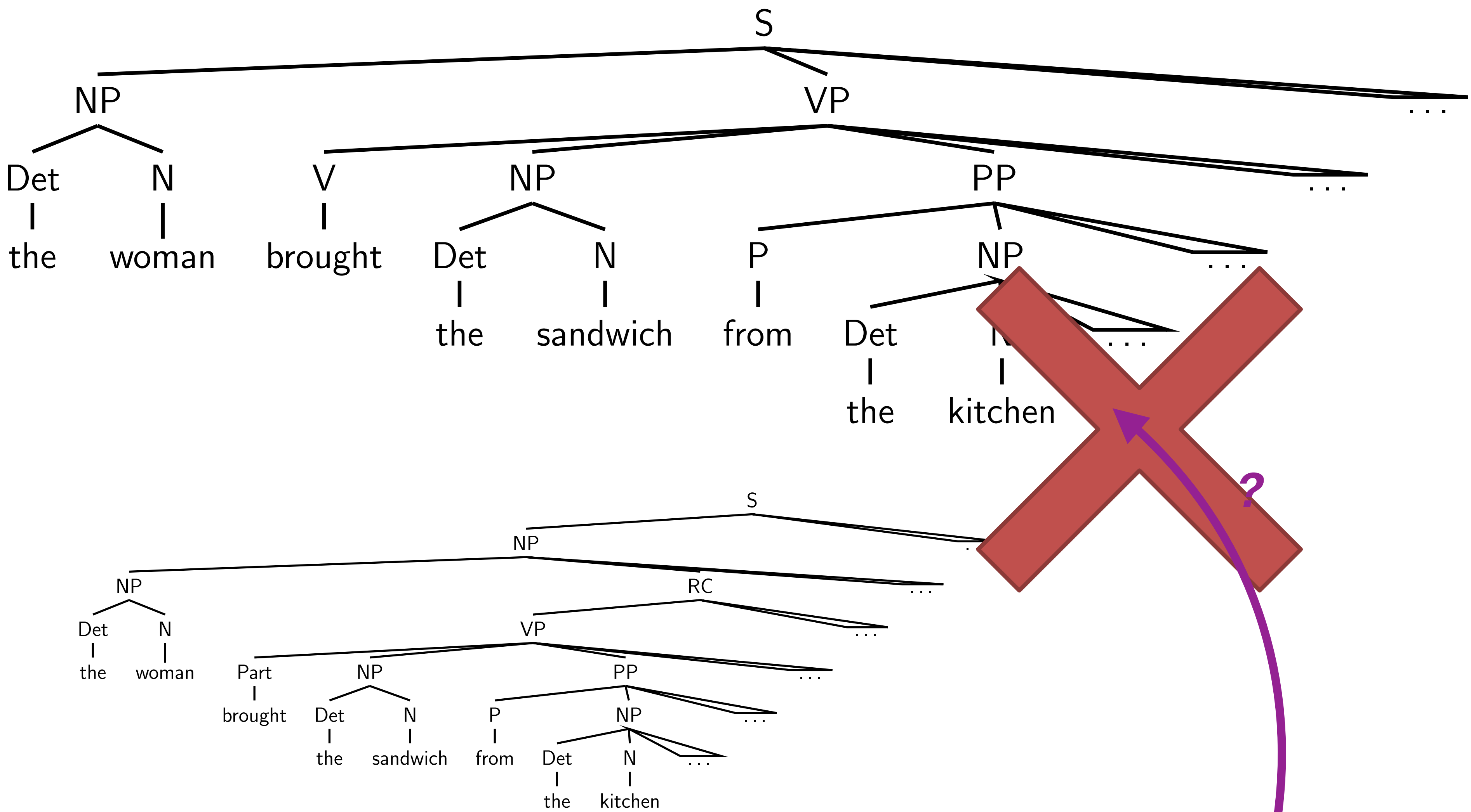
Strong garden-pathing



The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

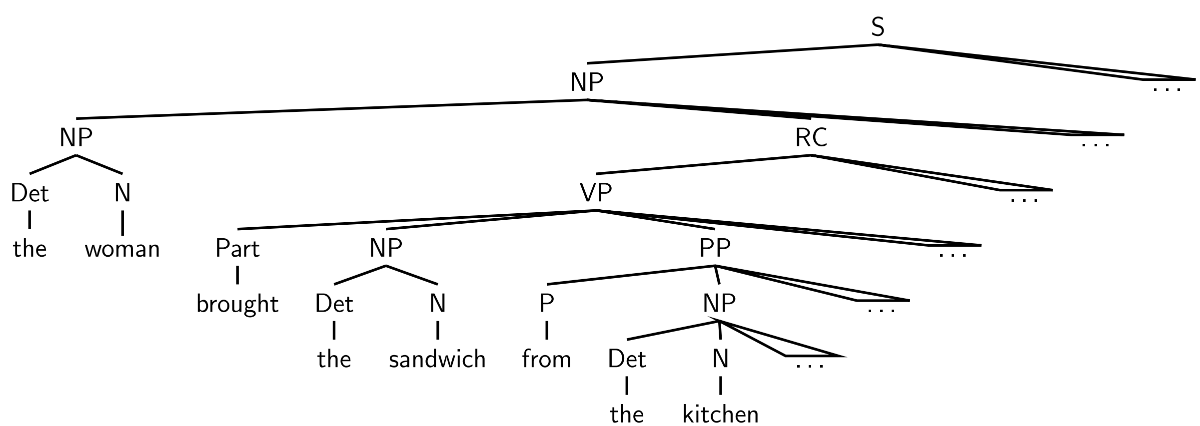
Strong garden-pathing



The woman brought the sandwich from the kitchen tripped.

(*Levy, Reali, & Griffiths, 2009*)

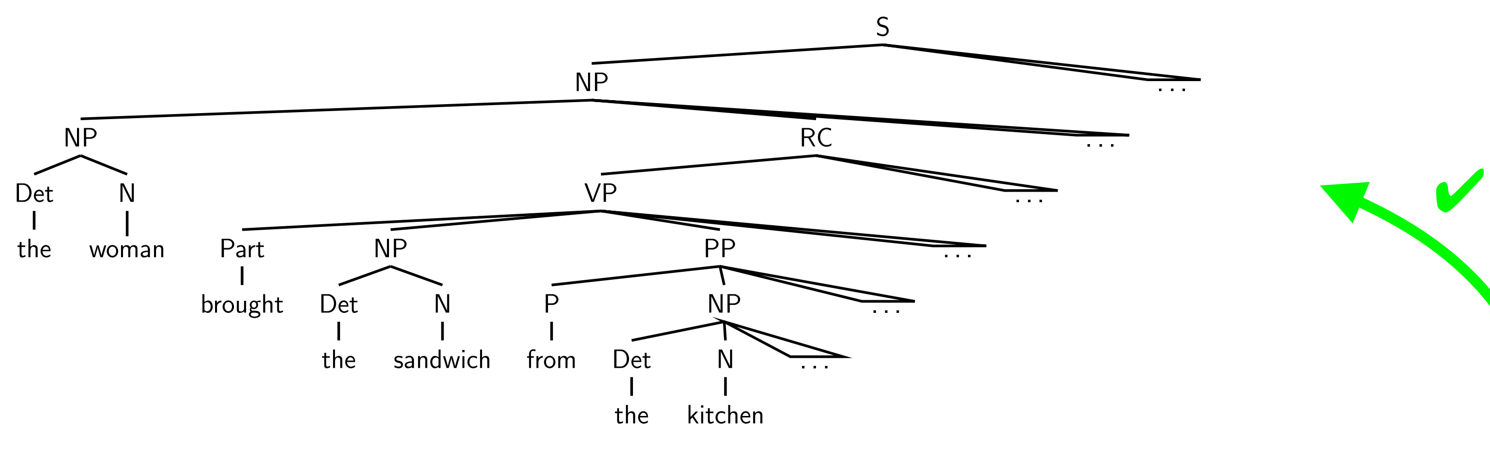
Strong garden-pathing



The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

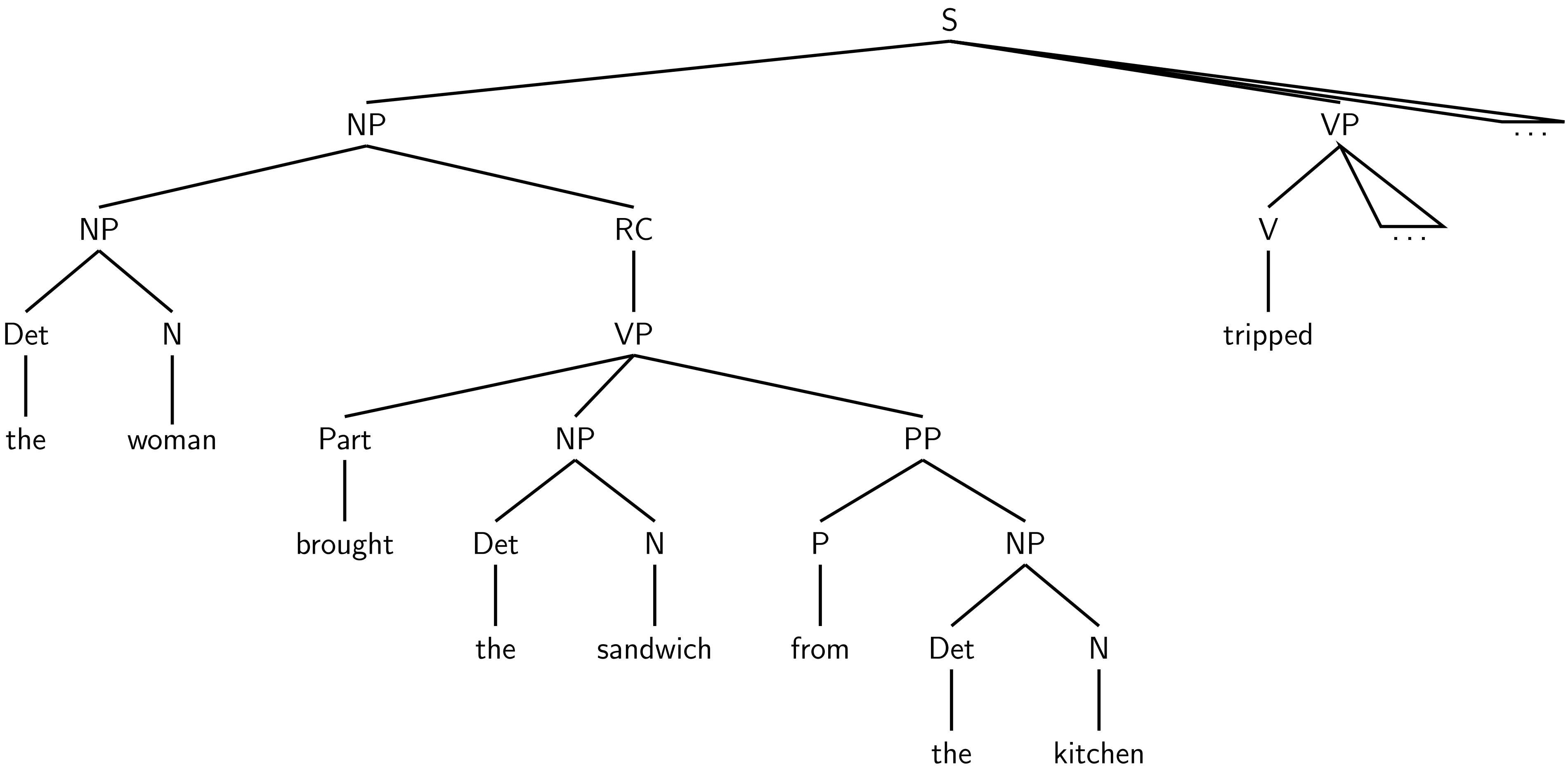
Strong garden-pathing



The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

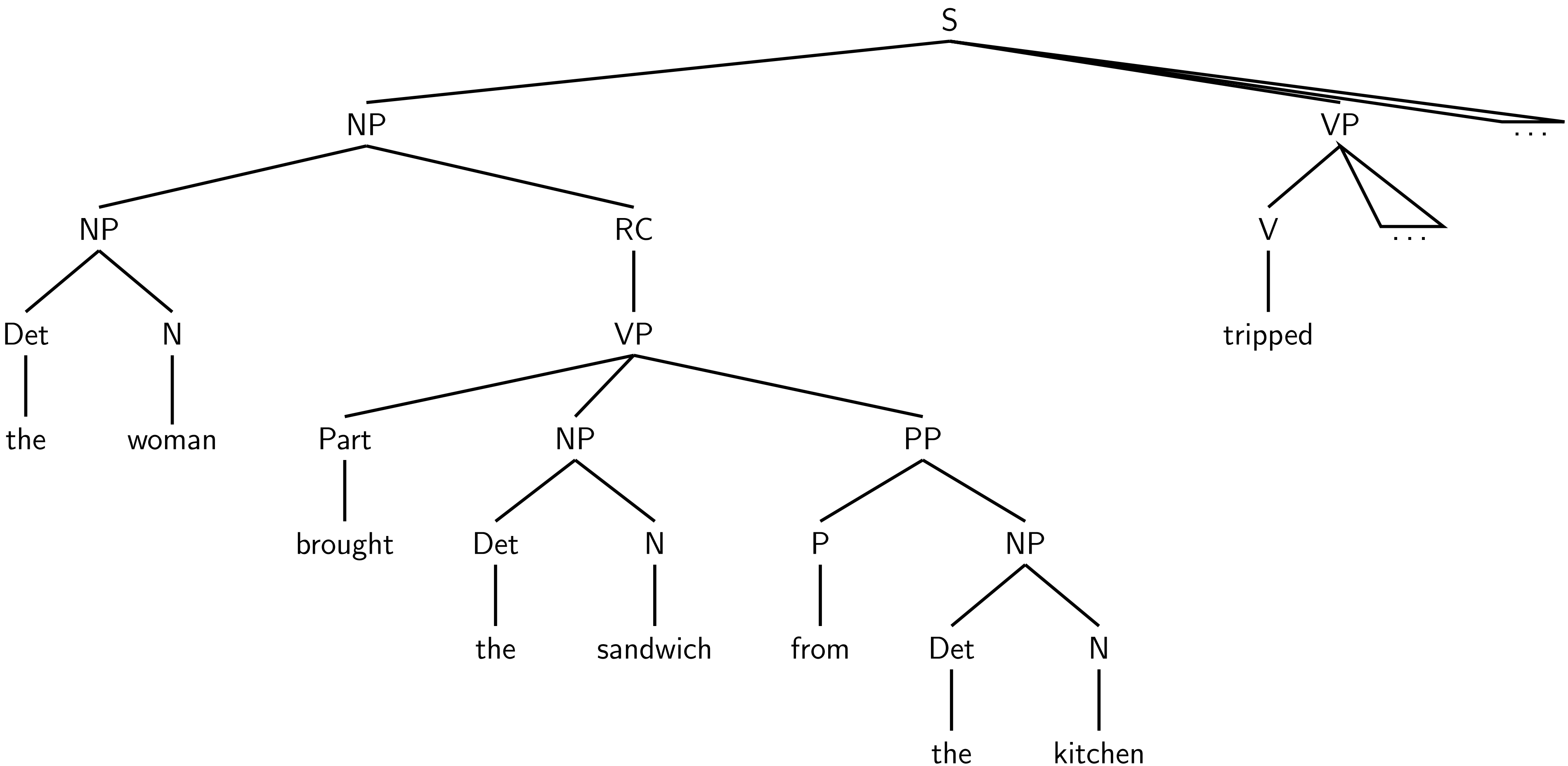
Strong garden-pathing



The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

Strong garden-pathing



Comprehension only successful if the earlier-disfavored interpretation is still available!!!

The woman brought the sandwich from the kitchen tripped.

(Levy, Reali, & Griffiths, 2009)

But not all garden paths are catastrophic:

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

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

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

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

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

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

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

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

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

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

When the dog scratched the vet and his new assistant $\boxed{\text{removed}}$ the muzzle.

difficulty here
(68ms/char)

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of 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.

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

When the dog scratched the vet and his new assistant $\boxed{\text{removed}}$ 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.

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of 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.

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of 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.

(Frazier & Rayner, 1982)

But not all garden paths are catastrophic:

- Here's another type of local syntactic ambiguity:

When the dog scratched the vet and his new assistant $\boxed{\text{removed}}$ 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 $\boxed{\text{removed}}$ the muzzle.

easier
(50ms/char)

(Frazier & Rayner, 1982)

A first-cut theory of incremental comprehension:

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

- Captures the *expectation* intuition: the more we expect an event, the easier it is to process

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

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

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

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

- 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…

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But 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?

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

- 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*?

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But 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*?

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But 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…

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

$$\begin{aligned} \text{Surprisal}(w_i) \quad &\equiv \quad \log \frac{1}{P(w_i|\text{CONTEXT})} \\ \left[\approx \right. \quad &\left. \log \frac{1}{P(w_i|w_{1\ldots i-1})} \right] \end{aligned}$$

- 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*

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But 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)

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

A first-cut theory of incremental comprehension:

- Stick with probabilistic grammars and Bayesian inference
- But let a word's difficulty be its *surprisal* given its context:

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

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

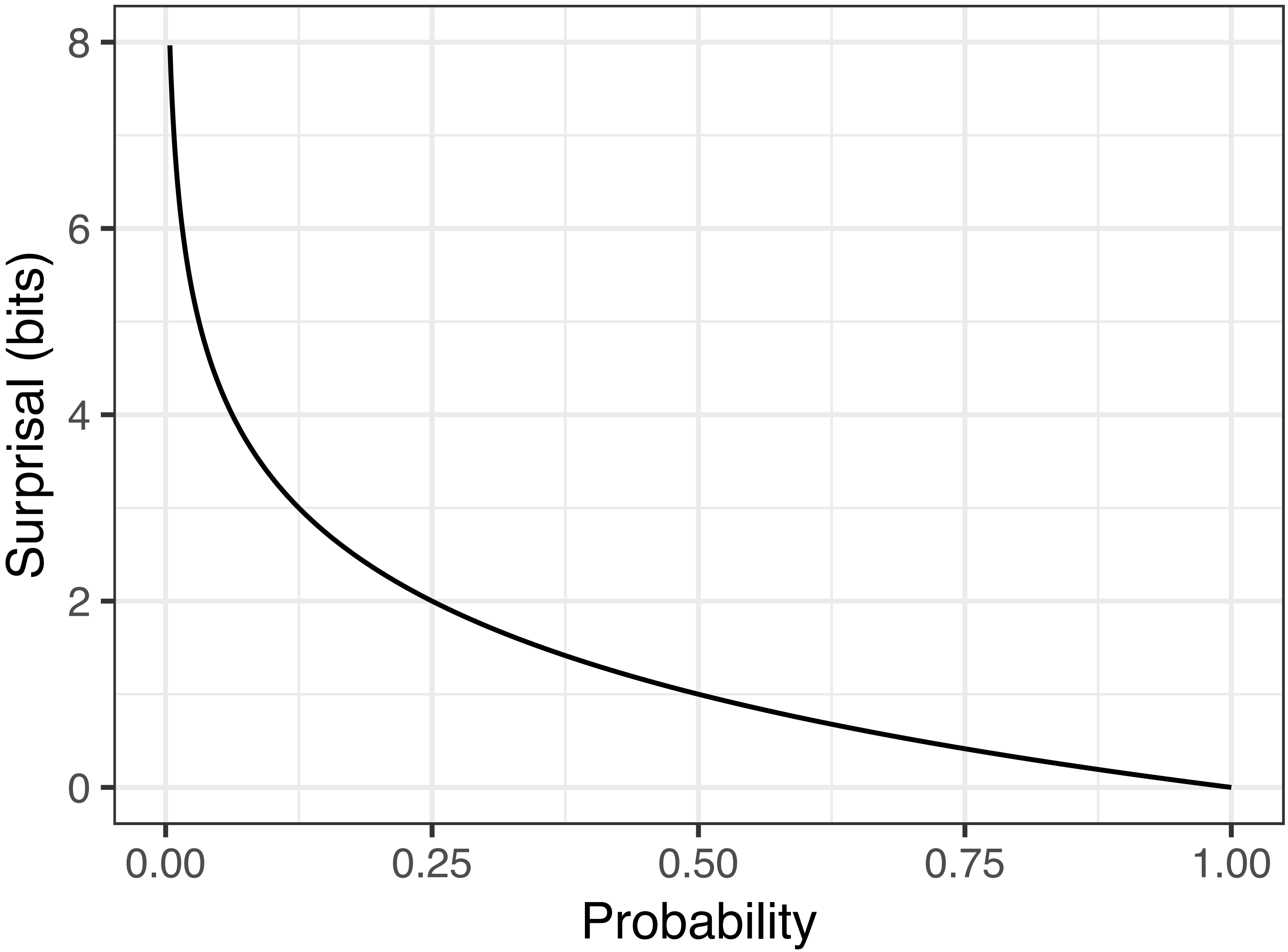
    my brother came inside to… chat? wash? get warm?

    the children went outside to… play

  - Predictable words are read faster (Ehrlich & Rayner, 1981) and have distinctive EEG responses (Kutas & Hillyard 1980)
- Probabilistic grammars give *grammatical expectations*

(*Hale, 2001, NAACL; Levy, 2008, Cognition*)

The surprisal graph



A small PCFG for this sentence type

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

(analysis in Levy, 2013)

A small PCFG for this sentence type

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				

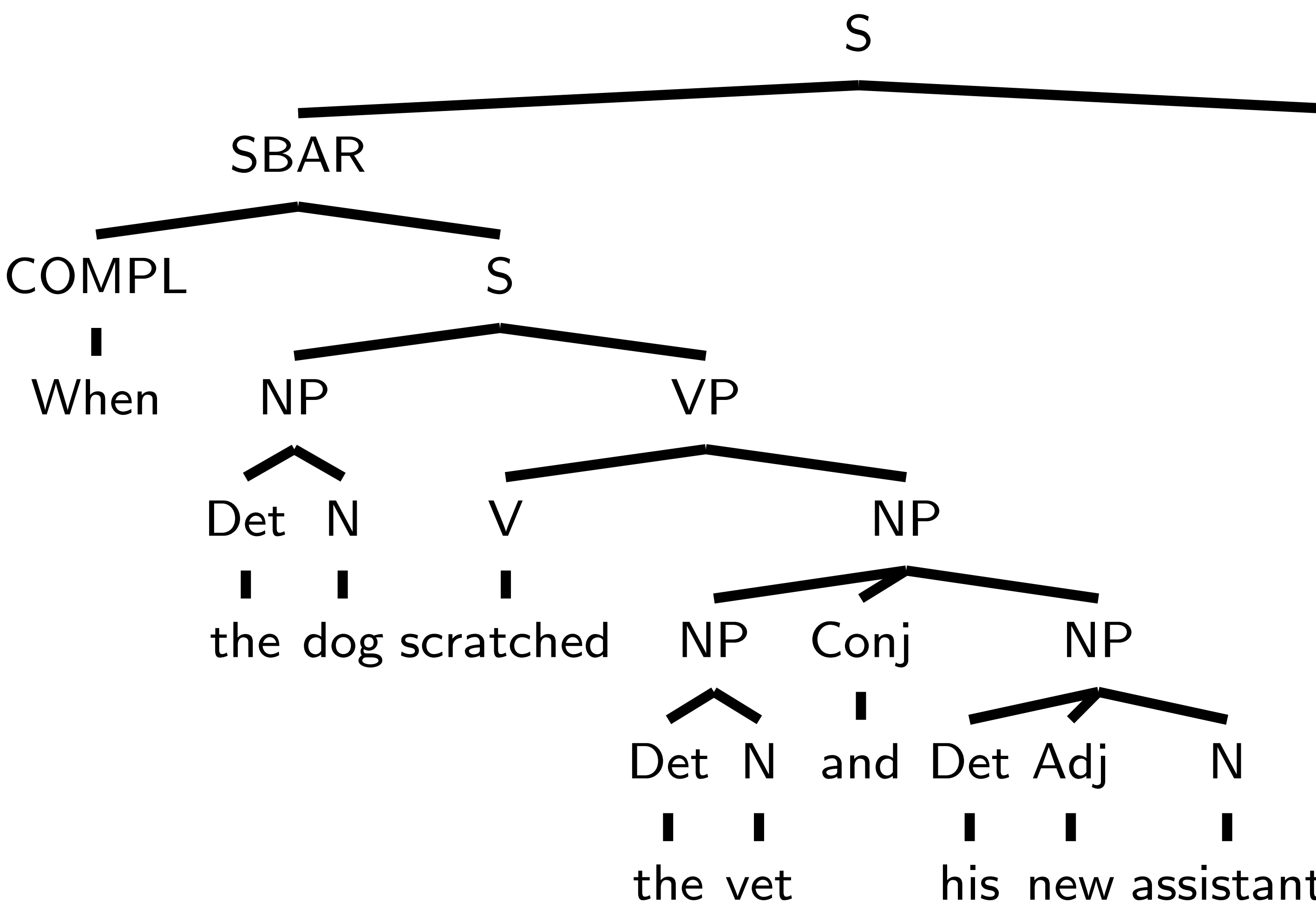
(analysis in Levy, 2013)

Two incremental trees

(analysis in Levy, 2013)

Two incremental trees

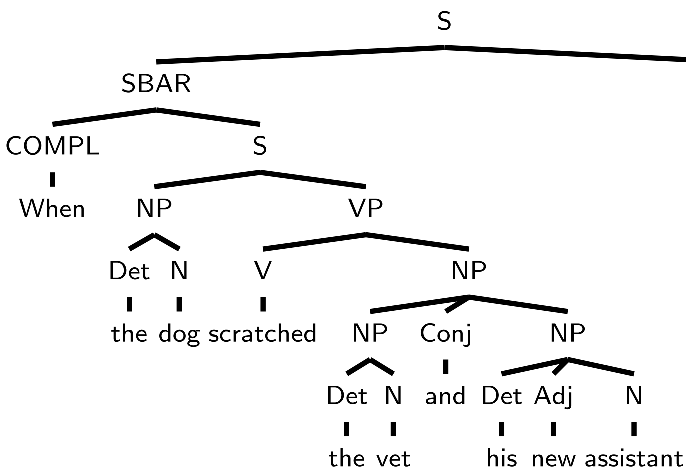
- "Garden-path" analysis:



(analysis in Levy, 2013)

Two incremental trees

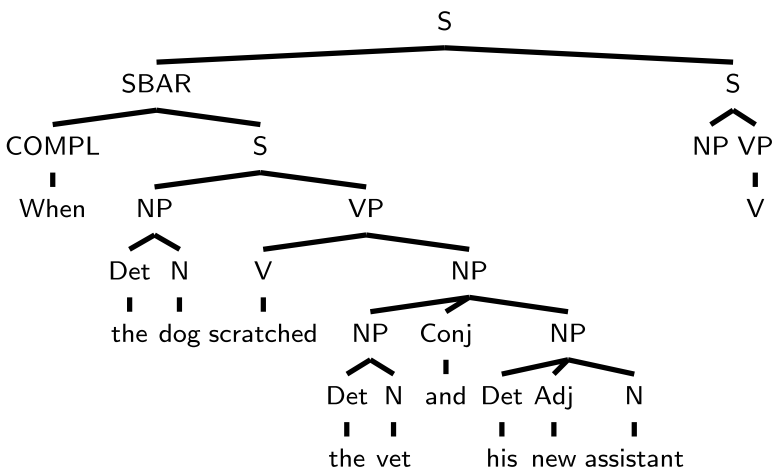
- “Garden-path” analysis:



(analysis in Levy, 2013)

Two incremental trees

- "Garden-path" analysis:

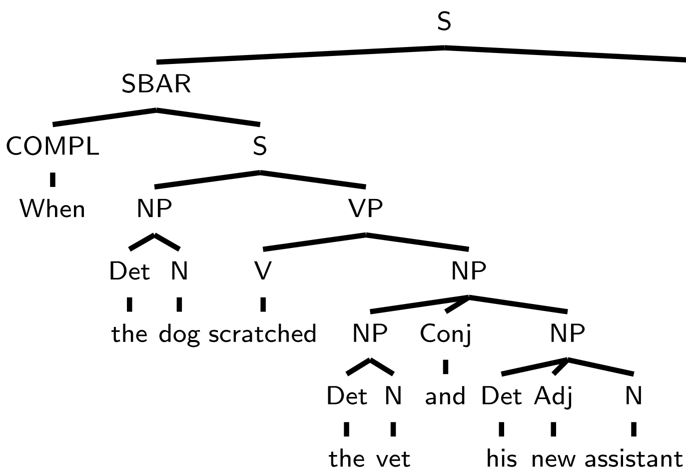


$$P(T|w_{1\dots10}) = 0.826$$

(analysis in Levy, 2013)

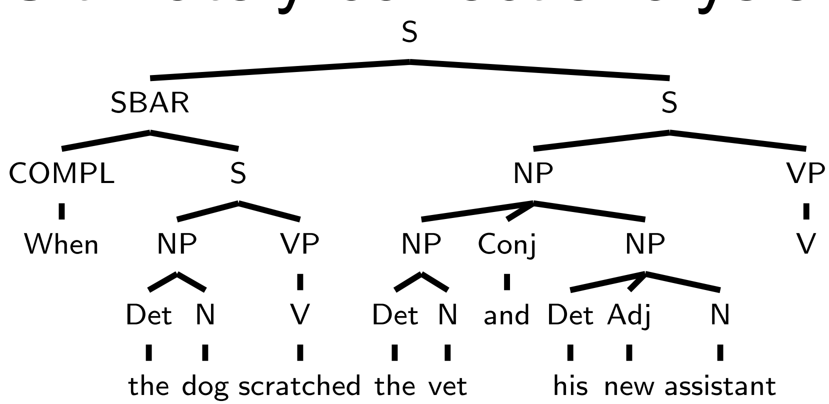
Two incremental trees

- “Garden-path” analysis:



$$P(T|w_{1\ldots10}) = 0.826$$

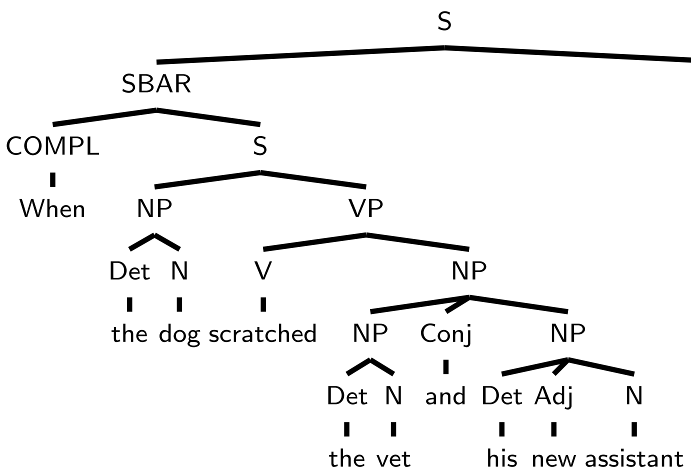
- Ultimately-correct analysis



(analysis in Levy, 2013)

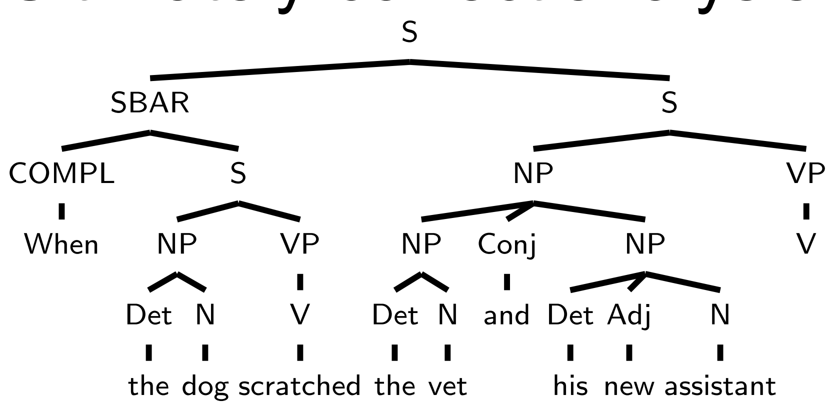
Two incremental trees

- "Garden-path" analysis:



$$P(T|w_{1...10}) = 0.826$$

- Ultimately-correct analysis

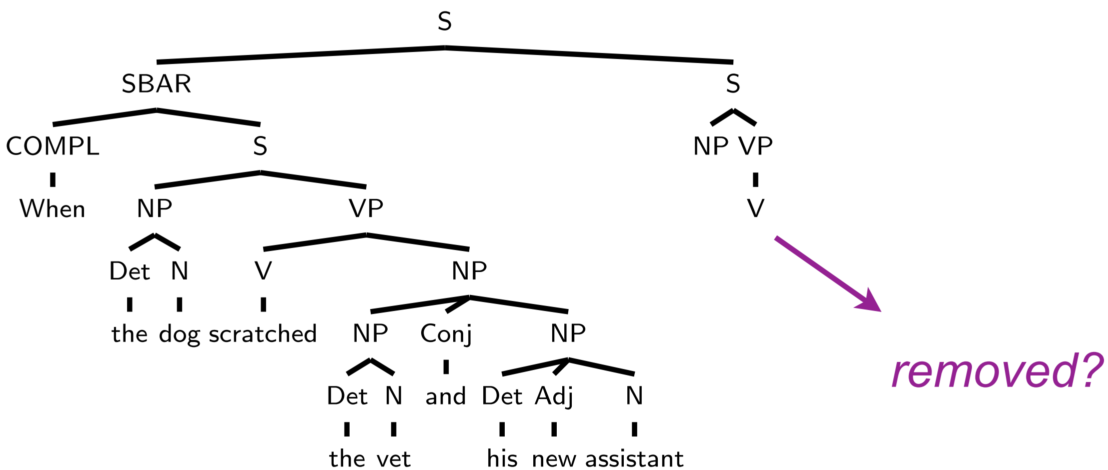


$$P(T|w_{1...10}) = 0.174$$

(analysis in Levy, 2013)

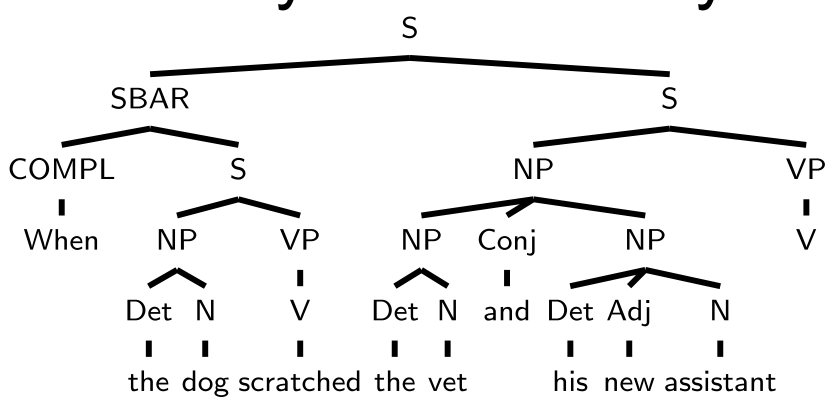
Two incremental trees

- "Garden-path" analysis:



$$P(T|w_{1\ldots10}) = 0.826$$

- Ultimately-correct analysis

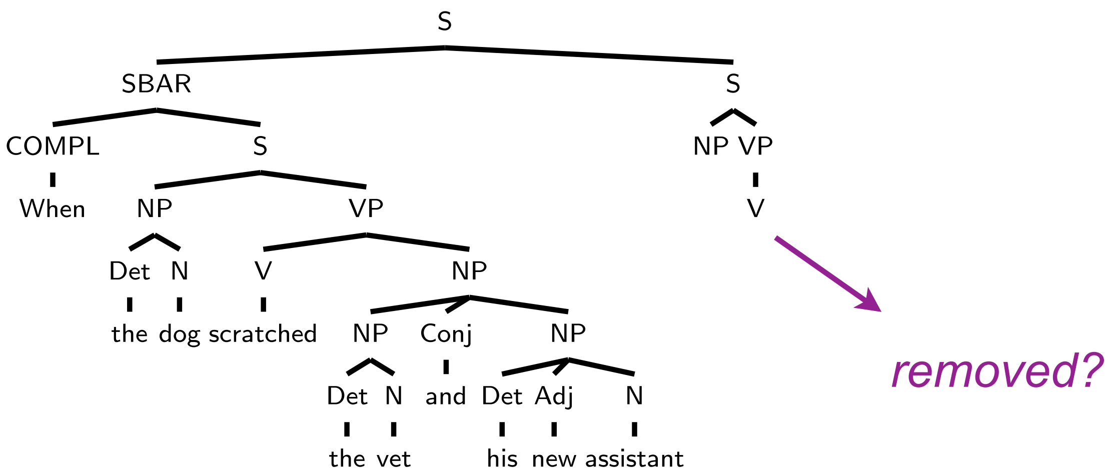


$$P(T|w_{1\ldots10}) = 0.174$$

(analysis in Levy, 2013)

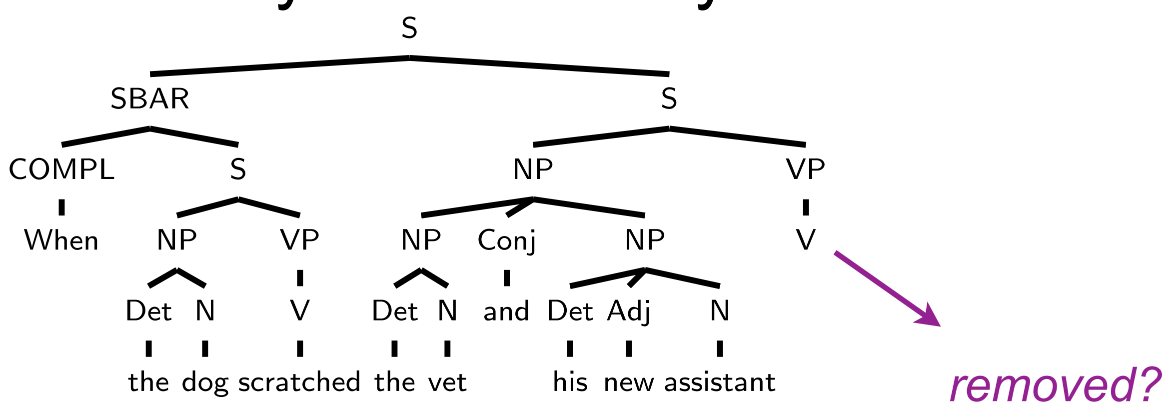
Two incremental trees

- “Garden-path” analysis:



$$P(T|w_{1...10}) = 0.826$$

- Ultimately-correct analysis

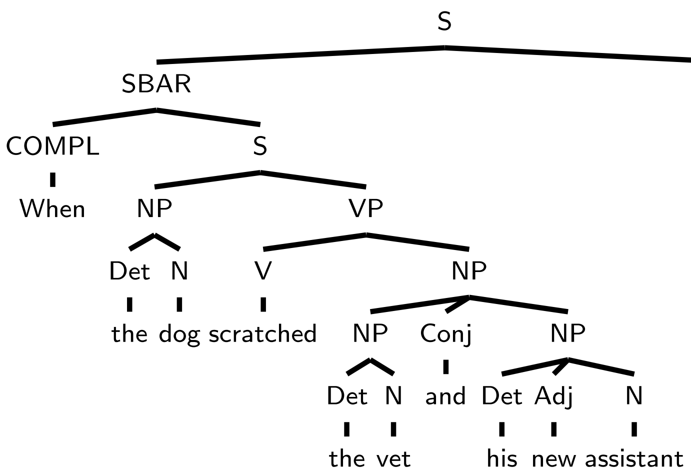


$$P(T|w_{1...10}) = 0.174$$

(analysis in Levy, 2013)

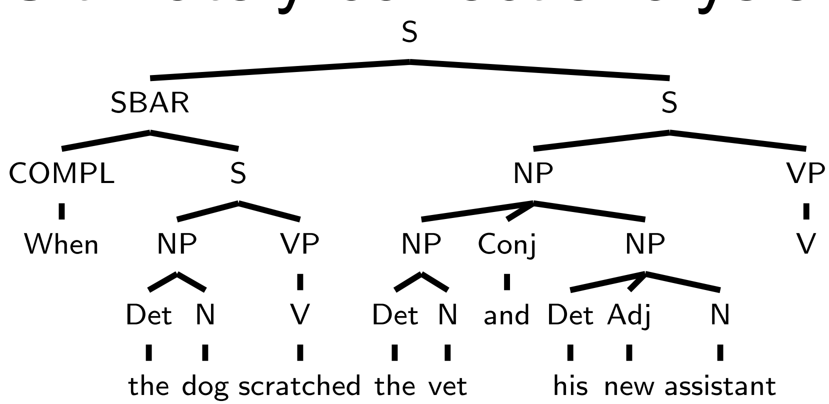
Two incremental trees

- “Garden-path” analysis:



$$P(T|w_{1...10}) = 0.826$$

- Ultimately-correct analysis



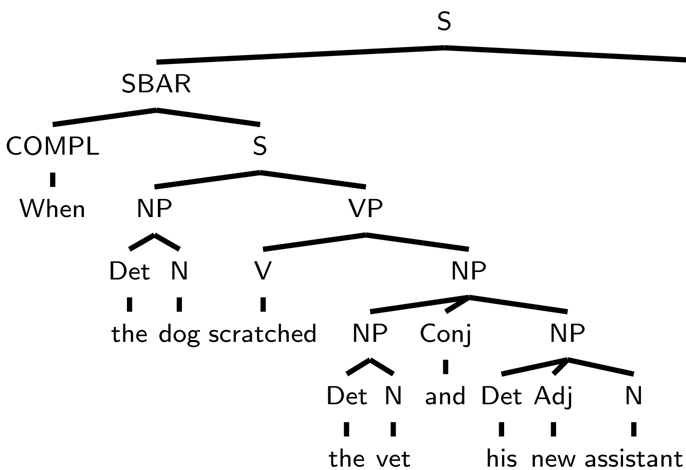
$$P(T|w_{1...10}) = 0.174$$

Disambiguating word probability marginalizes over incremental trees:

(analysis in Levy, 2013)

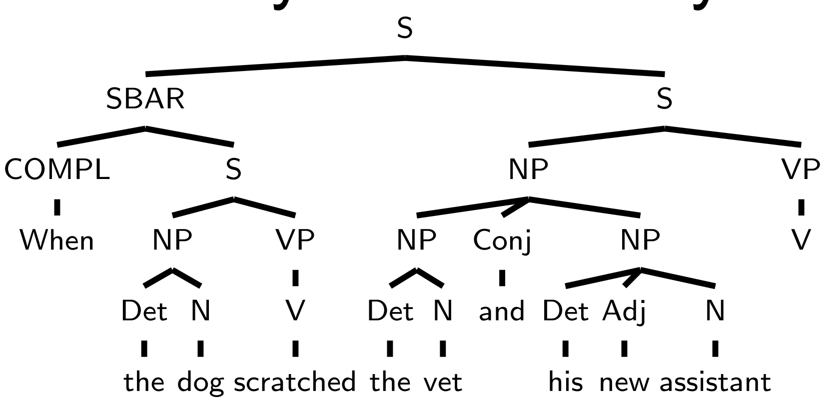
Two incremental trees

- "Garden-path" analysis:



$$P(T|w_{1\ldots10}) = 0.826$$

- Ultimately-correct analysis



$$P(T|w_{1\ldots10}) = 0.174$$

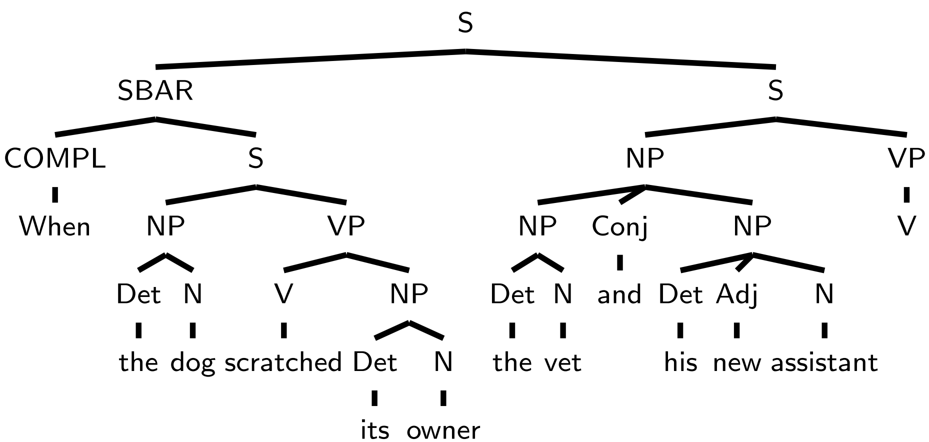
Disambiguating word probability marginalizes over incremental trees:

$$P(\text{removed}|w_{1\ldots10}) = \sum_T P(\text{removed}|T)P(T|w_{1\ldots10})$$
$$= 0 \times 0.826 + 0.25 \times 0.174$$

(analysis in Levy, 2013)

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.

(Staub, 2007)

Sensitivity to verb argument structure

- A superficially similar example:

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

Easier here

(Staub, 2007)

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

(Staub, 2007)

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

Modeling argument-structure sensitivity

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

- The “context-free” assumption doesn’t preclude relaxing probabilistic locality:

(Johnson, 1998; Klein & Manning, 2003)

Modeling argument-structure sensitivity

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

- The “context-free” assumption doesn’t preclude relaxing probabilistic locality:

(Johnson, 1998; Klein & Manning, 2003)

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, 1998; Klein & Manning, 2003)

Result

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

ambiguity onset

ambiguity resolution

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

Surprisal vs. predictability in general

$$\begin{aligned}\text{Surprisal}(w_i) &\equiv \log \frac{1}{P(w_i|\text{CONTEXT})} \\ \left[\approx\right. &\left.\log \frac{1}{P(w_i|w_{1\ldots i-1})}\right]\end{aligned}$$

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

(Smith & Levy, 2013)

Estimating probability/time curve shape

Estimating probability/time curve shape

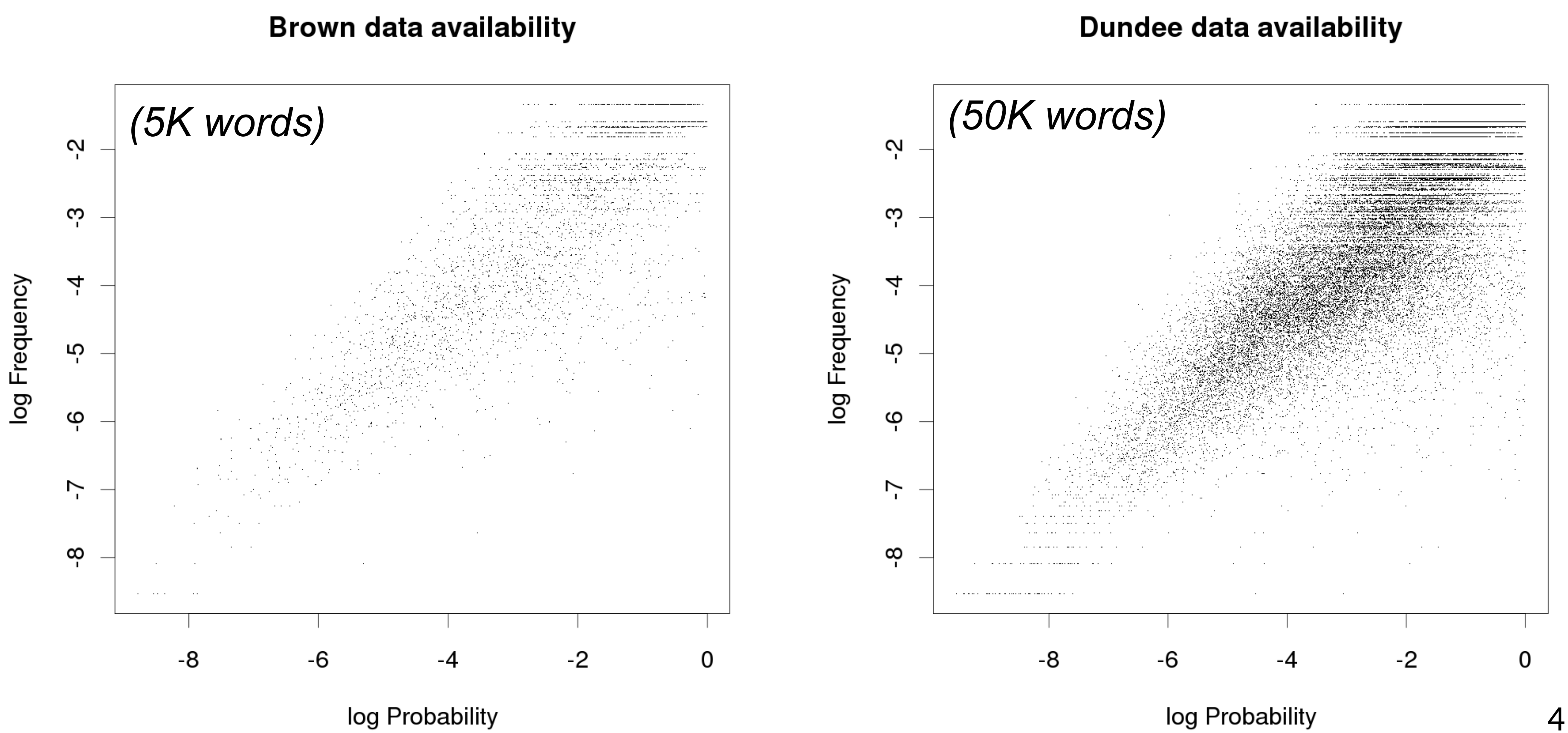
- As a proxy for "processing difficulty," reading time in two different methods: self-paced reading & eye-tracking

Estimating probability/time curve shape

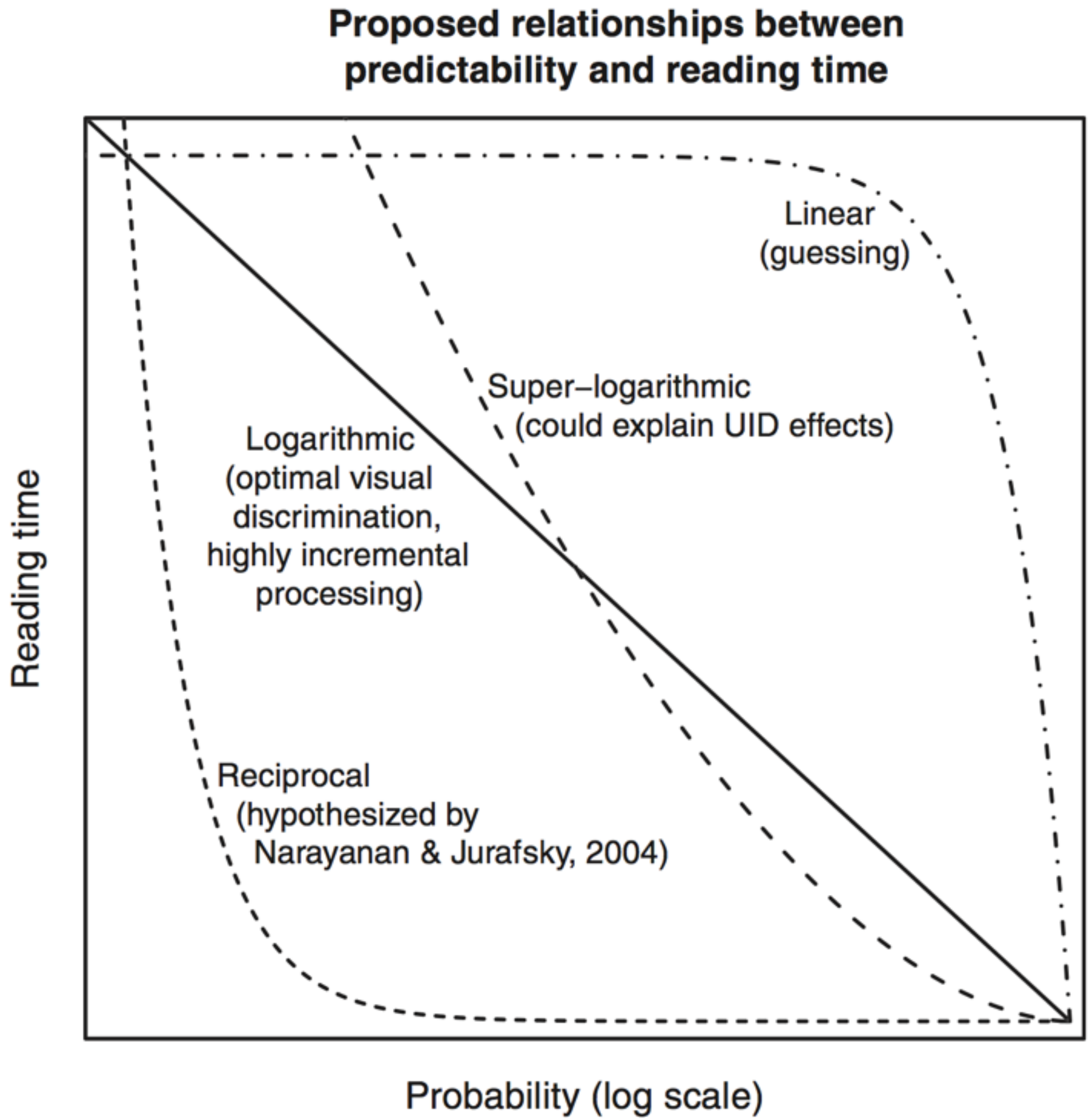
- As a proxy for “processing difficulty,” reading time in two different methods: self-paced reading & eye-tracking
- Challenge: we need big data to estimate curve shape, but probability correlated with confounding variables

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

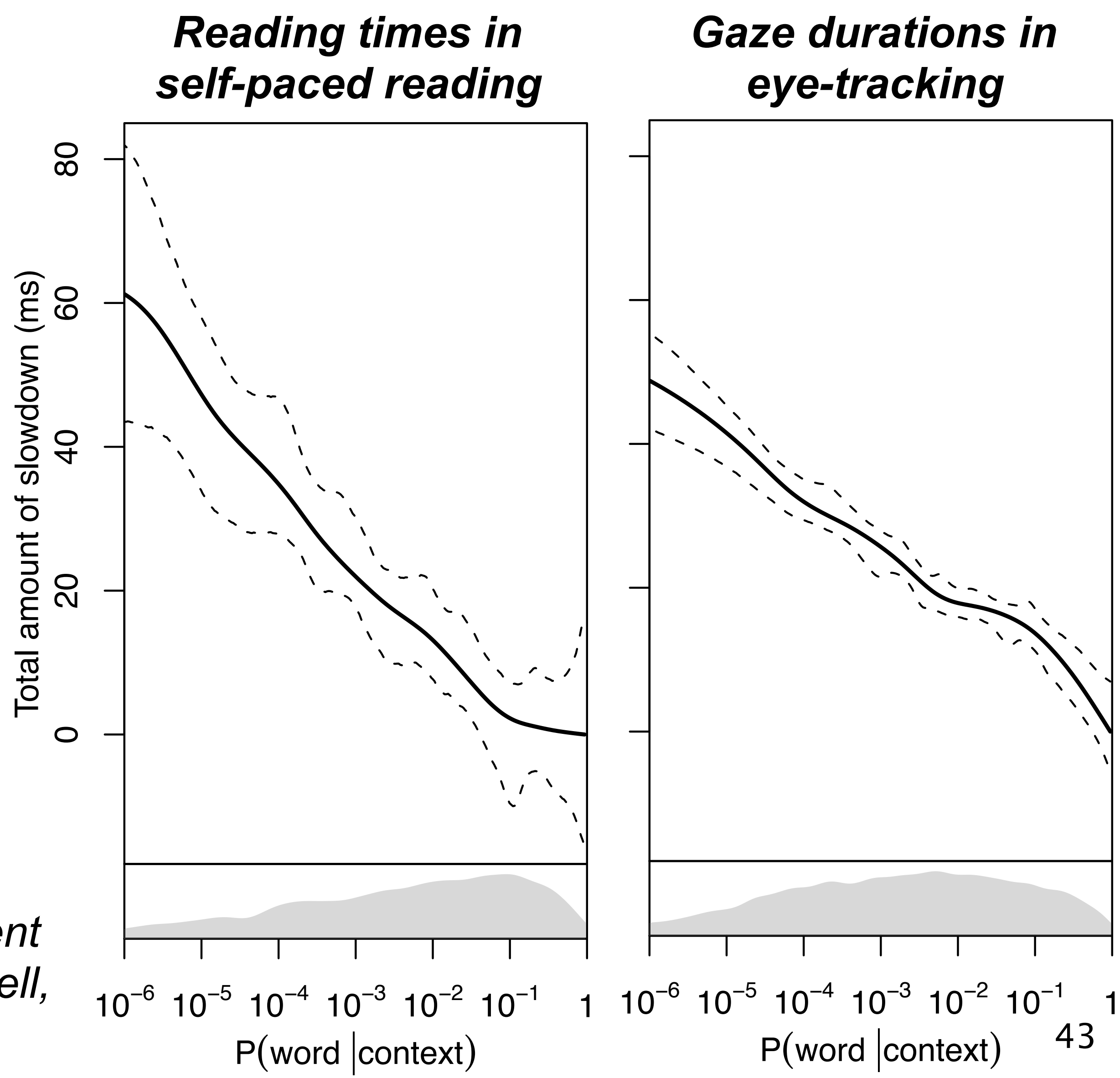


Hypothesized curve shapes



Estimating probability/time curve shape

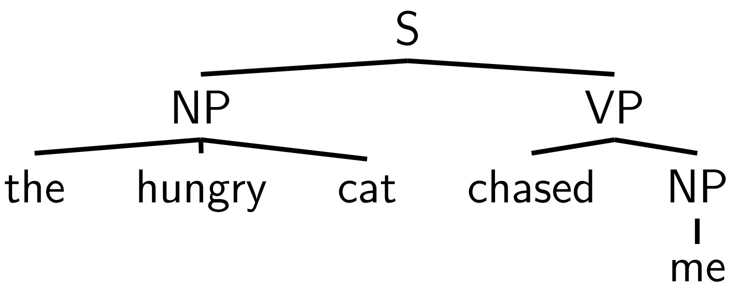
- GAM regression: total contribution of word (trigram) probability to RT near-linear over 6 orders of magnitude!



(Smith & Levy, 2013; more recent validation by Goodkind & Bicknell, 2018)

Integration with deep learning

- Humans condition extremely flexibly on context
- Goal: **symbolic grammars + neural generatization**
- Enabling step: **action sequence** for structure building



(S (NP the hungry cat) (VP chased (NP me)))

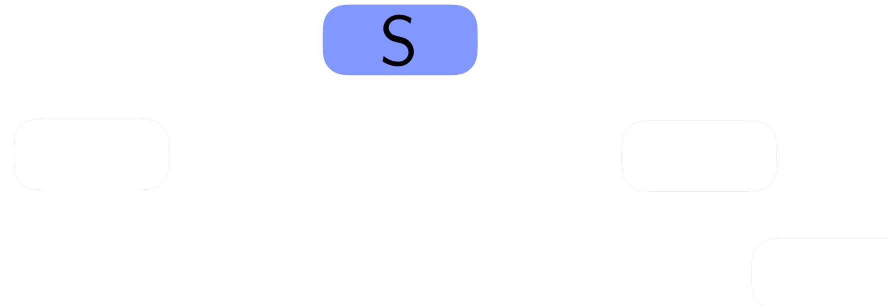
Action	**Meaning**	**String gloss**
NT(X)	Push a new **open** non-terminal on top of the **stack**	**(X**
Gen(*w*)	Generate word *w* as a terminal node and put it on top of the stack (as a **closed** node)	*w*
REDUCE	Pop **closed** nodes $N_{1\ldots i-1}$ from the top of the **stack** until encountering **open** node N_i; close N_i	**)**
END	Finish parsing (iff the sole stack element is a closed S)	**n/a**

(S (NP the hungry cat) (VP chased (NP me)))

(S (NP the hungry cat) (VP chased (NP me)))

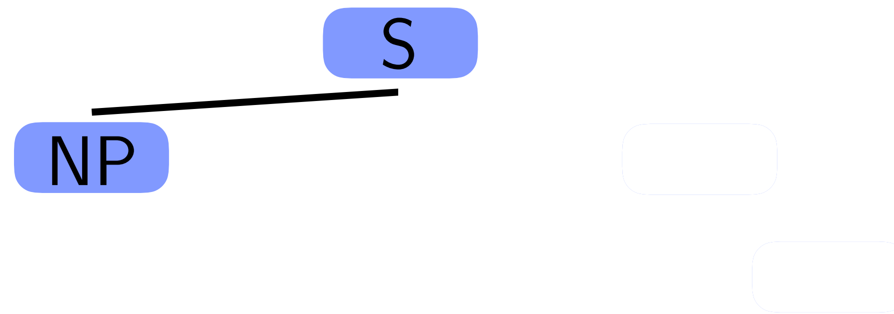
Action Stack

(S (NP the hungry cat) (VP chased (NP me)))



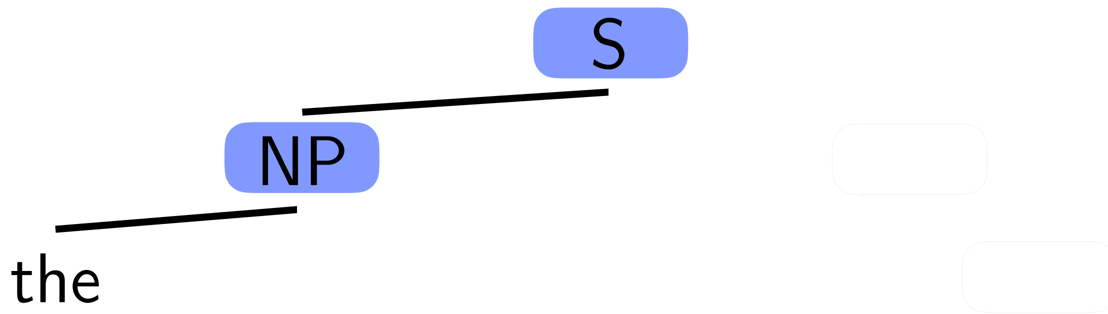
Action	Stack
NT(S)	(S

(S (NP the hungry cat) (VP chased (NP me)))



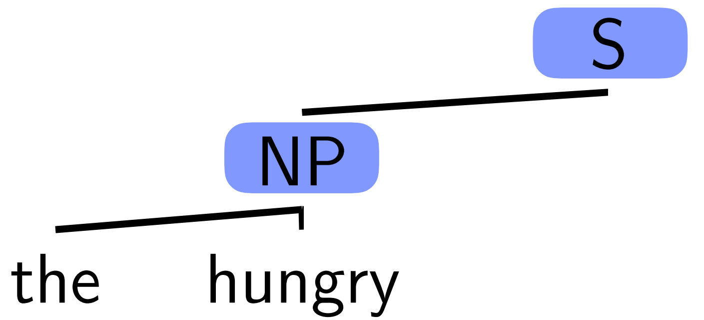
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP

(S (NP the hungry cat) (VP chased (NP me)))



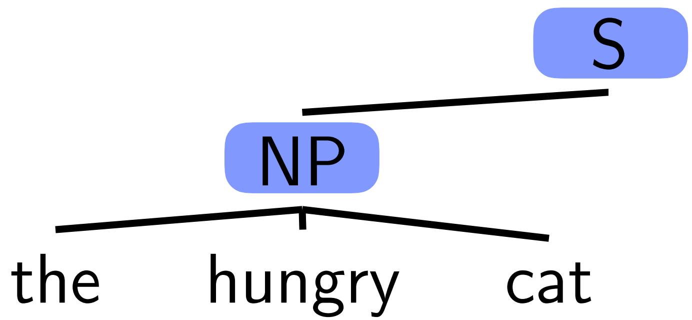
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the

(S (NP the hungry cat) (VP chased (NP me)))



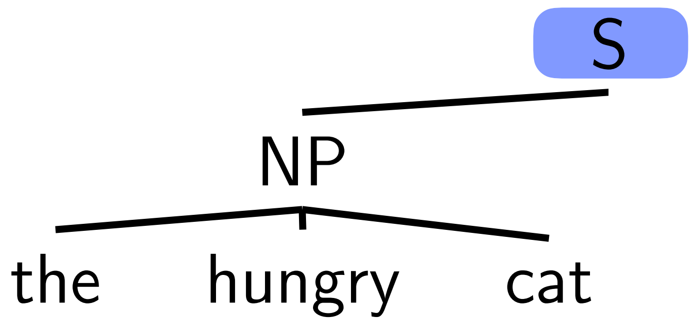
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry

(S (NP the hungry cat) (VP chased (NP me)))



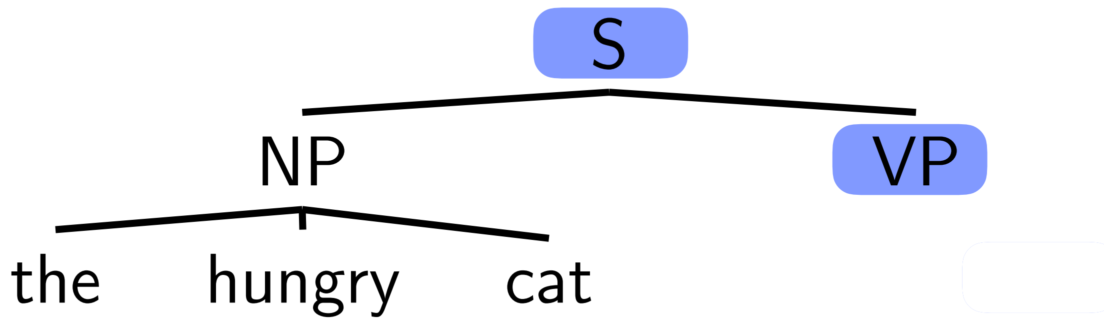
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat

(S (NP the hungry cat) (VP chased (NP me)))



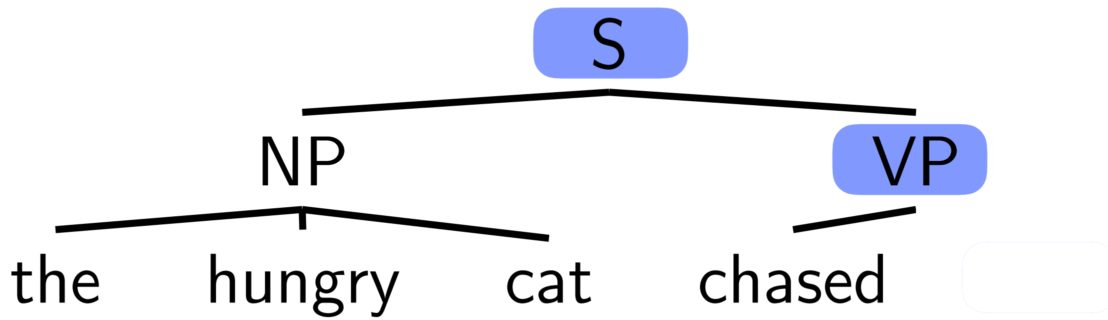
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)

(S (NP the hungry cat) (VP chased (NP me)))



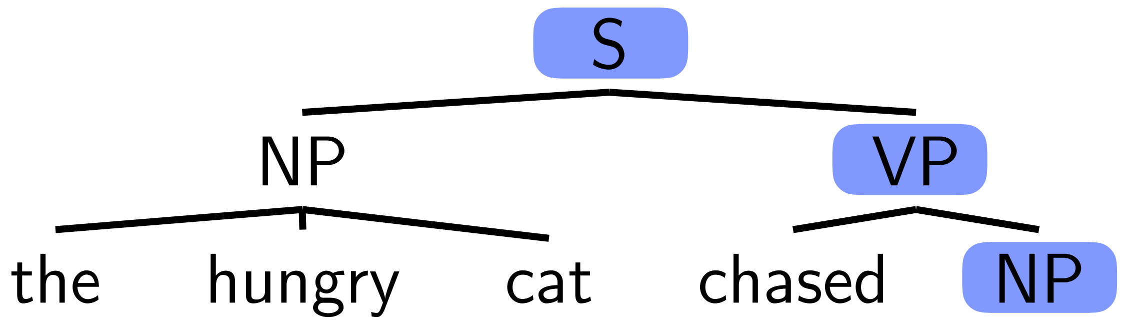
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP

(S (NP the hungry cat) (VP chased (NP me)))



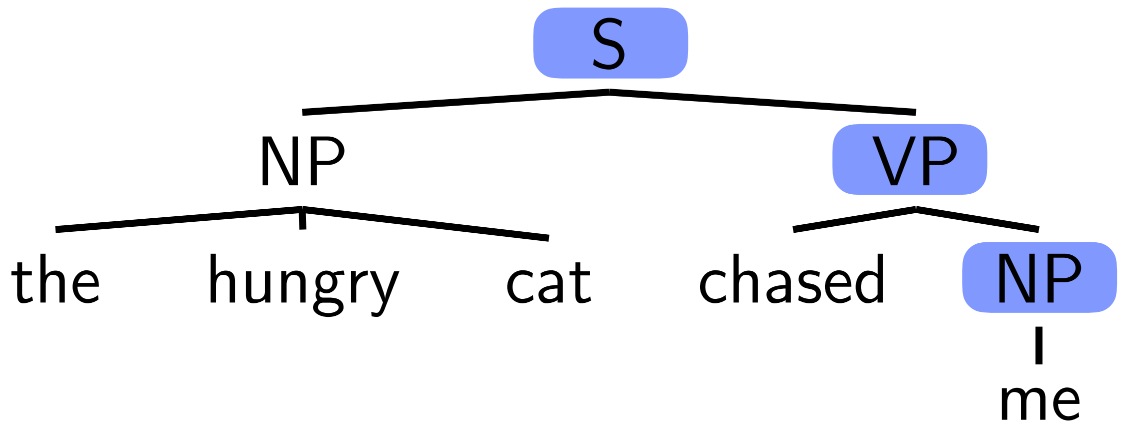
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased

(S (NP the hungry cat) (VP chased (NP me)))



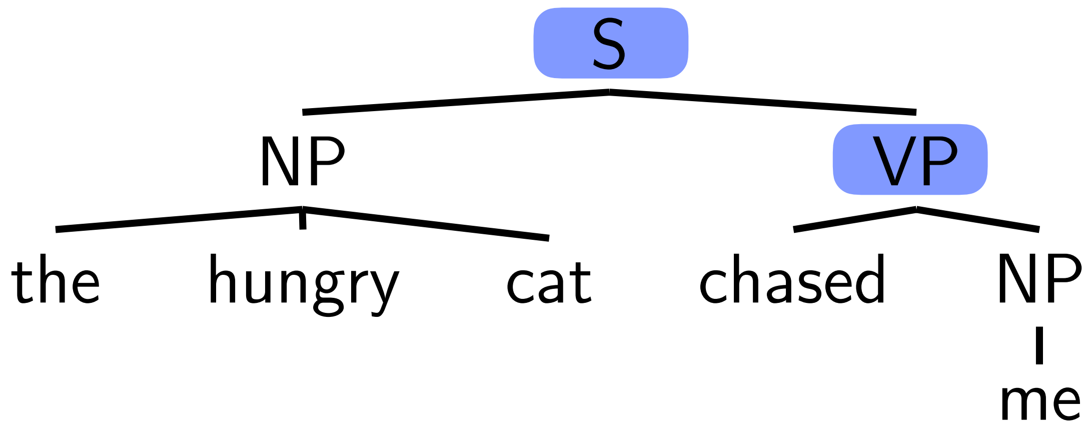
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP

(S (NP the hungry cat) (VP chased (NP me)))



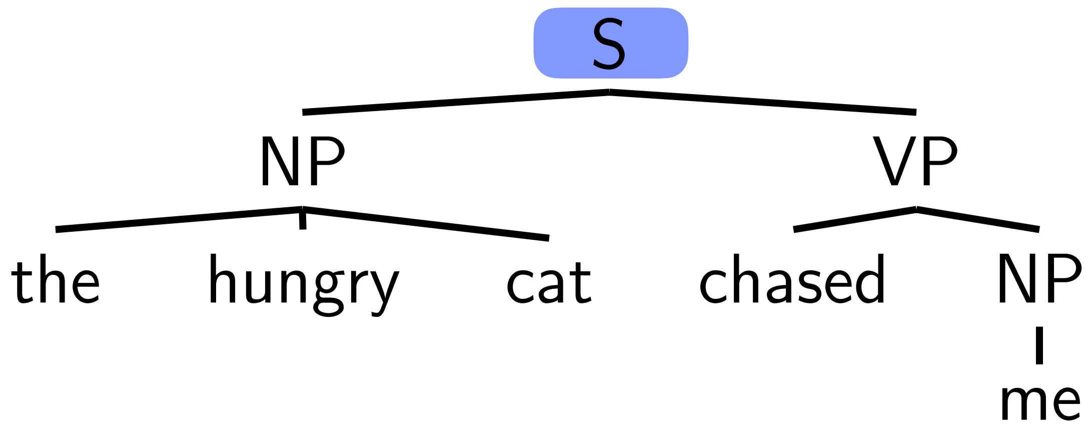
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me

(S (NP the hungry cat) (VP chased (NP me)))



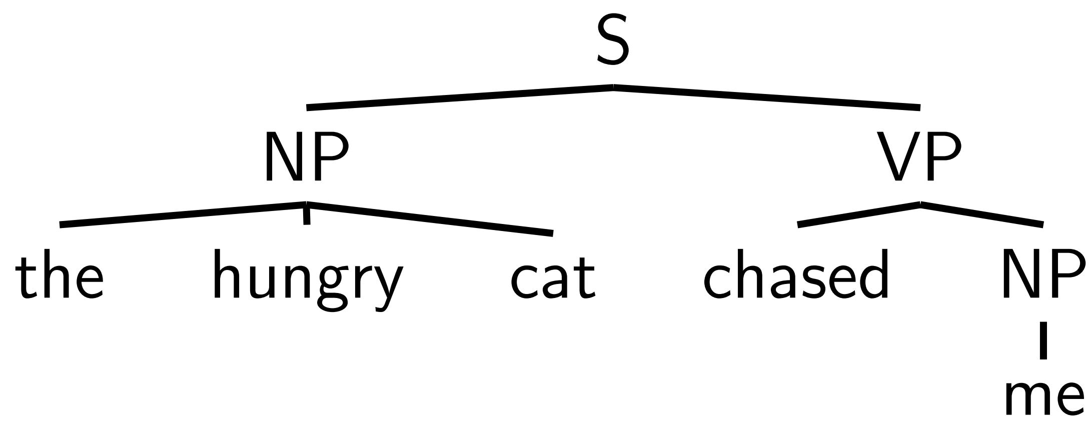
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me
REDUCE	(S | (NP the hungry cat) | (VP | chased | (NP me)

(S (NP the hungry cat) (VP chased (NP me)))



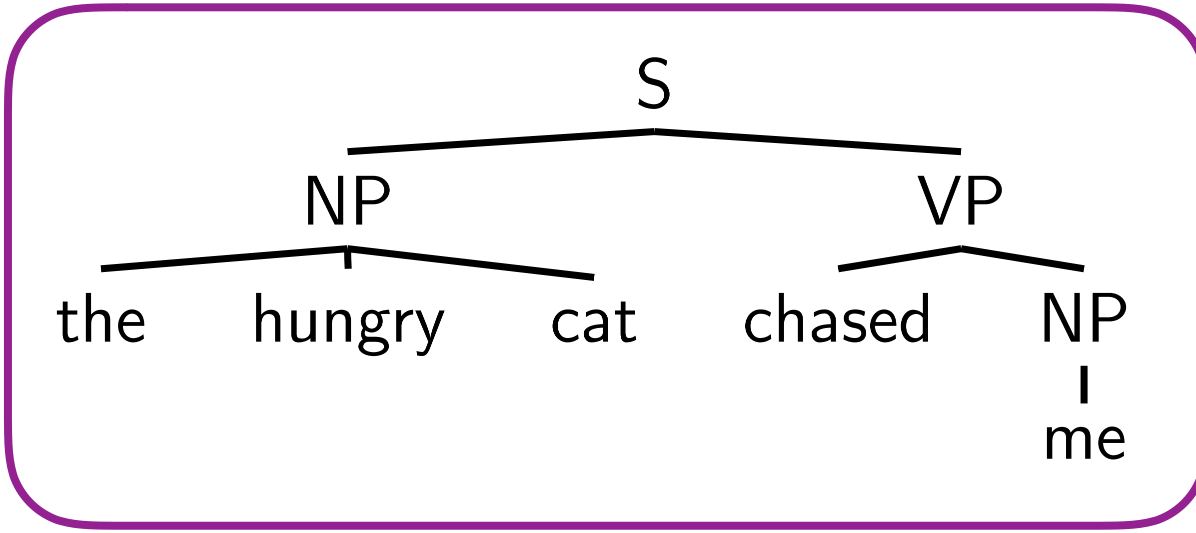
<table>
<tr><th>Action</th><th>Stack</th></tr>
<tr><td>NT(S)</td><td>(S</td></tr>
<tr><td>NT(NP)</td><td>(S | (NP</td></tr>
<tr><td>Gen(the)</td><td>(S | (NP | the</td></tr>
<tr><td>Gen(hungry)</td><td>(S | (NP | the | hungry</td></tr>
<tr><td>Gen(cat)</td><td>(S | (NP | the | hungry | cat</td></tr>
<tr><td>REDUCE</td><td>(S | (NP the hungry cat)</td></tr>
<tr><td>NT(VP)</td><td>(S | (NP the hungry cat) | (VP</td></tr>
<tr><td>Gen(chased)</td><td>(S | (NP the hungry cat) | (VP | chased</td></tr>
<tr><td>NT(NP)</td><td>(S | (NP the hungry cat) | (VP | chased | (NP</td></tr>
<tr><td>Gen(me)</td><td>(S | (NP the hungry cat) | (VP | chased | (NP | me</td></tr>
<tr><td>REDUCE</td><td>(S | (NP the hungry cat) | (VP | chased | (NP me)</td></tr>
<tr><td>REDUCE</td><td>(S | (NP the hungry cat) | (VP chased (NP me))</td></tr>
</table>

(S (NP the hungry cat) (VP chased (NP me)))



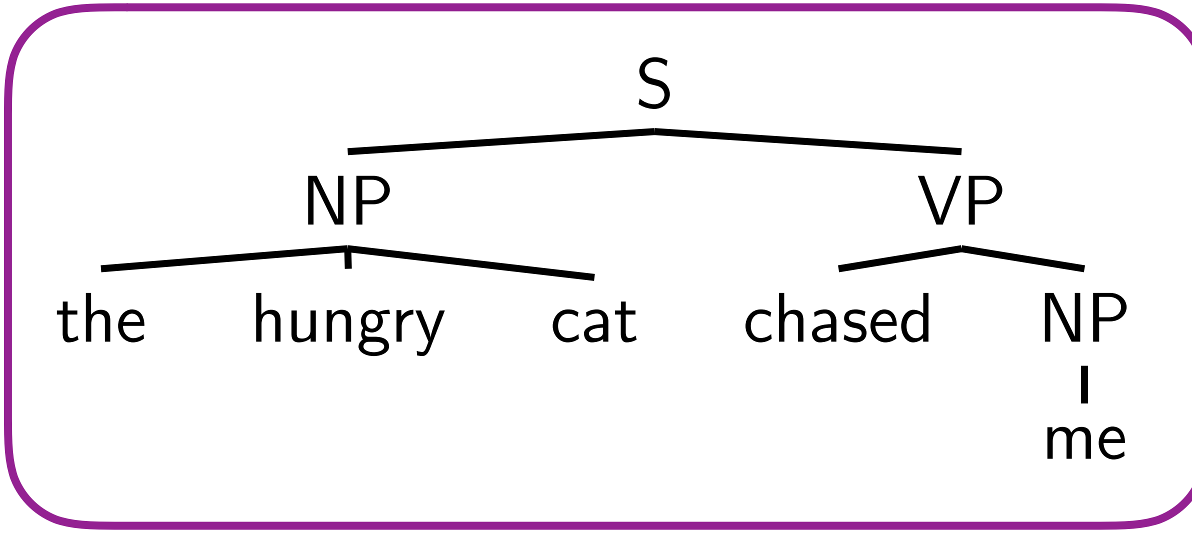
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me
REDUCE	(S | (NP the hungry cat) | (VP | chased | (NP me)
REDUCE	(S | (NP the hungry cat) | (VP chased (NP me))
REDUCE	(S (NP the hungry cat) (VP chased (NP me)))

(S (NP the hungry cat) (VP chased (NP me)))



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me
REDUCE	(S | (NP the hungry cat) | (VP | chased | (NP me)
REDUCE	(S | (NP the hungry cat) | (VP chased (NP me))
REDUCE	(S (NP the hungry cat) (VP chased (NP me)))
END	

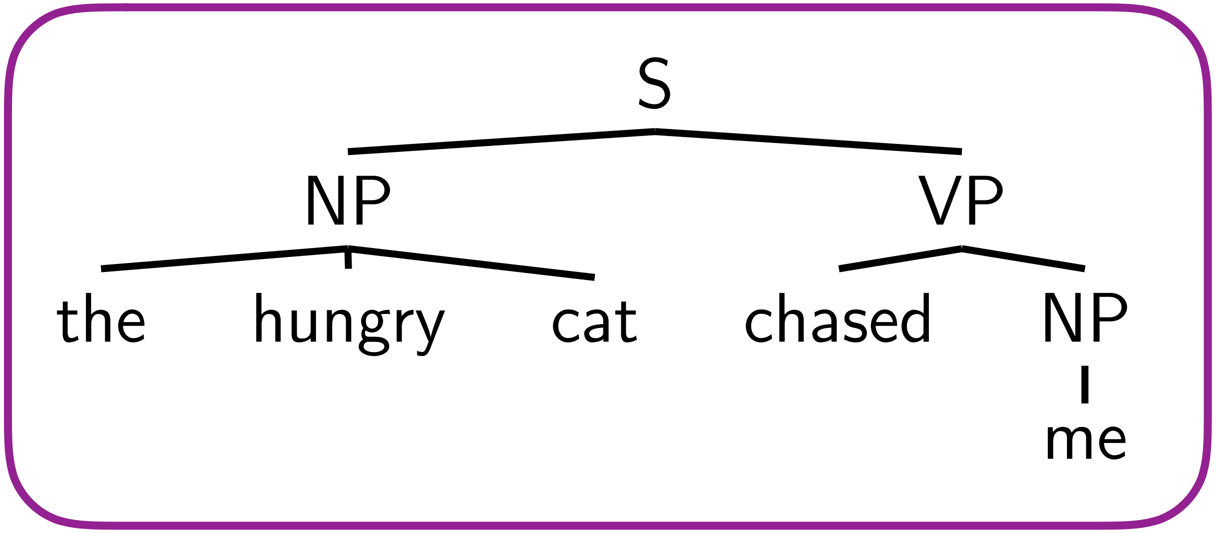
(S (NP the hungry cat) (VP chased (NP me)))



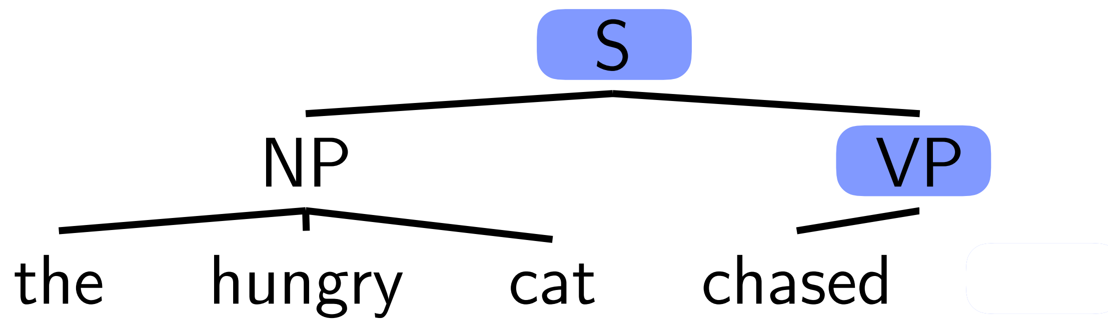
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me
REDUCE	(S | (NP the hungry cat) | (VP | chased | (NP me)
REDUCE	(S | (NP the hungry cat) | (VP chased (NP me))
REDUCE	(S (NP the hungry cat) (VP chased (NP me)))
END	

(S (NP the hungry cat) (VP chased (NP me)))

If we put a conditional probability distribution on actions, we have a probabilistic grammar!

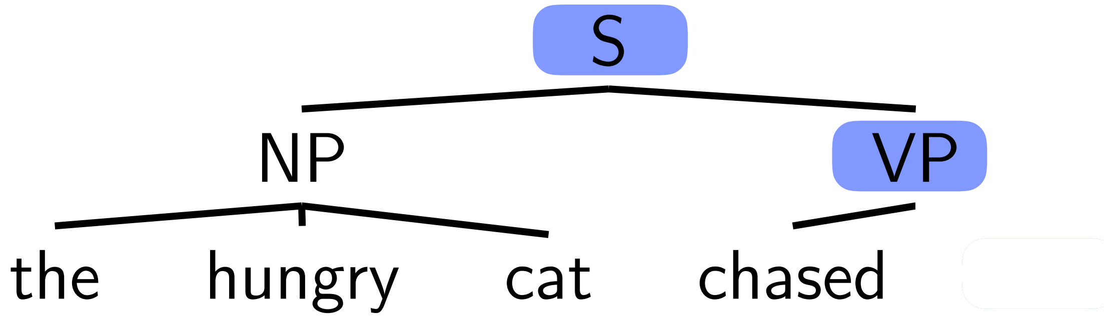


Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
NT(NP)	(S | (NP the hungry cat) | (VP | chased | (NP
Gen(me)	(S | (NP the hungry cat) | (VP | chased | (NP | me
REDUCE	(S | (NP the hungry cat) | (VP | chased | (NP me)
REDUCE	(S | (NP the hungry cat) | (VP chased (NP me))
REDUCE	(S (NP the hungry cat) (VP chased (NP me)))
END	



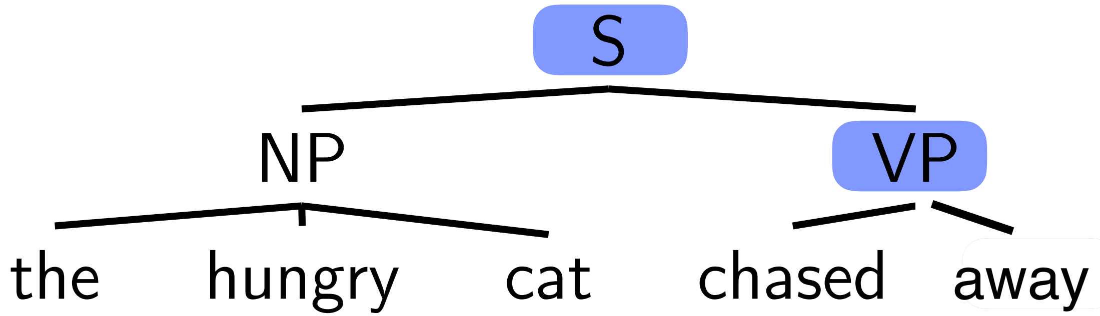
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



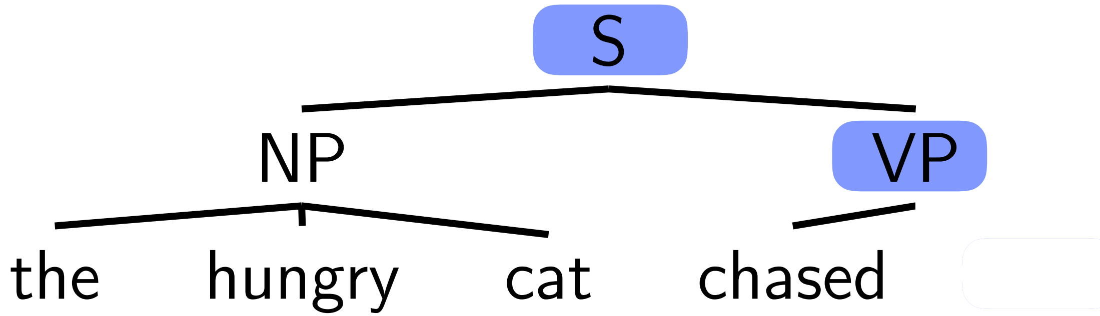
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	
Gen(away)	

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



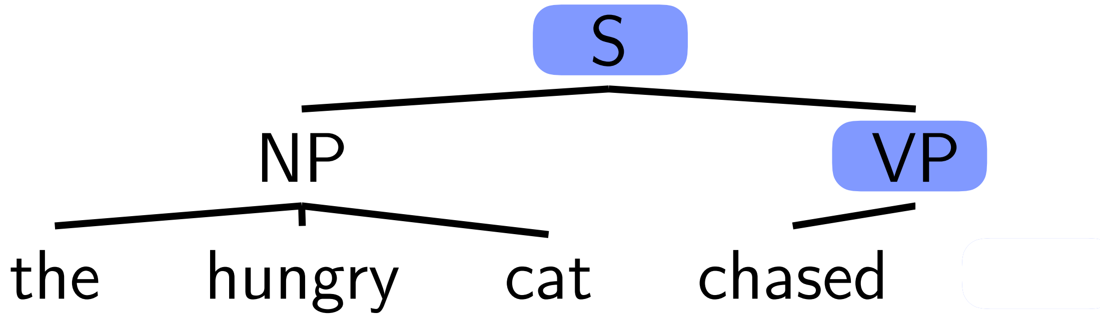
Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	
Gen(away)	

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	
Gen(away)	

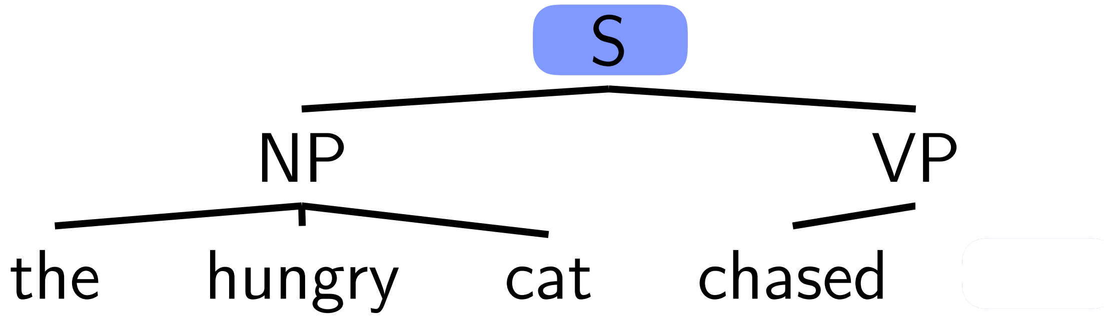
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE

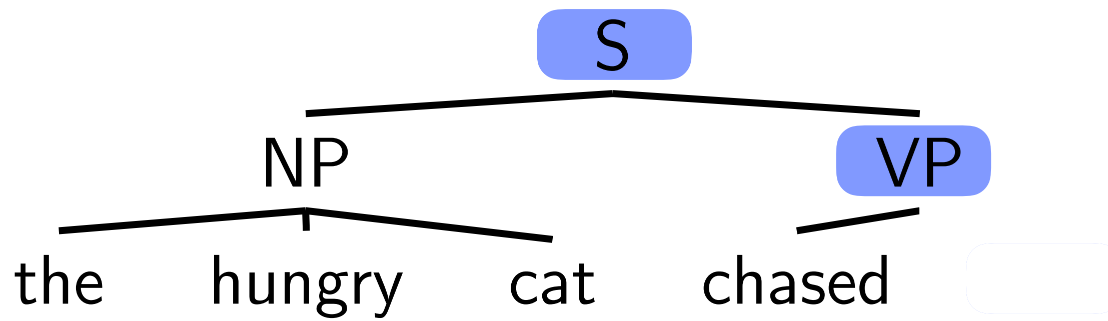
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE

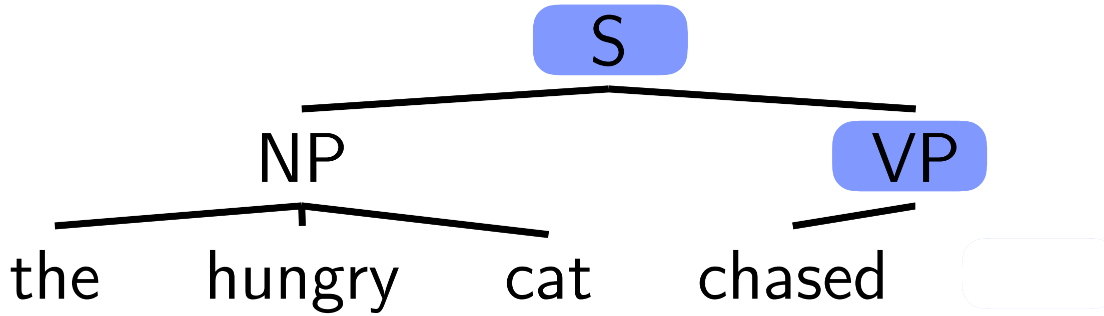
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE

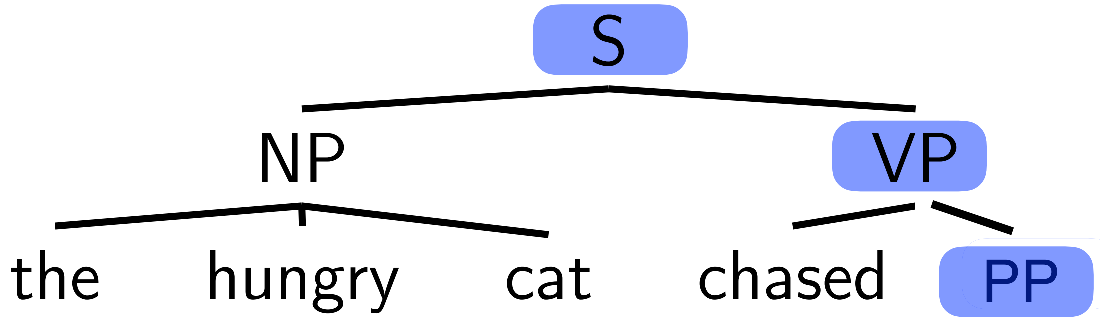
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP)

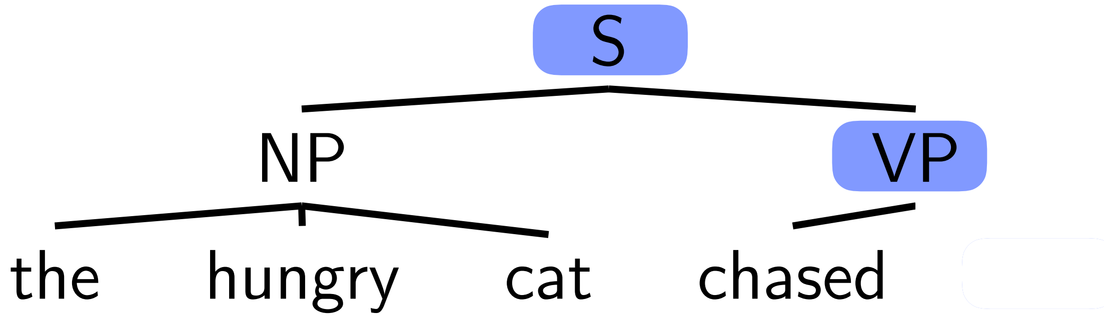
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP)

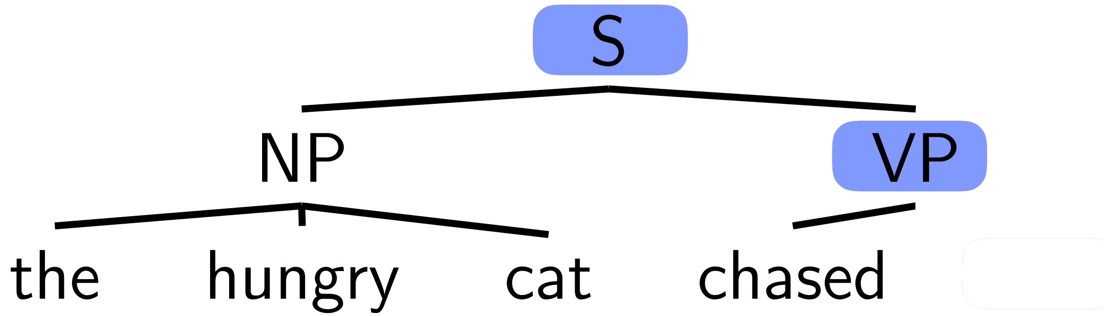
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP)

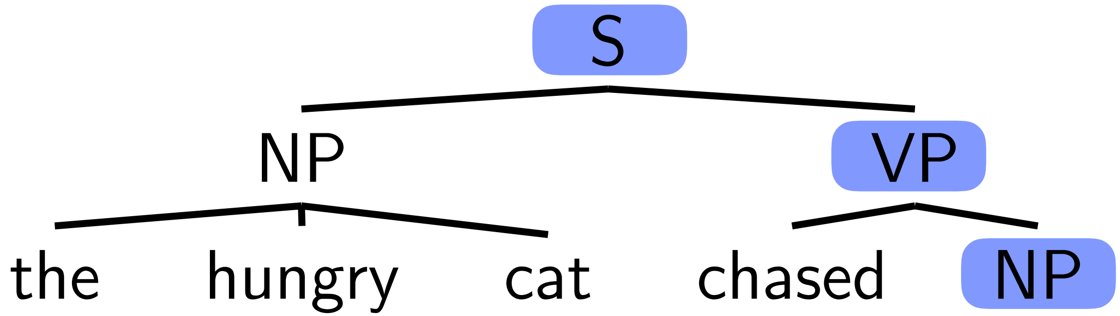
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP) NT(NP)

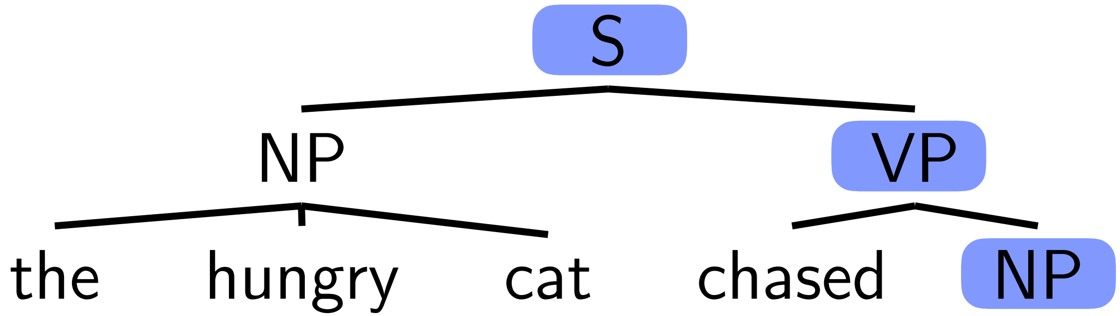
(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP) NT(NP)

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)

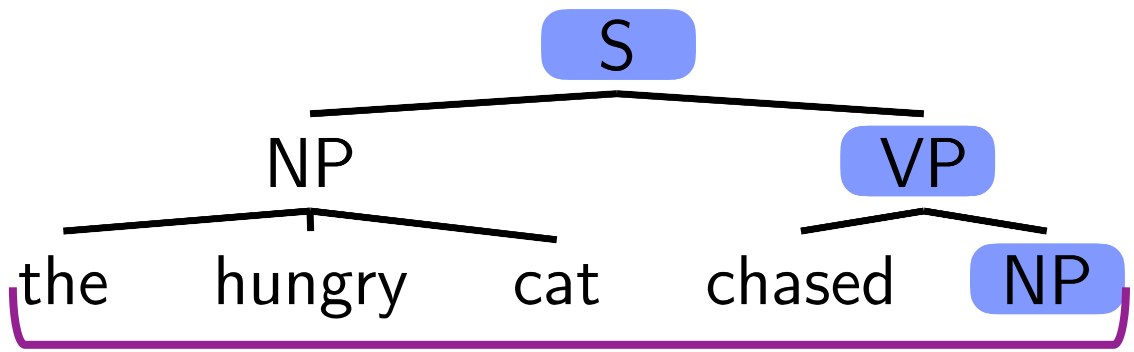


Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

Gen(away) REDUCE NT(PP) NT(NP)

Knowledge characterization: P(action|context)

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)

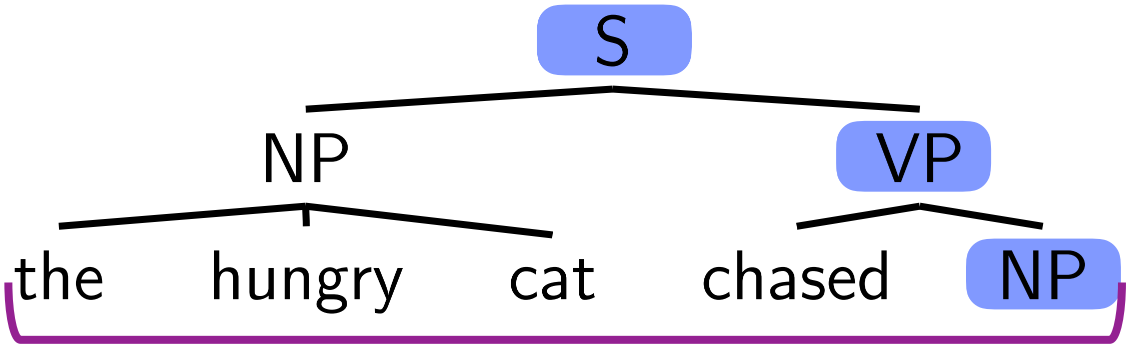


Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

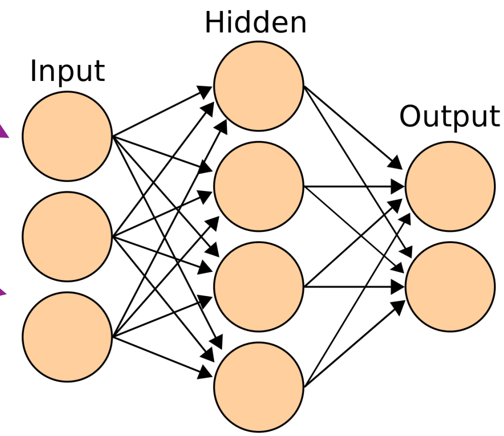
Gen(away) REDUCE NT(PP) NT(NP)

Knowledge characterization: P(action|context)

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)



Action	Stack
NT(S)	(S
NT(NP)	(S | (NP
Gen(the)	(S | (NP | the
Gen(hungry)	(S | (NP | the | hungry
Gen(cat)	(S | (NP | the | hungry | cat
REDUCE	(S | (NP the hungry cat)
NT(VP)	(S | (NP the hungry cat) | (VP
Gen(chased)	(S | (NP the hungry cat) | (VP | chased
???	

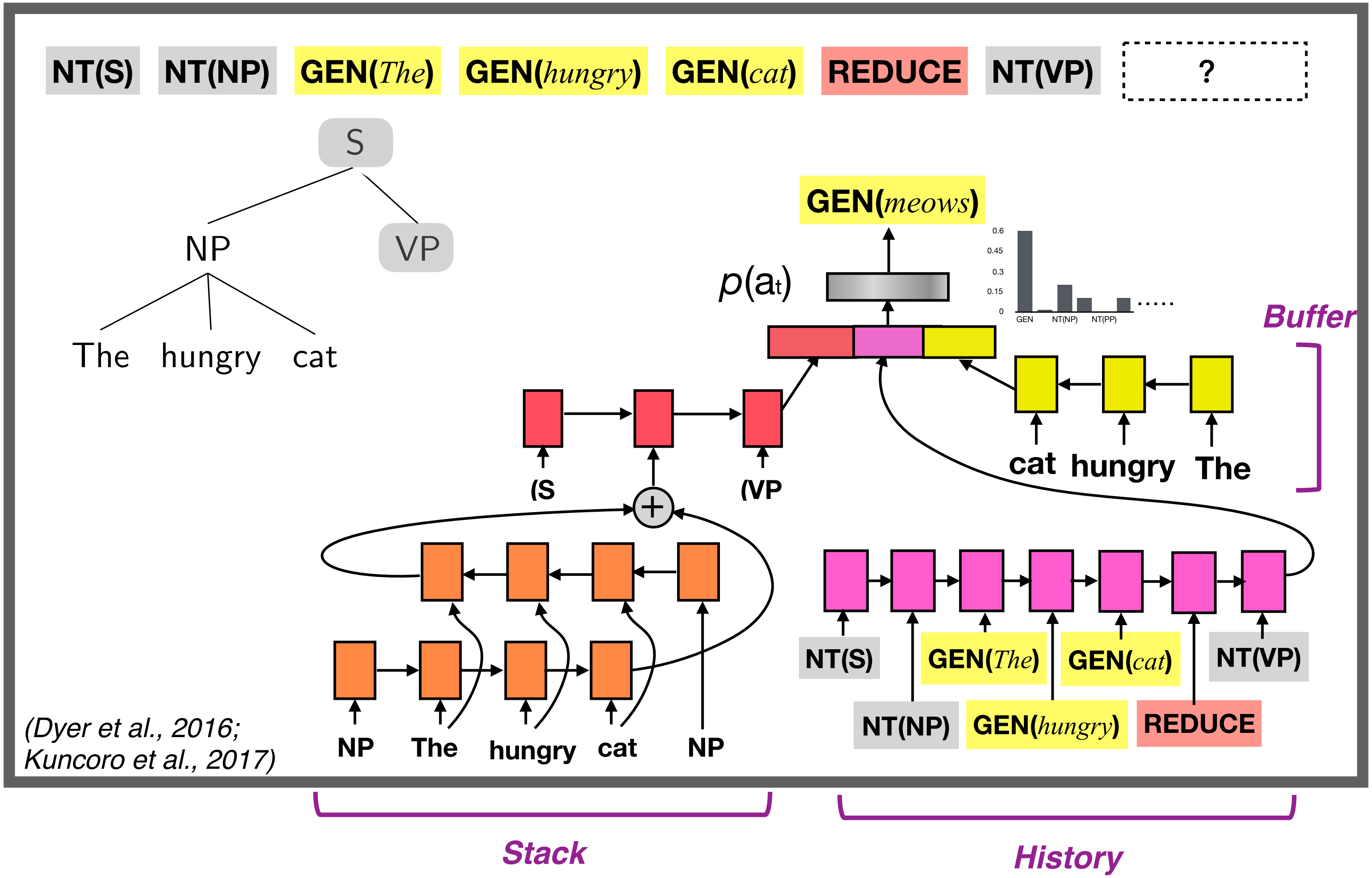


Gen(away) REDUCE NT(PP) NT(NP)

Knowledge characterization: P(action|context)

(Henderson, 2003; Dyer et al. 2016; Kuncoro et al., 2017)

Recurrent Neural Network Grammars (RNNGs)



Evidence of human-like language processing:

Kuncoro et al., 2018 (ACL)

Hale et al., 2018 (ACL)

Futrell et al., 2019 (NAACL)

Wilcox et al., 2019 (NAACL)

An inferential challenge

```
(S (NP I ) (VP saw
```

An inferential challenge

```
(S (NP I ) (VP saw (NP the
```

An inferential challenge

`(S (NP I ) (VP saw (NP the`

I saw the child

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child’s dog*

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child's dog*
`(S (NP I ) (VP saw (S (NP the`	*I saw the child leave*

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child’s dog*
`(S (NP I ) (VP saw (S (NP the`	*I saw the child leave*
`(S (NP I ) (VP saw (S (NP (NP the`	*I saw the child’s dog leave*

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child’s dog*
`(S (NP I ) (VP saw (S (NP the`	*I saw the child leave*
`(S (NP I ) (VP saw (S (NP (NP the`	*I saw the child’s dog leave*
`(S (NP I ) (VP saw (SBAR (NP the`	*I saw the child left*

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child’s dog*
`(S (NP I ) (VP saw (S (NP the`	*I saw the child leave*
`(S (NP I ) (VP saw (S (NP (NP the`	*I saw the child’s dog leave*
`(S (NP I ) (VP saw (SBAR (NP the`	*I saw the child left*
`(S (NP I ) (VP saw (SBAR (NP (NP the`	*I saw the child’s dog left*

An inferential challenge

`(S (NP I ) (VP saw (NP the`	*I saw the child*
`(S (NP I ) (VP saw (NP (NP the`	*I saw the child's dog*
`(S (NP I ) (VP saw (S (NP the`	*I saw the child leave*
`(S (NP I ) (VP saw (S (NP (NP the`	*I saw the child's dog leave*
`(S (NP I ) (VP saw (SBAR (NP the`	*I saw the child left*
`(S (NP I ) (VP saw (SBAR (NP (NP the`	*I saw the child's dog left*

There is a potentially unbounded number of tree-generation operations just to get to the next word!

Inference using beam search

```
(S (NP I ) (VP saw (NP the

(S (NP I ) (VP saw (NP (NP the

(S (NP I ) (VP saw (S (NP the

(S (NP I ) (VP saw (S (NP (NP the

(S (NP I ) (VP saw (SBAR (NP the

(S (NP I ) (VP saw (SBAR (NP (NP the
```

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Inference using beam search

Context *C*

```
(S (NP I ) (VP saw (NP the
(S (NP I ) (VP saw (NP (NP the
(S (NP I ) (VP saw (S (NP the
(S (NP I ) (VP saw (S (NP (NP the
(S (NP I ) (VP saw (SBAR (NP the
(S (NP I ) (VP saw (SBAR (NP (NP the
```

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Inference using beam search

$\overbrace{\texttt{(S (NP I) (VP saw}}^{\text{Context } \boldsymbol{C}}$ `(NP the`

`(S (NP I ) (VP saw` `(NP (NP the`

`(S (NP I ) (VP saw` `(S (NP the`

`(S (NP I ) (VP saw` `(S (NP (NP the`

`(S (NP I ) (VP saw` `(SBAR (NP the`

`(S (NP I ) (VP saw` `(SBAR (NP (NP the`

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Inference using beam search

Context ***C*** Actions ***A***

```
(S (NP I ) (VP saw (NP the
(S (NP I ) (VP saw (NP (NP the
(S (NP I ) (VP saw (S (NP the
(S (NP I ) (VP saw (S (NP (NP the
(S (NP I ) (VP saw (SBAR (NP the
(S (NP I ) (VP saw (SBAR (NP (NP the
```

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Inference using beam search

Context C	Actions A	$\log P(A \mid C)$
`(S (NP I ) (VP saw`	`(NP the`	-5.1
`(S (NP I ) (VP saw`	`(NP (NP the`	-6.3
`(S (NP I ) (VP saw`	`(S (NP the`	-5.8
`(S (NP I ) (VP saw`	`(S (NP (NP the`	-7.2
`(S (NP I ) (VP saw`	`(SBAR (NP the`	-6.2
`(S (NP I ) (VP saw`	`(SBAR (NP (NP the`	-7.8

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Inference using beam search

Context *C*	Actions *A*	$\log P(A \mid C)$	Rank on beam
`(S (NP I ) (VP saw`	`(NP the`	-5.1	1
`(S (NP I ) (VP saw`	`(NP (NP the`	-6.3	4
`(S (NP I ) (VP saw`	`(S (NP the`	-5.8	2
`(S (NP I ) (VP saw`	`(S (NP (NP the`	-7.2	✘
`(S (NP I ) (VP saw`	`(SBAR (NP the`	-6.2	3
`(S (NP I ) (VP saw`	`(SBAR (NP (NP the`	-7.8	✘

A "word-synchronous" beam, beam size=4

Natural account of **strong** garden-pathing effects (*the woman brought the sandwich tripped*):

The needed analysis "falls off the beam"

(Stern et al., 2017)

Challenges for surprisal theory

- Limitations in the **memory representations** available during real-time comprehension

- Accounting for **input uncertainty** from noise & speaker error

Structural Forgetting and the Noisy Channel

(Futrell & Levy, 2017)

Structural Forgetting and the Noisy Channel

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

(Futrell & Levy, 2017)

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

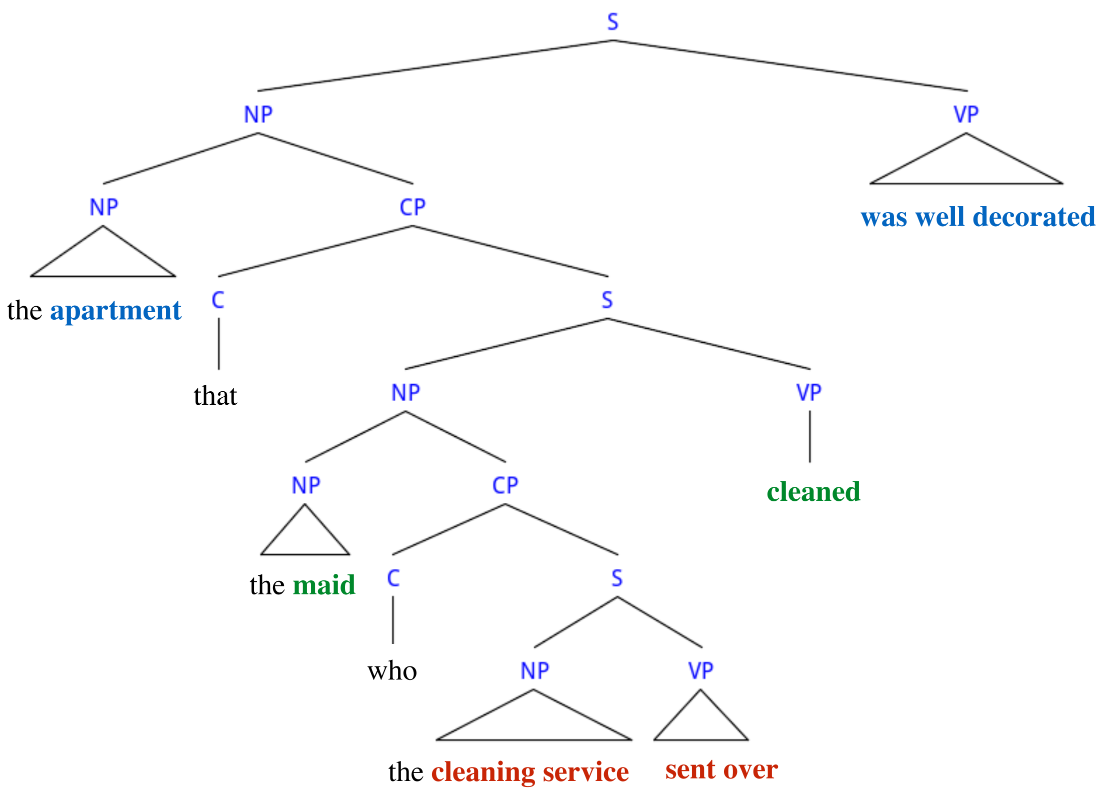
1. *The **apartment** that the **maid** who the **cleaning service sent over was well-decorated**. 👍

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**. 👍

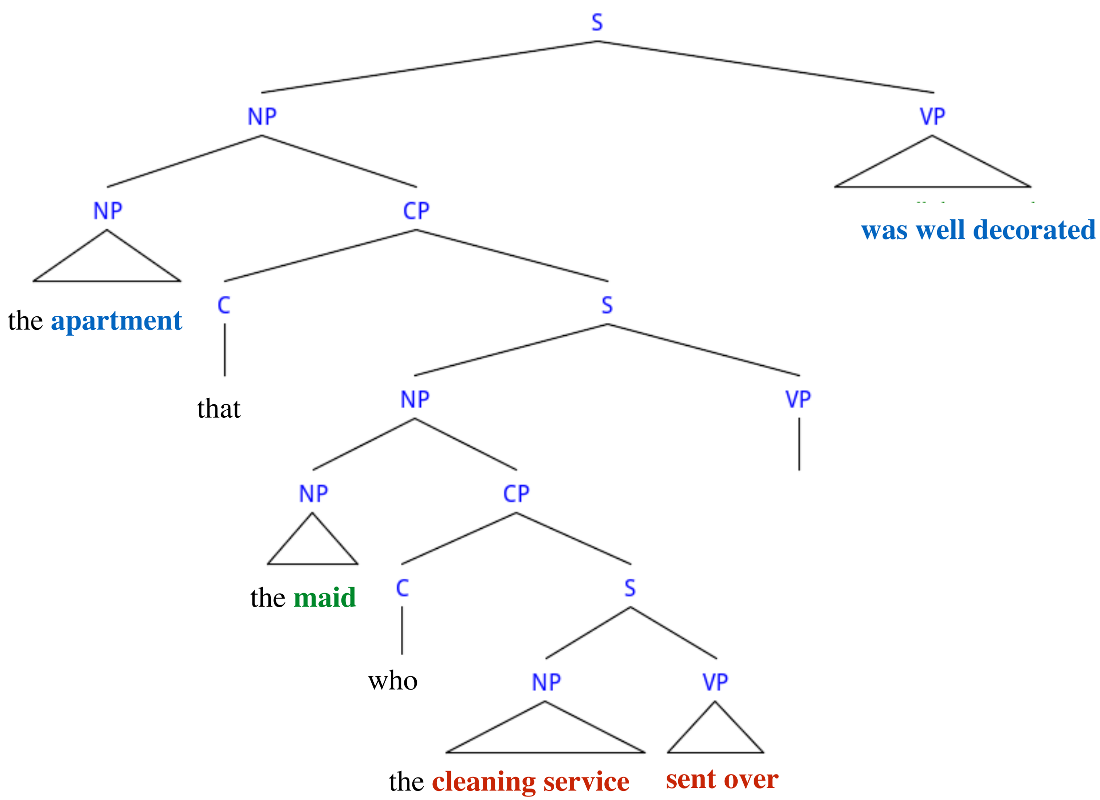
2. The **apartment** that the **maid** who the **cleaning service** **sent over** **cleaned** **was well-decorated**. 👎



Structural Forgetting

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

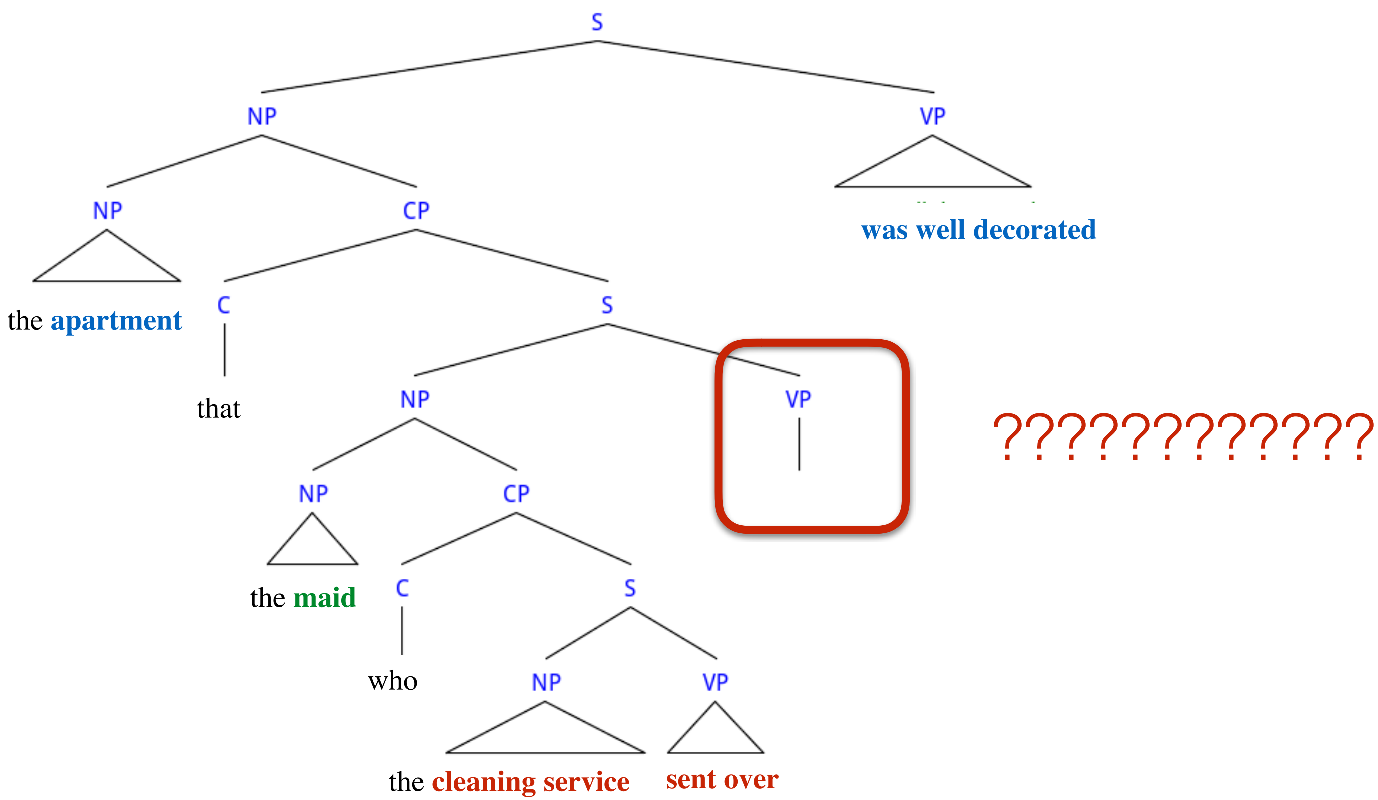
2. The **apartment** that the **maid** who the **cleaning service** **sent over** **cleaned** **was well-decorated**. 👎



Structural Forgetting

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

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



Structural Forgetting

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

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

Structural Forgetting

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

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

- **Structural forgetting effect**: part of the sentence is forgotten by the time you get to the end (Gibson & Thomas, 1999; Frazier, 1985; Fodor, p.c.)

Structural Forgetting

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

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

- **Structural forgetting effect**: part of the sentence is forgotten by the time you get to the end (Gibson & Thomas, 1999; Frazier, 1985; Fodor, p.c.)
- The ungrammatical sentence seems better than the grammatical one.
    - A "**grammaticality illusion**": how could we define grammaticality in this case?

Structural Forgetting

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

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

Structural Forgetting

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

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

- But the effect is **language-dependent** (Vasishth et al., 2010; Frank et al., 2016).

Structural Forgetting

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

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

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

Structural Forgetting

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

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

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

Structural Forgetting

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

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

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

Structural Forgetting

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

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

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

Structural Forgetting

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

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

Structural Forgetting

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

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

- These contexts are more common in German than English (Roland et al., 2007).

Structural Forgetting

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

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

- These contexts are more common in German than English (Roland et al., 2007).
  - English: the maid [that cleaned the apartment]
    the apartment [that the maid cleaned]

Structural Forgetting

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

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

- These contexts are more common in German than English (Roland et al., 2007).
  - English: the maid [that cleaned the apartment] the apartment [that the maid cleaned] **80%**

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%**

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>]

An incremental inference puzzle for surprisal

An incremental inference puzzle for surprisal

- Try to understand this sentence:

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

…and contrast this with:

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

…and contrast this with:

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

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

…and contrast this with:

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

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

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

…and contrast this with:

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

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

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

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

…and contrast this with:

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

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

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

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

Tabor et al. (2004, JML)

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)

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

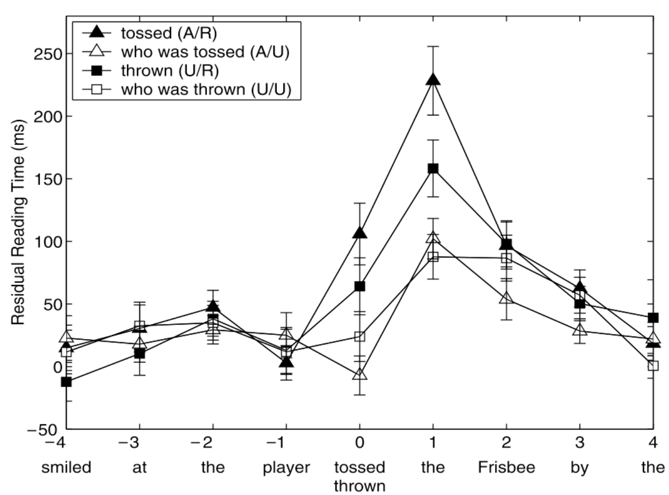
…and contrast this with:

  (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)

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

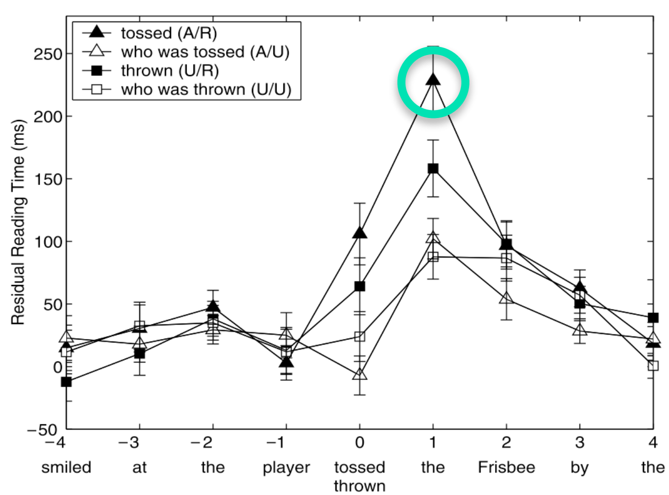
…and contrast this with:

  (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)

An incremental inference puzzle for surprisal

- Try to understand this sentence:

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

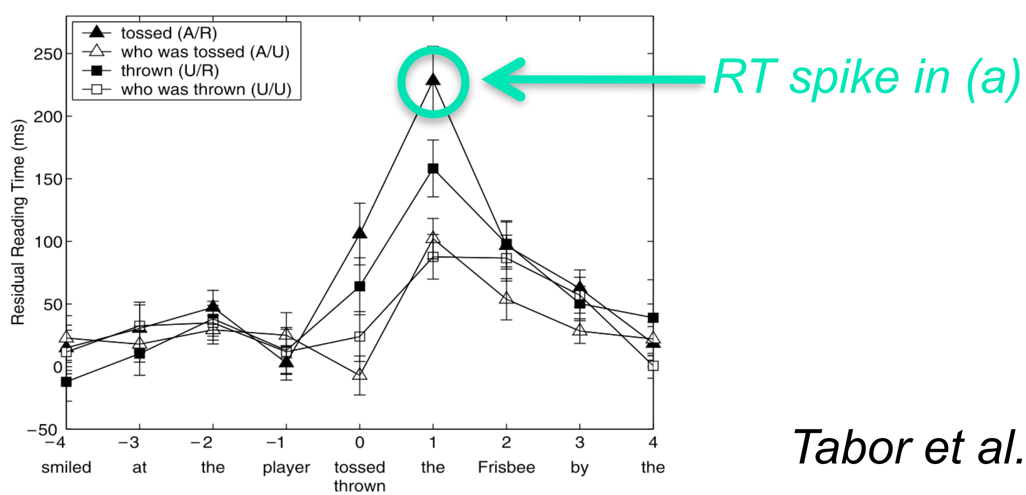
…and contrast this with:

  (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?

Why is *tossed/thrown* interesting?

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

Why is *tossed/thrown* interesting?

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

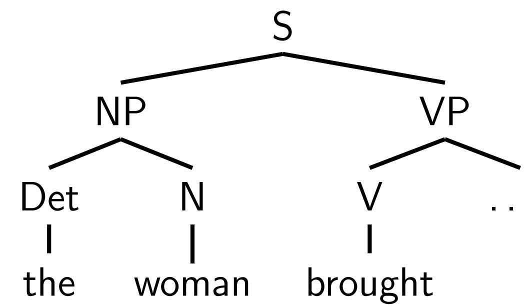
    verb?

    participle?

Why is *tossed/thrown* interesting?

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

    verb?
    participle?

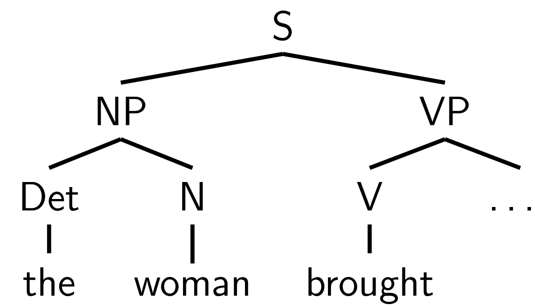


Why is *tossed/thrown* interesting?

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

    verb?

    participle?



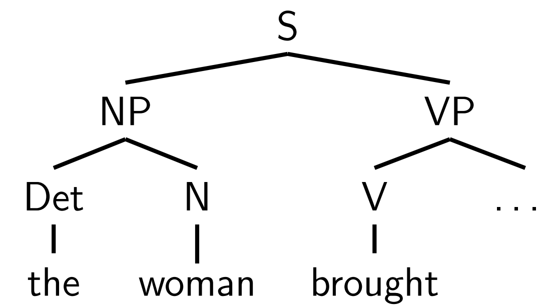
- But now context "should" rule out the garden path:

Why is *tossed/thrown* interesting?

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

    verb?

    participle?

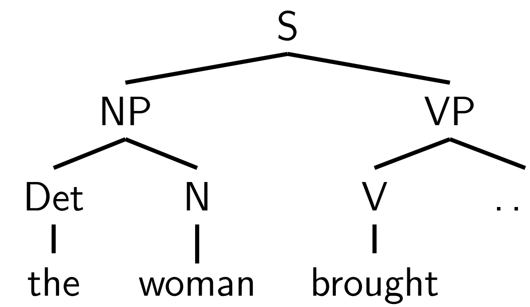


- But now context "should" rule out the garden path:
  - *The coach smiled at the player tossed…*

Why is *tossed/thrown* interesting?

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

    verb?
    participle?



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

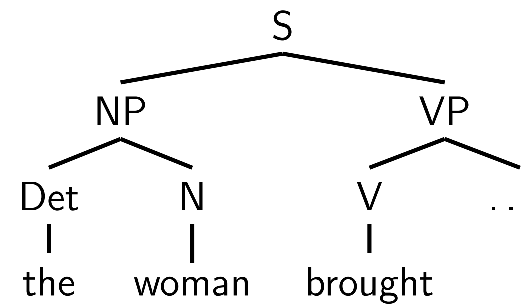
    verb?
    participle?

Why is *tossed/thrown* interesting?

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

    verb?

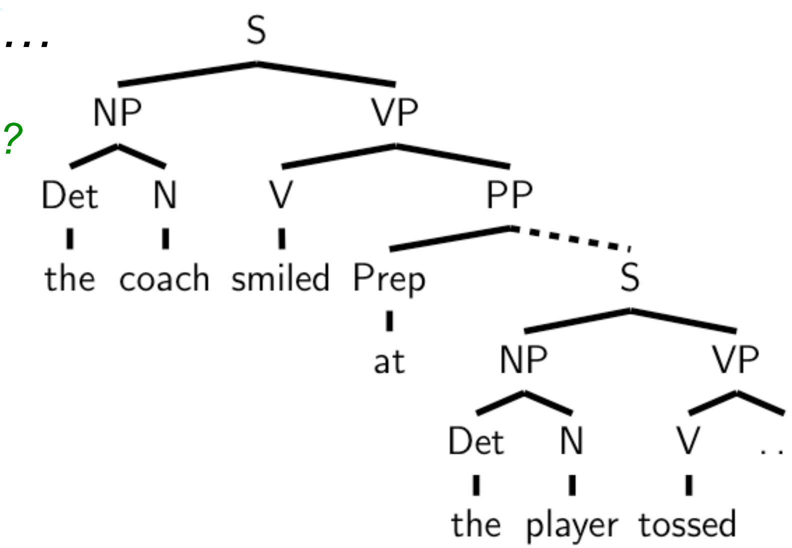
    participle?



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

    verb?

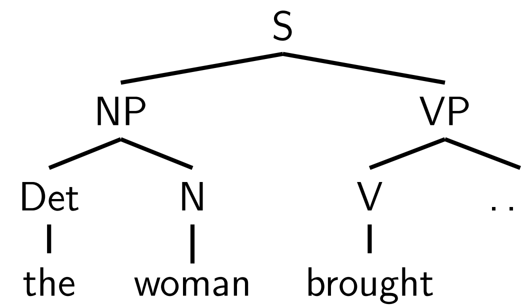
    participle?



Why is *tossed/thrown* interesting?

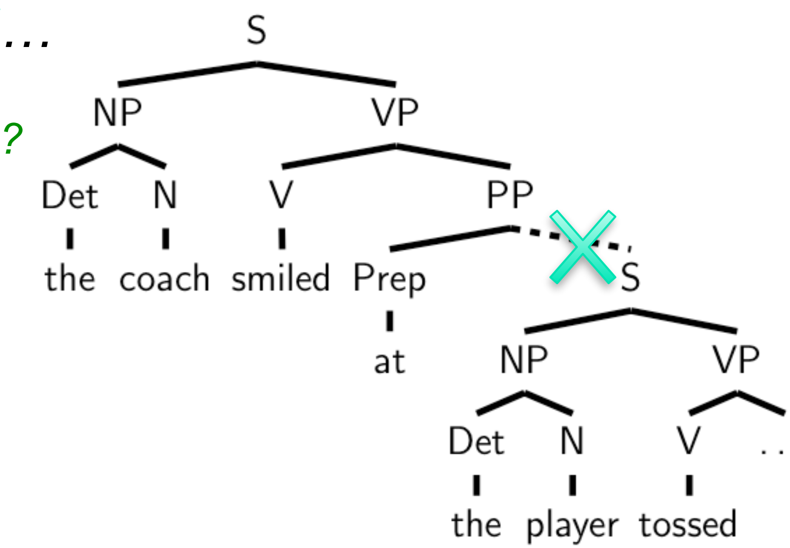
- As with classic garden-paths, part-of-speech ambiguity leads to misinterpretation
  - *The woman brought the sandwich…tripped*

    verb?
    participle?



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

    verb?
    participle?

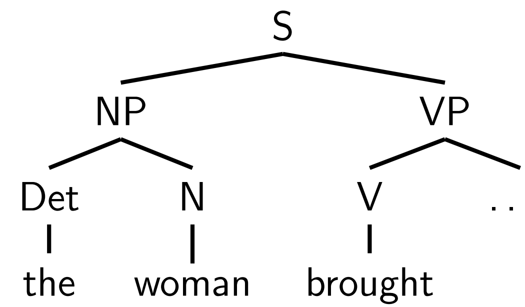


Why is *tossed/thrown* interesting?

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

    verb?

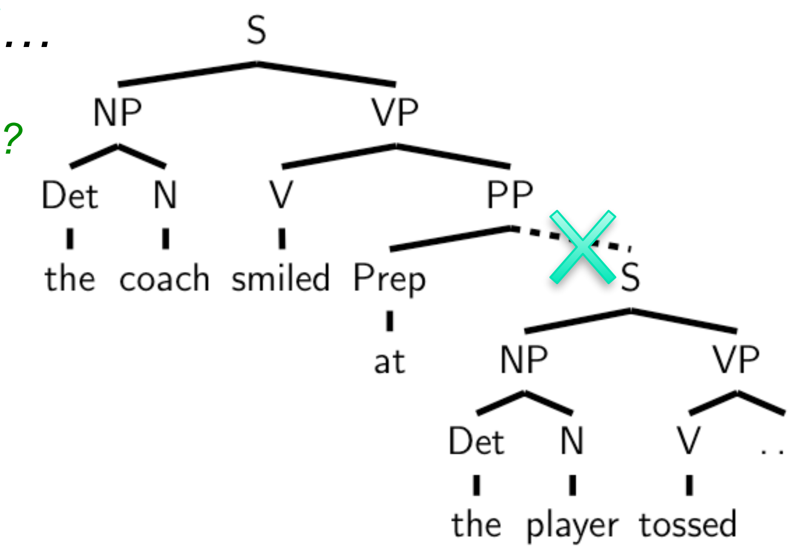
    participle?



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

    verb?

    participle?



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

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)

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

Failures!

(Anderson, 1990, 1991)

Rational analysis

Revise somehow

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

Failures!

(Anderson, 1990, 1991)

Rational analysis

Revise somehow

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

Failures!

Our case study: revise #2, the model of the environment to which the cognitive agent is adapted

(Anderson, 1990, 1991)

Uncertain input in language comprehension

Uncertain input in language comprehension

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

Uncertain input in language comprehension

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

Uncertain input in language comprehension

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

Uncertain input in language comprehension

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

1. What might a model of sentence comprehension under uncertain input look like?
2. What interesting consequences might such a model have?

Noisy-channel language comprehension

$$P(\text{T} \mid \text{words}) \propto P(\text{words} \mid T)P(T)$$

Levy (2008, EMNLP); Futrell & Levy (2017, EACL)

Noisy-channel language comprehension

- Standard probabilistic language comprehension

$$P(\text{T} \mid \text{words}) \propto P(\text{words} \mid T)P(T)$$

Levy (2008, EMNLP); Futrell & Levy (2017, EACL)

Noisy-channel language comprehension

- Standard probabilistic language comprehension

$$P(\text{T} \mid \text{words}) \propto P(\text{words} \mid T)P(T)$$

- **Revision**: probabilistic language comprehension where the input is subject to *noise* and *imperfect memory*

Levy (2008, EMNLP); Futrell & Levy (2017, EACL)

Noisy-channel language comprehension

- Standard probabilistic language comprehension

$$P(\text{T} \mid \text{words}) \propto P(\text{words} \mid T)P(T)$$

- **Revision**: probabilistic language comprehension where the input is subject to *noise* and *imperfect memory*

$$P(\text{T} \mid \text{input}) \propto P(\text{input} \mid T)P(T)$$

Levy (2008, EMNLP); Futrell & Levy (2017, EACL)

Noisy-channel language comprehension

- Standard probabilistic language comprehension

$$P(\text{T} \mid \text{words}) \propto P(\text{words} \mid T)P(T)$$

- **Revision**: probabilistic language comprehension where the input is subject to *noise* and *imperfect memory*

$$P(\text{T} \mid \text{input}) \propto P(\text{input} \mid T)P(T)$$

$$= \sum_{w} P(\text{input} \mid w, T)P(w, T)$$

Ranges over possible word sequences

Levy (2008, EMNLP); Futrell & Levy (2017, EACL)

Incremental inference under uncertain input

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

Incremental inference under uncertain input

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

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

Incremental inference under uncertain input

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

(and?)

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

Incremental inference under uncertain input

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

(and?)
(as?)

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

Incremental inference under uncertain input

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

(and?)

(and?)
(as?)

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

Incremental inference under uncertain input

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

(and?) (that?)

(and?) (as?)

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

Incremental inference under uncertain input

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

(and?) (as?) — (and?) (that?) (who?)

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

Incremental inference under uncertain input

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

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

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

Incremental inference under uncertain input

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

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

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

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*

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*

The core of the intuition

the coach smiled…

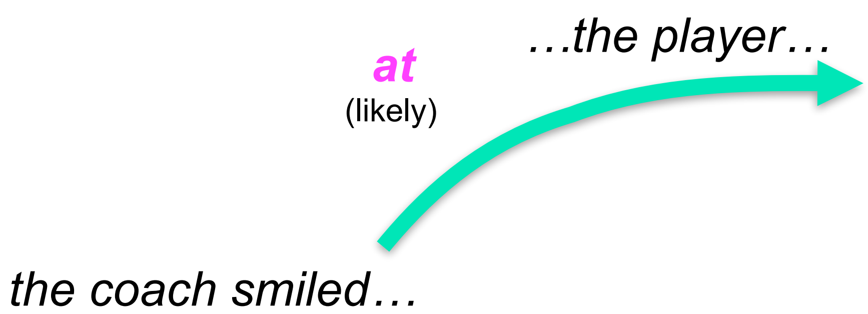
The core of the intuition

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

the coach smiled…

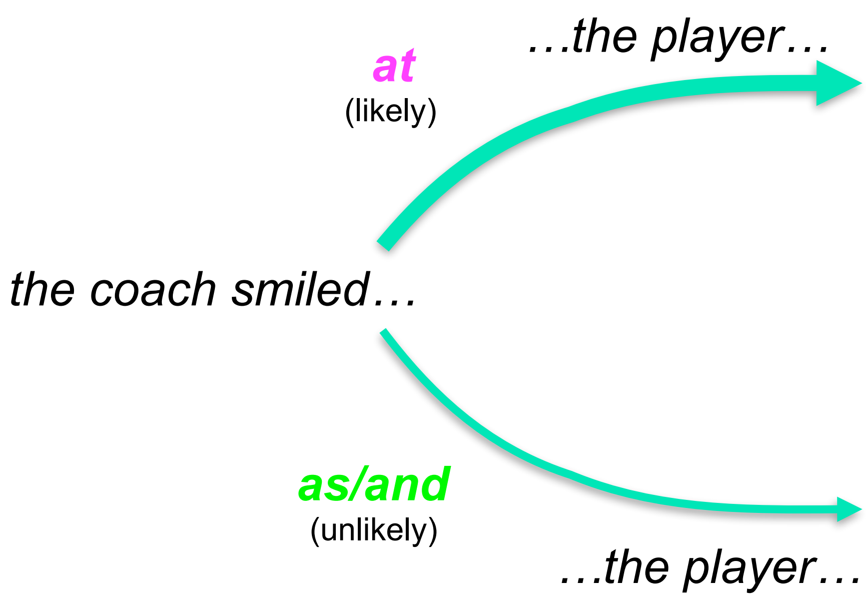
The core of the intuition

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



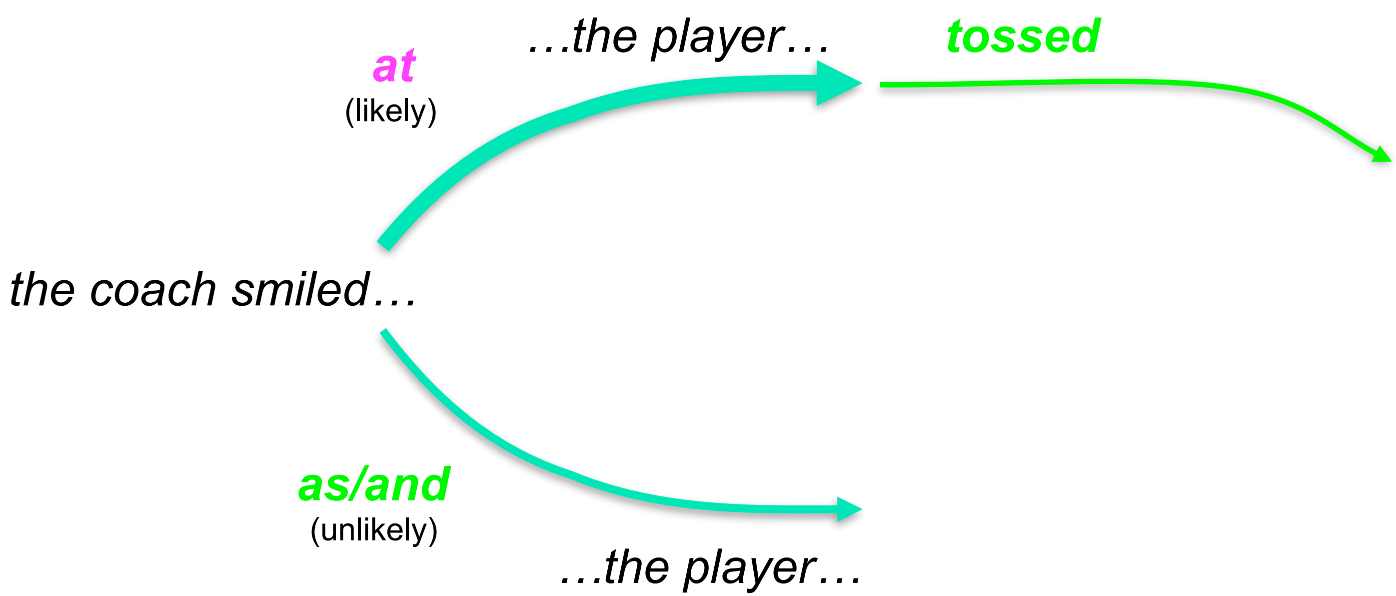
The core of the intuition

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



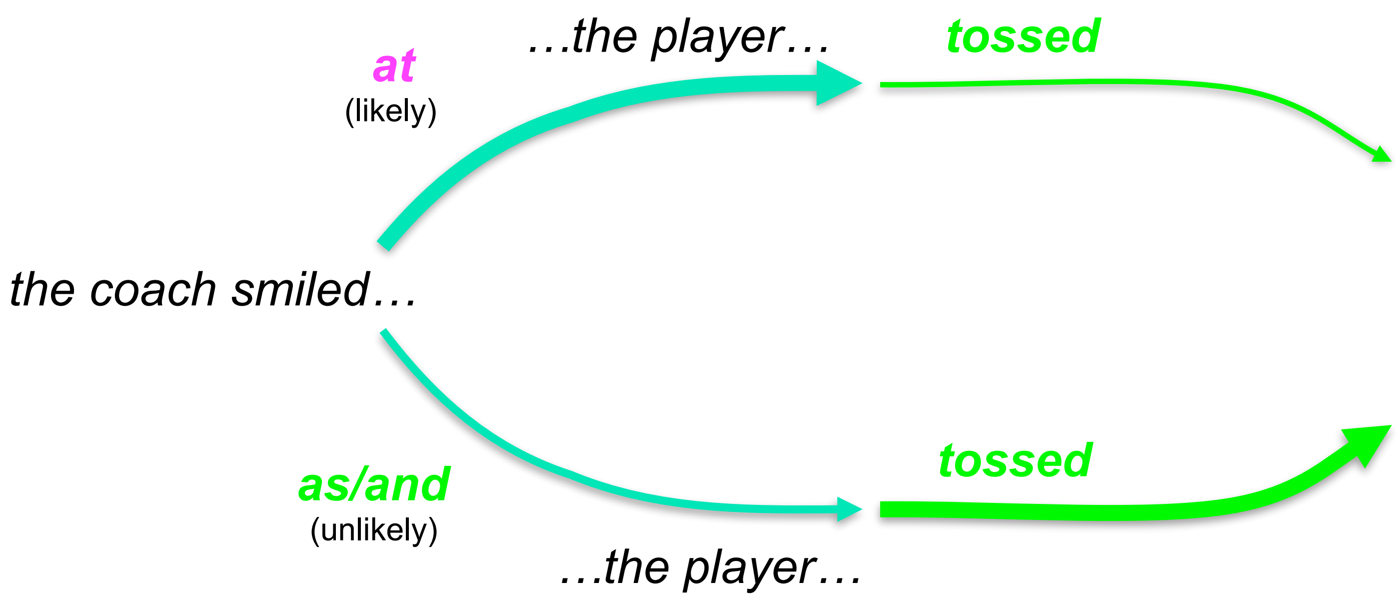
The core of the intuition

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



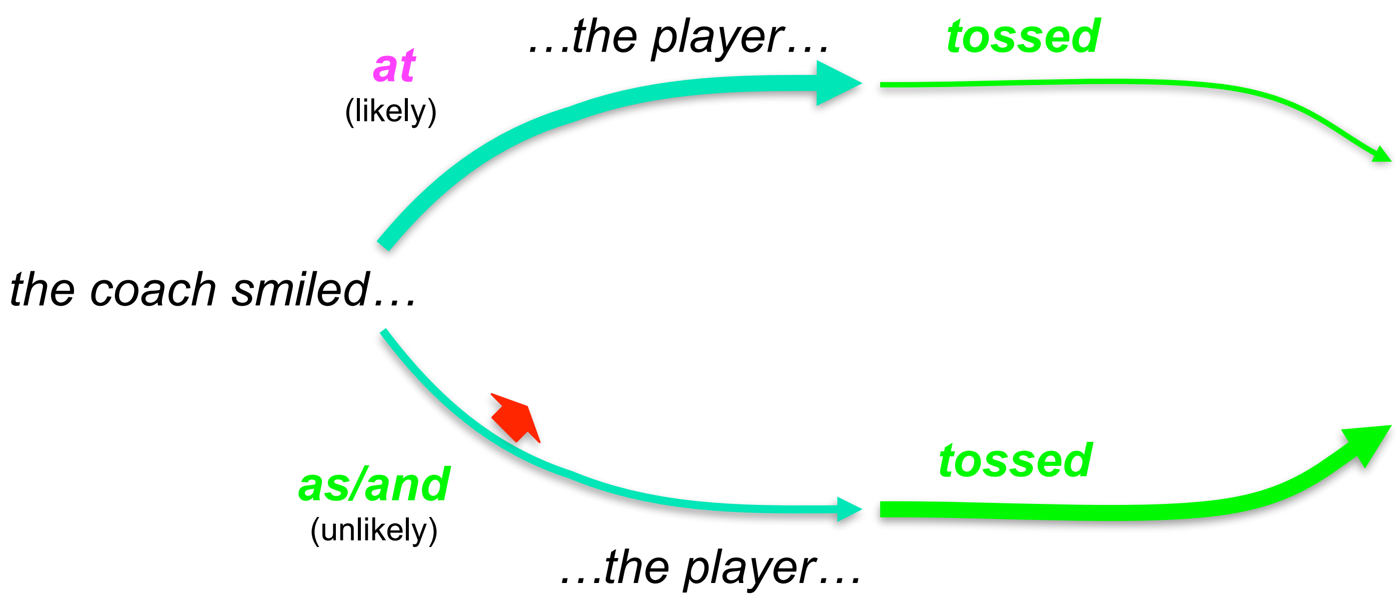
The core of the intuition

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



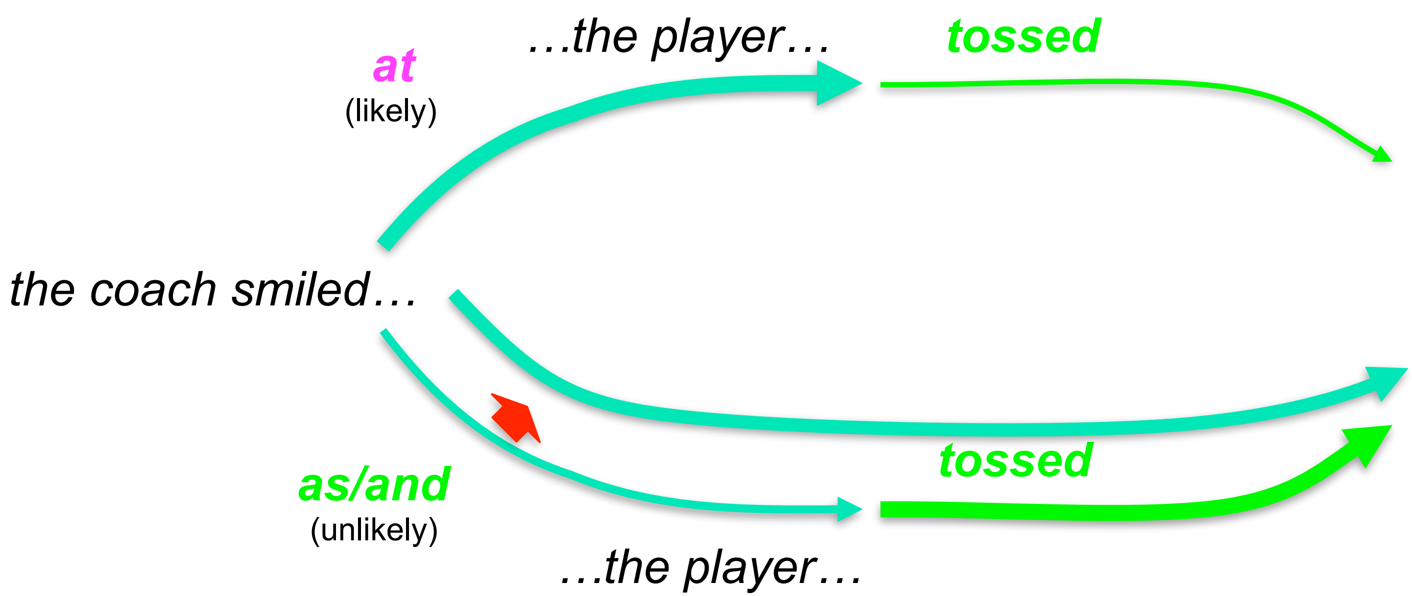
The core of the intuition

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



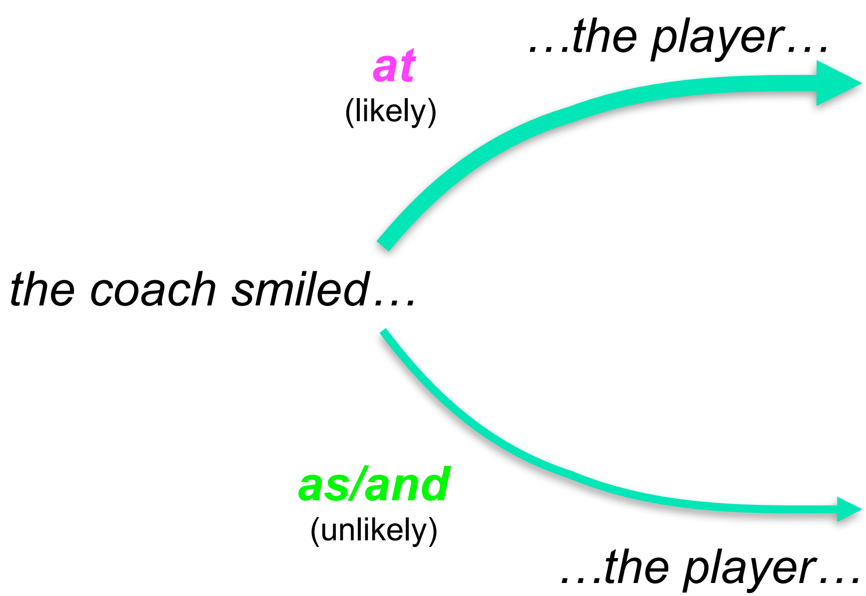
The core of the intuition

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



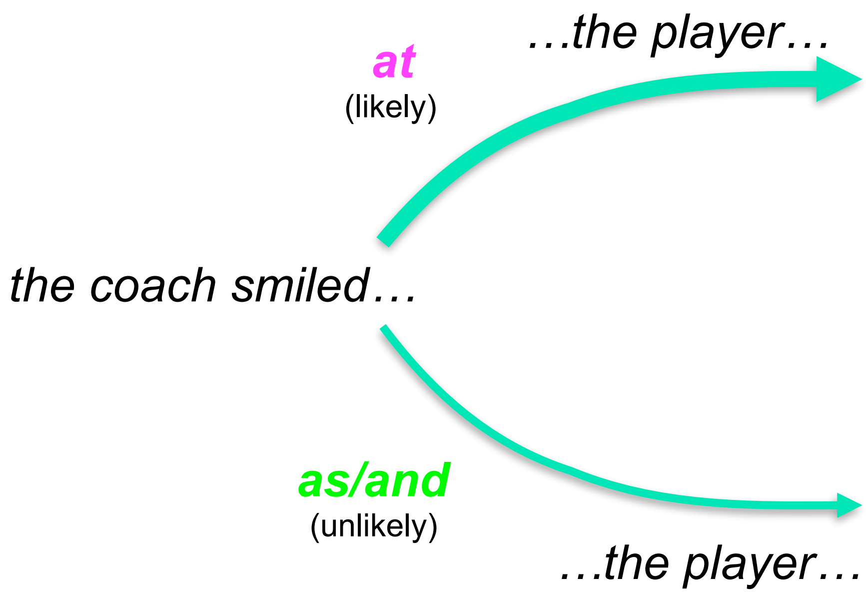
The core of the intuition

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



The core of the intuition

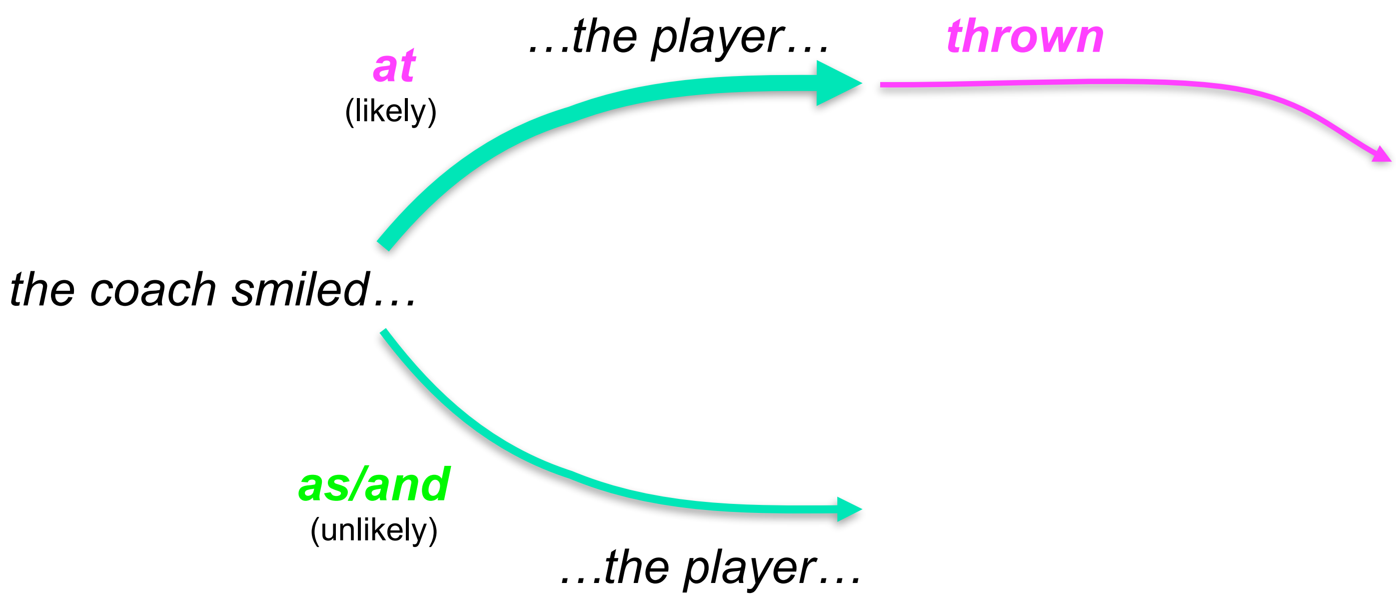
- Grammar & input come together to determine two possible "paths" through the partial sentence: *(line thickness ≈ probability)*



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

The core of the intuition

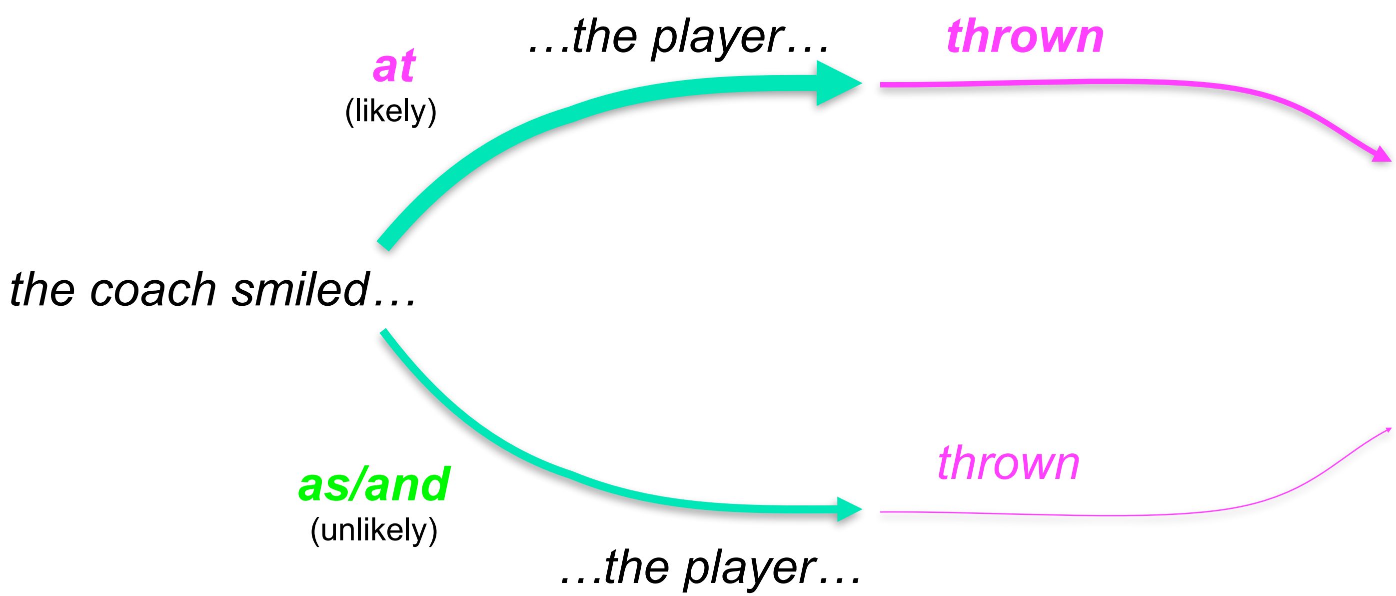
- Grammar & input come together to determine two possible "paths" through the partial sentence: *(line thickness ≈ probability)*



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

The core of the intuition

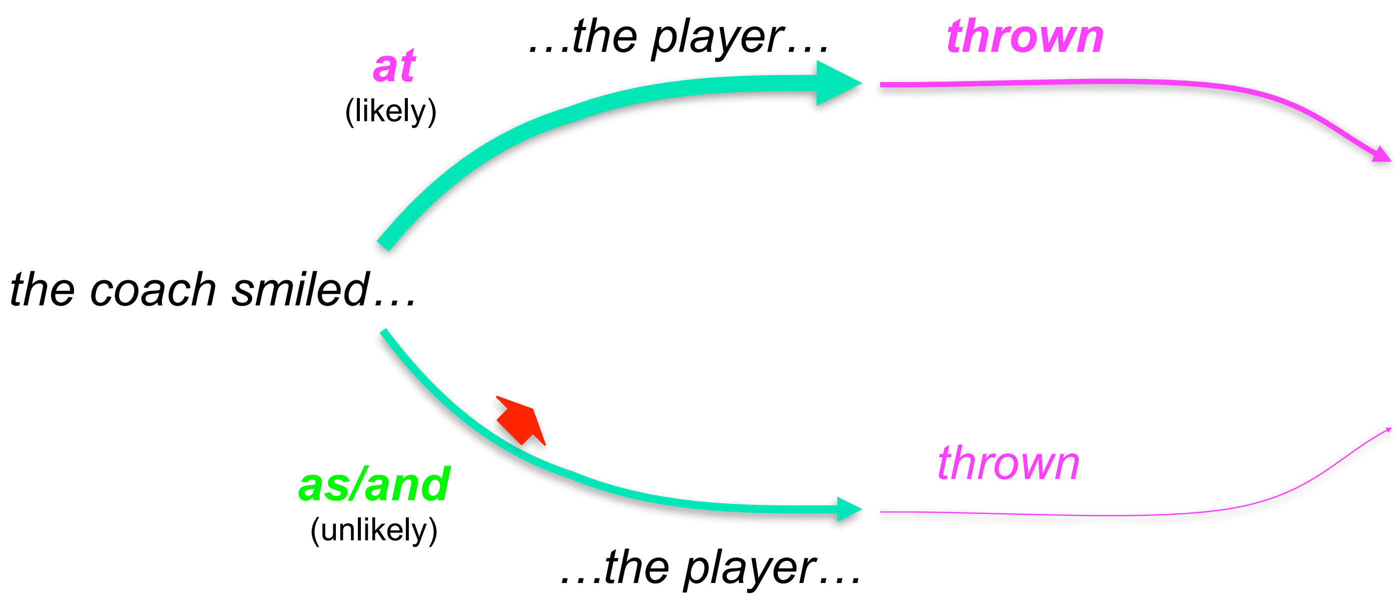
- Grammar & input come together to determine two possible "paths" through the partial sentence: *(line thickness ≈ probability)*



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

The core of the intuition

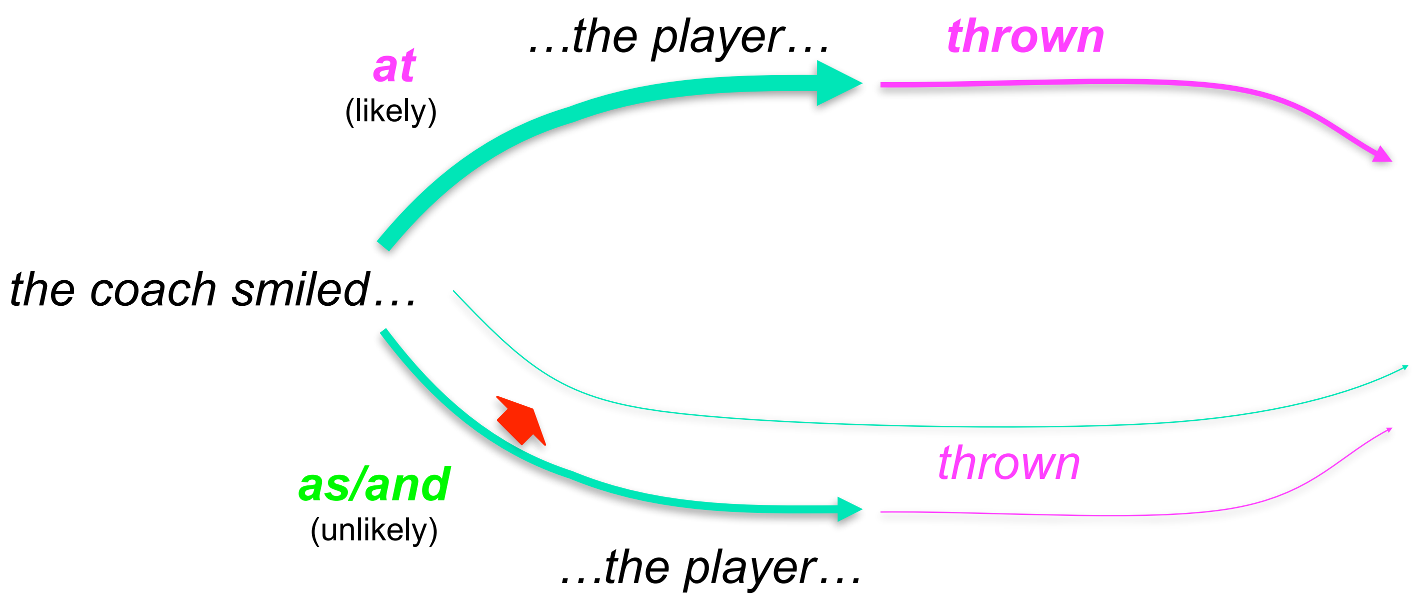
- Grammar & input come together to determine two possible "paths" through the partial sentence: *(line thickness ≈ probability)*



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

The core of the intuition

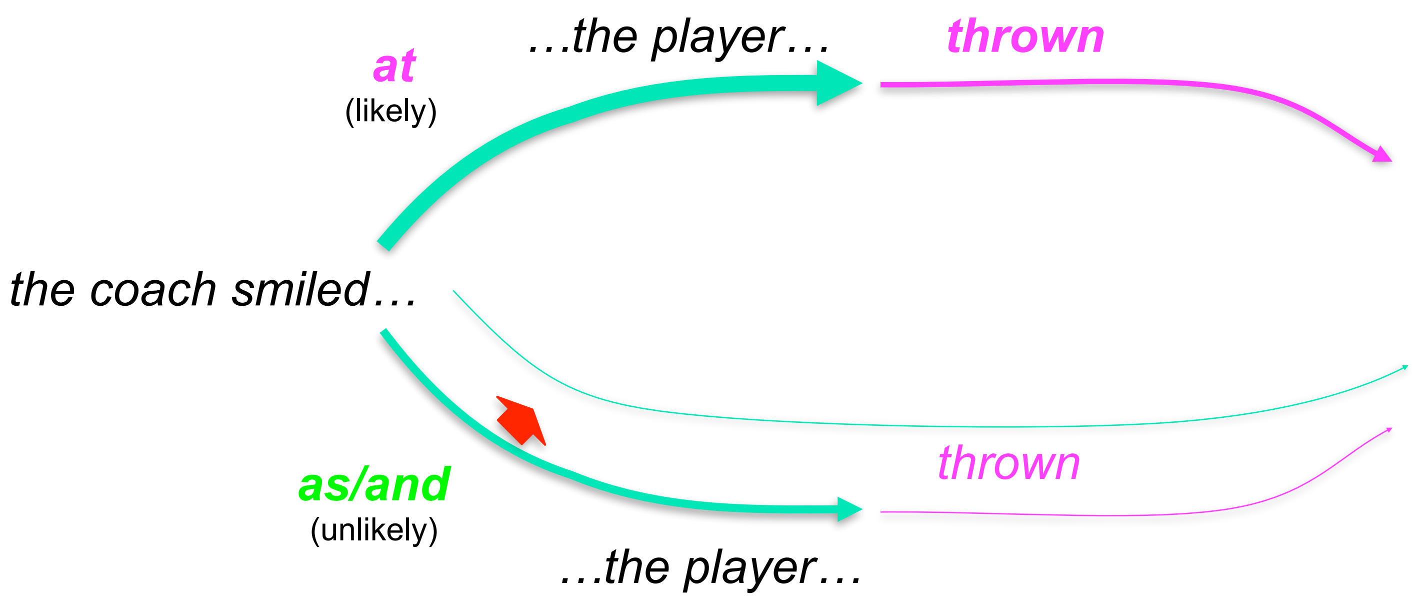
- Grammar & input come together to determine two possible "paths" through the partial sentence: *(line thickness ≈ probability)*



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

The core of the intuition

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



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

Experimental design

Experimental design

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

Experimental design

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

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

Experimental design

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

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

Experimental design

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

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

- Prediction: **interaction** between preposition & part-of-speech ambiguity in eye movements upon encountering participle

Experimental design

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

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

- Prediction: **interaction** between preposition & part-of-speech ambiguity in eye movements upon encountering participle

Experimental results

The coach smiled <u>*at*</u> *the player* <u>*tossed*</u>*…*

Experimental results

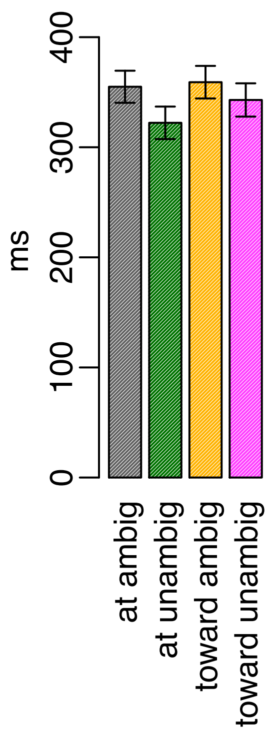
The coach smiled at the player tossed…

Experimental results

The coach smiled at the player tossed…

Experimental results

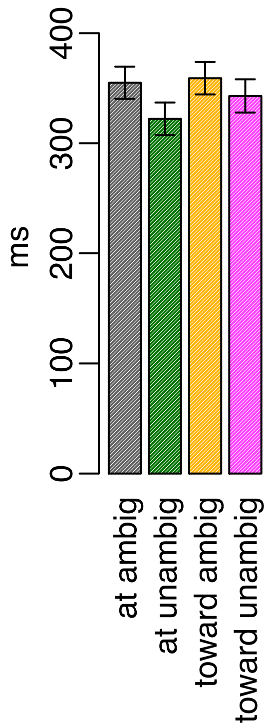
The coach smiled at the player tossed…



First-pass RT

Experimental results

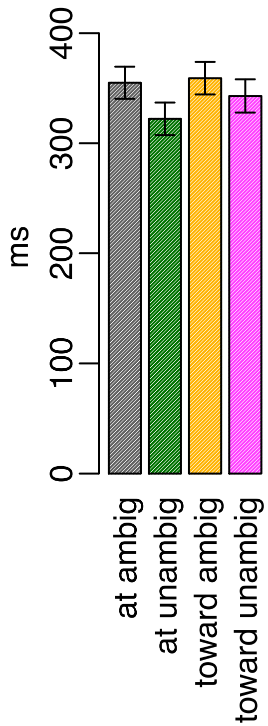
The coach smiled <u>at</u> the player <u>tossed</u>…



First-pass RT

Experimental results

The coach smiled at the player tossed…

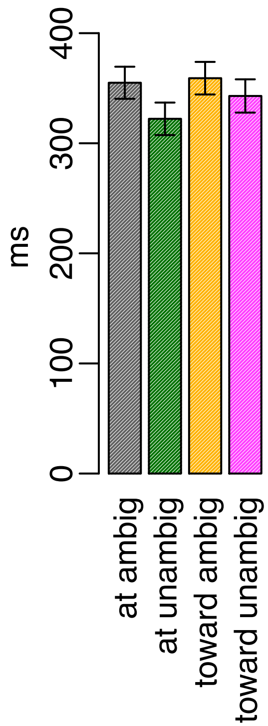


First-pass RT

Experimental results

?

The coach smiled <u>at</u> the player <u>tossed</u>…

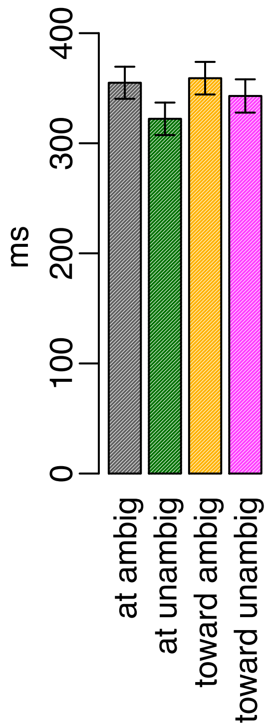


First-pass RT

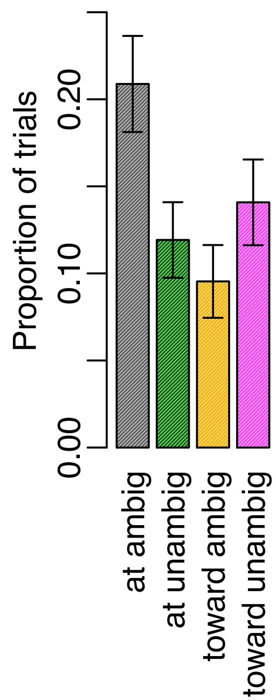
Experimental results

?

The coach smiled <u>at</u> the player <u>tossed</u>…



First-pass RT

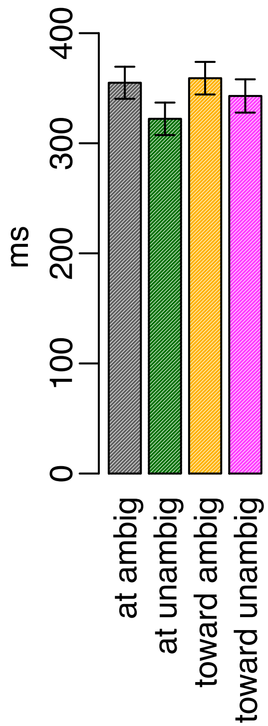


Regressions out

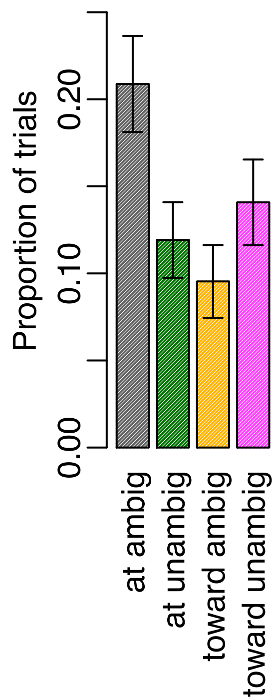
Experimental results

?

The coach smiled at the player tossed…



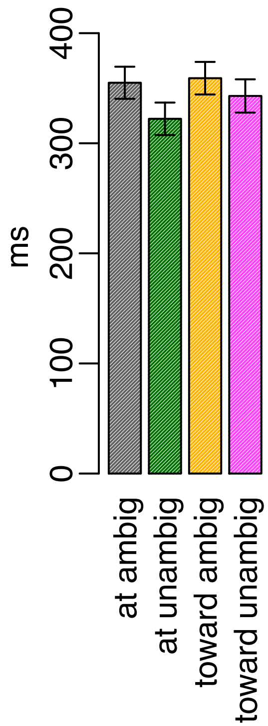
First-pass RT



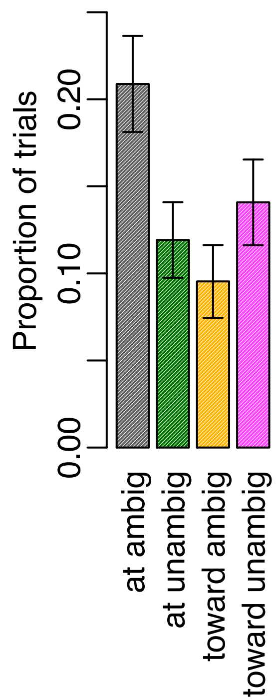
Regressions out

Experimental results

The coach smiled <u>at</u> the player <u>tossed</u>…



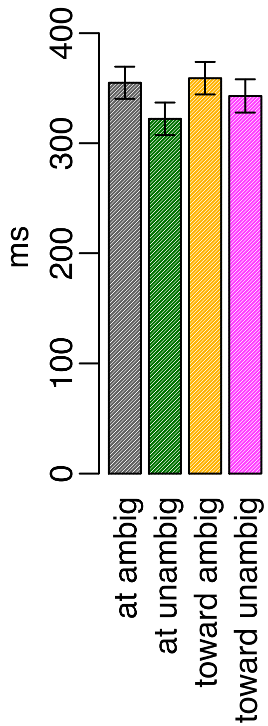
First-pass RT



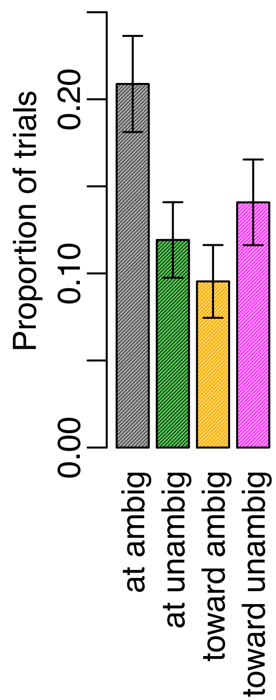
Regressions out

Experimental results

The coach smiled at the player tossed…



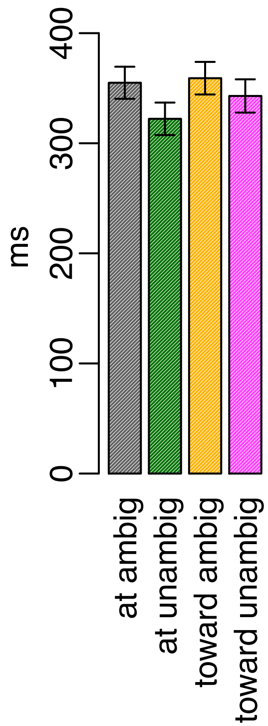
First-pass RT



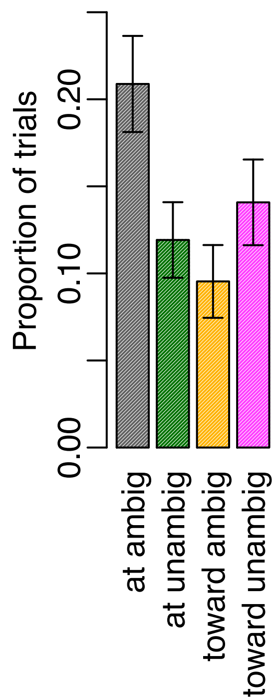
Regressions out

Experimental results

The coach smiled at the player tossed…



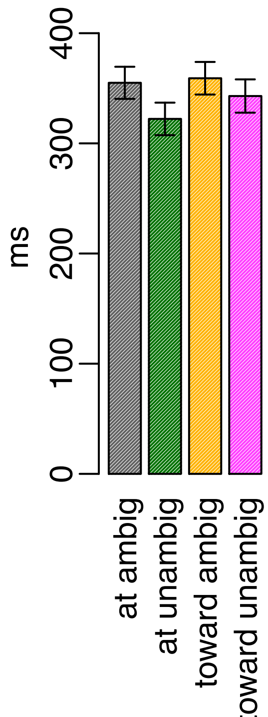
First-pass RT



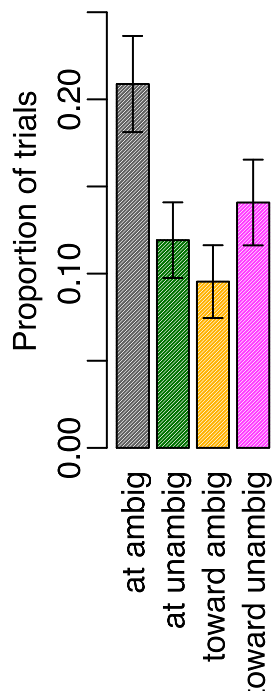
Regressions out

Experimental results

The coach smiled at the player tossed…



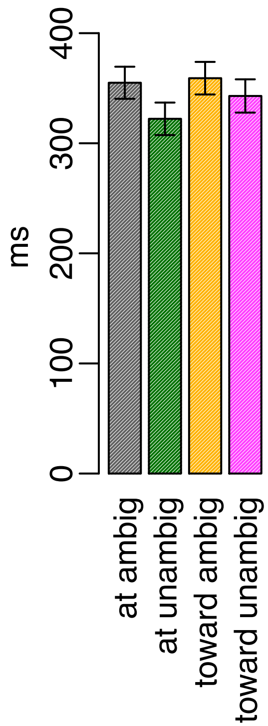
First-pass RT



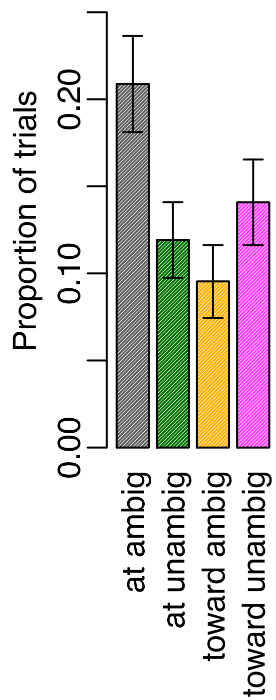
Regressions out

Experimental results

The coach smiled at the player tossed…



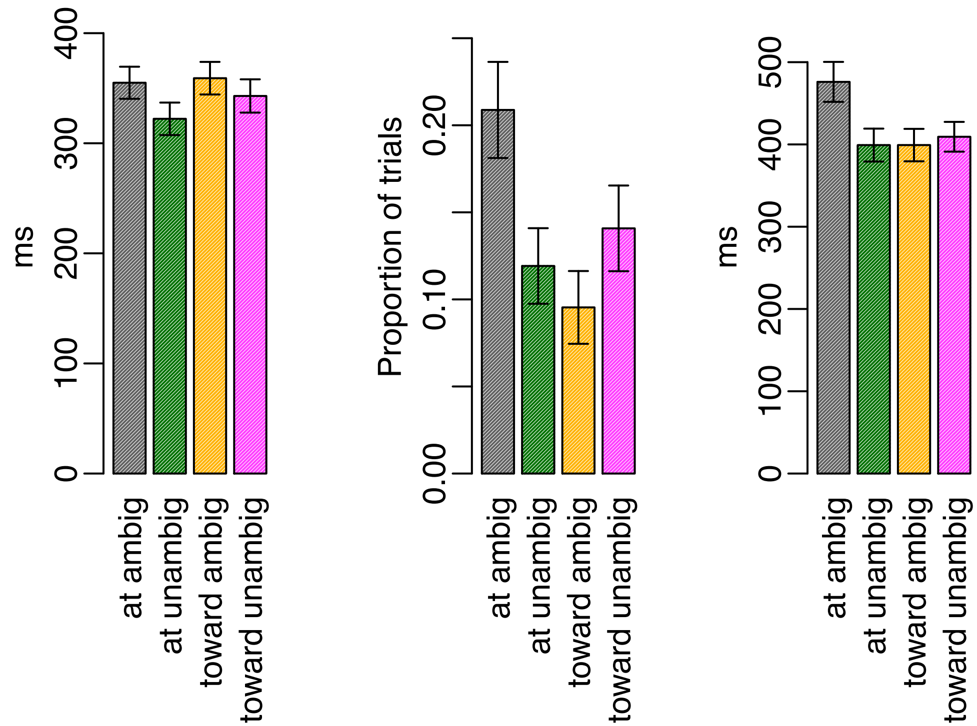
First-pass RT



Regressions out

Experimental results

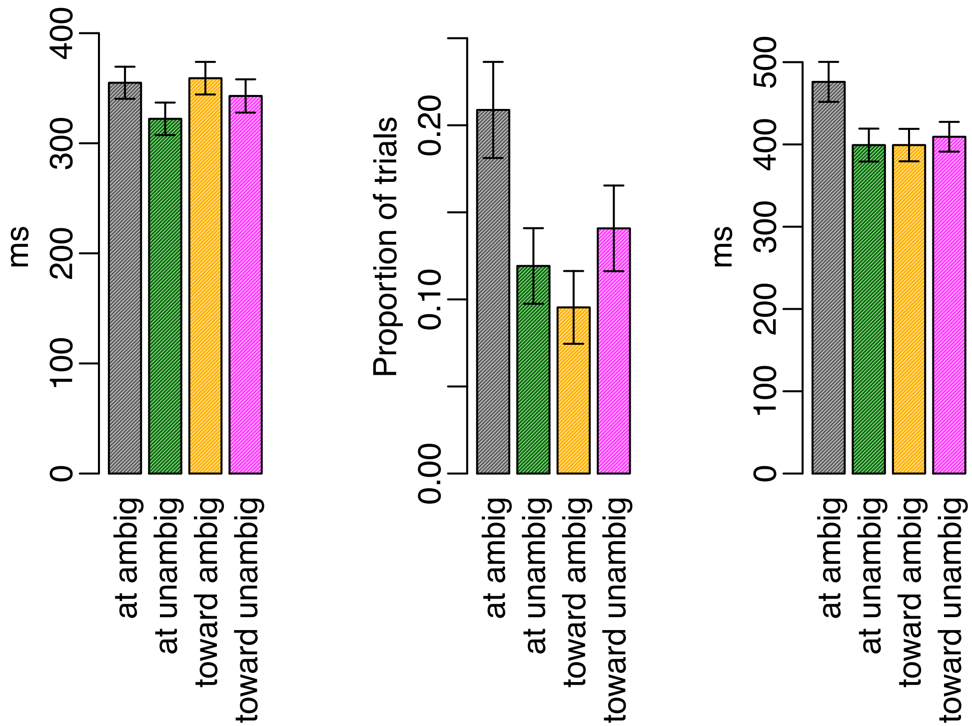
The coach smiled at the player tossed…



First-pass RT *Regressions out* *Go-past RT*


Experimental results

The coach smiled at the player tossed…



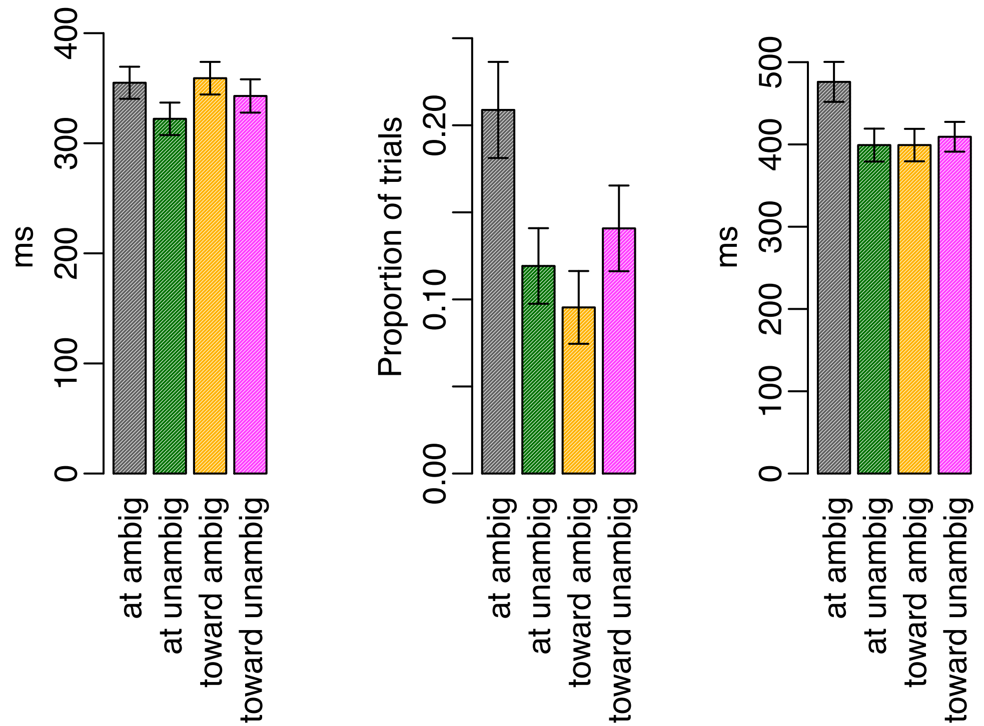
First-pass RT

Regressions out

Go-past RT

Experimental results

The coach smiled at the player tossed …

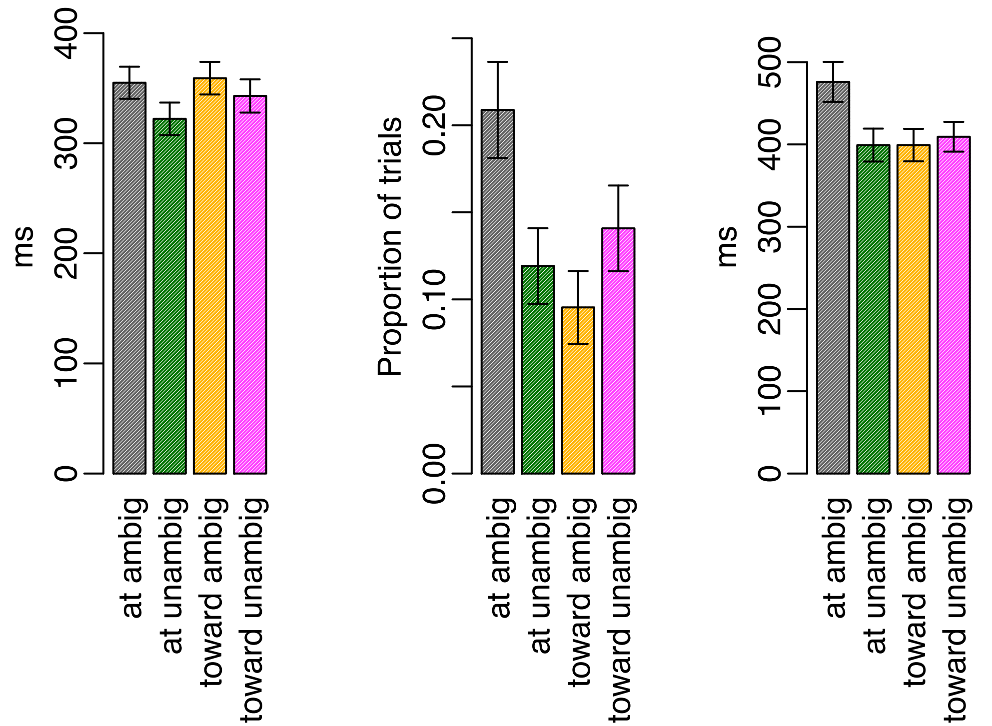


First-pass RT *Regressions out* *Go-past RT*

Experimental results

?

The coach smiled at the player tossed…



First-pass RT

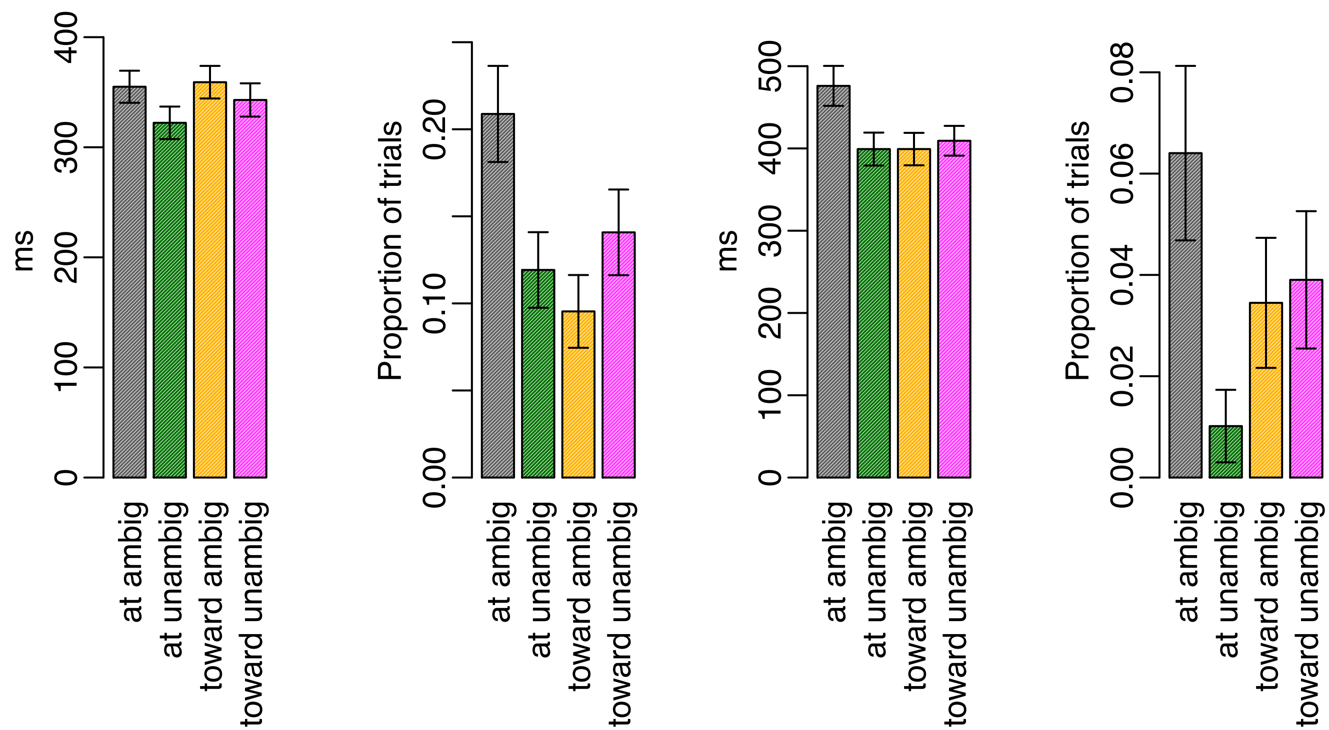
Regressions out

Go-past RT

Experimental results

?

The coach smiled at the player tossed …

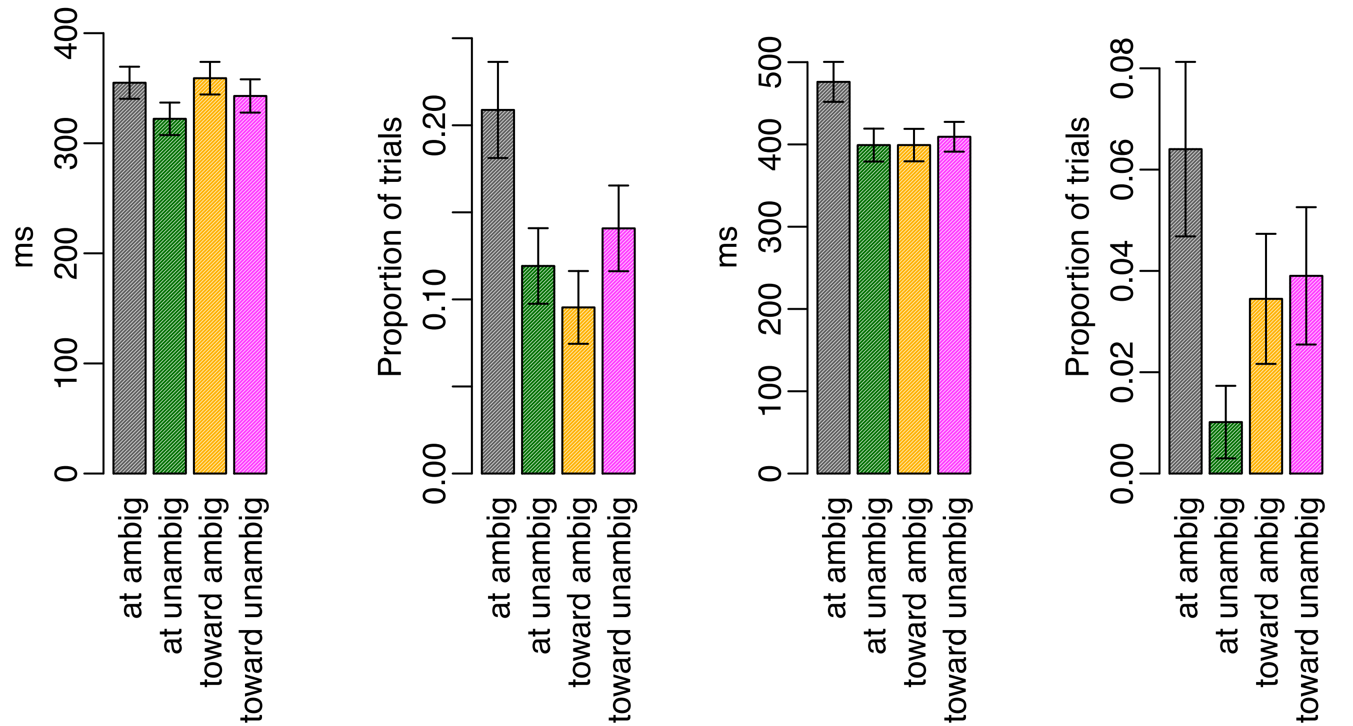


First-pass RT | *Regressions out* | *Go-past RT* | *Go-past regressions*

Experimental results

?

The coach smiled at the player tossed …



First-pass RT

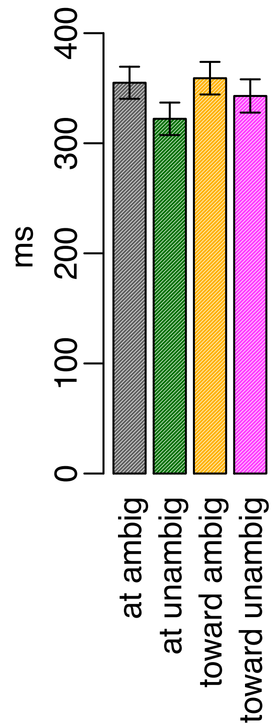
Regressions out

Go-past RT

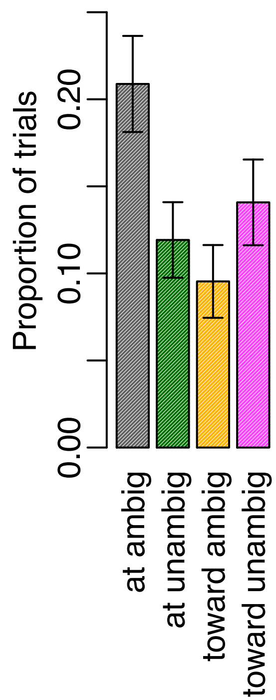
Go-past regressions

Experimental results

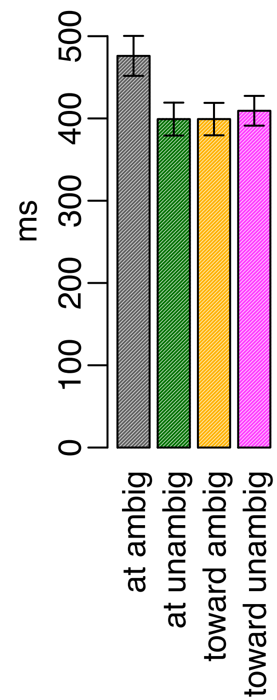
The coach smiled at the player tossed…



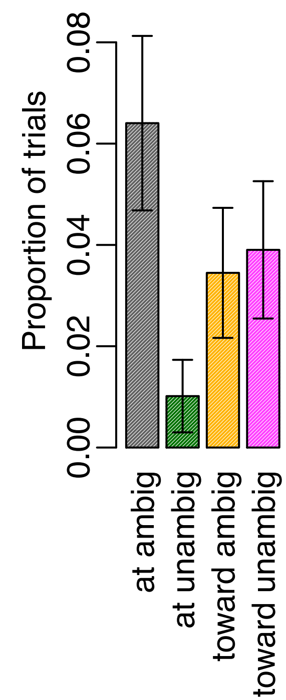
First-pass RT



Regressions out



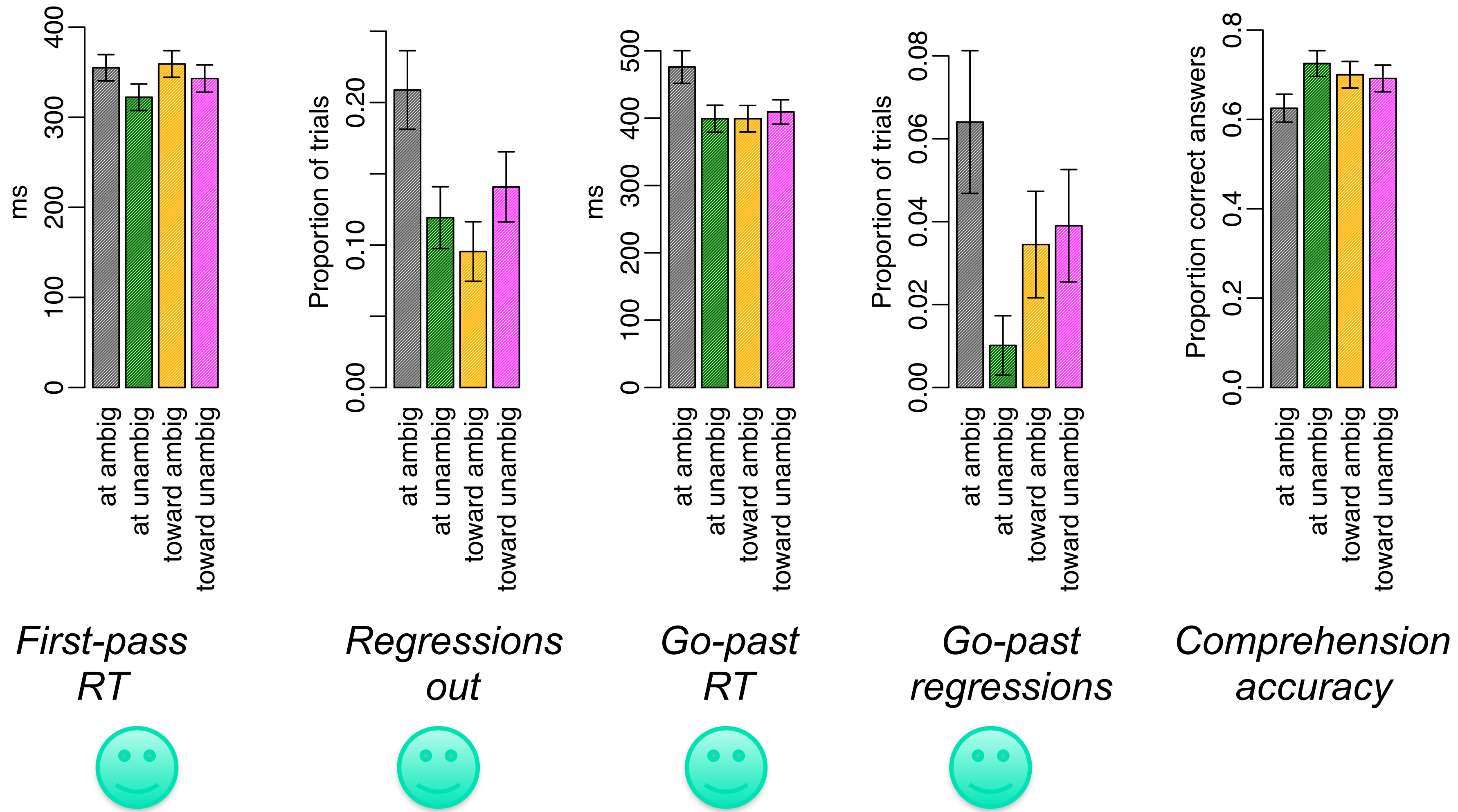
Go-past RT



Go-past regressions

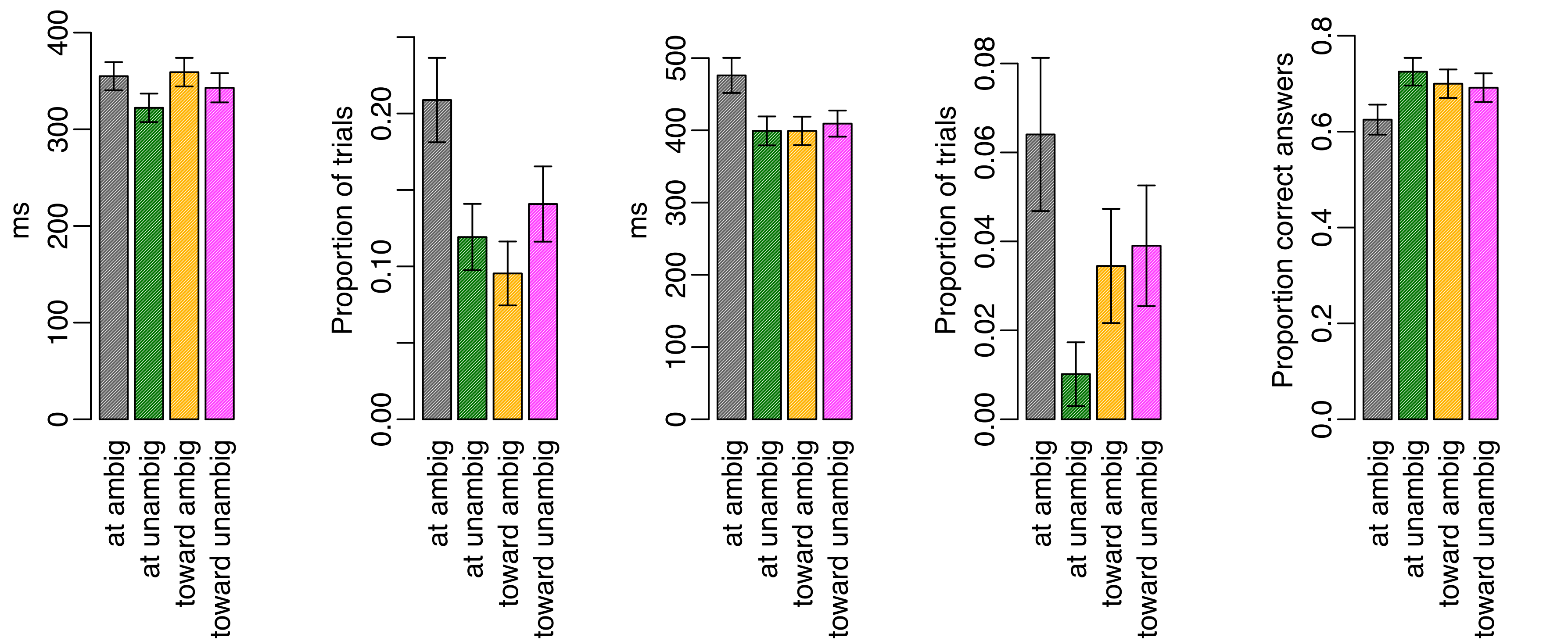
Experimental results

The coach smiled <u>at</u> the player <u>tossed</u>…



Experimental results

The coach smiled *<u>at</u>* *the player* *<u>tossed</u>*…



First-pass RT

Regressions out

Go-past RT

Go-past regressions

Comprehension accuracy

Application to structural forgetting

$$P(w_i \mid C) = \sum_{w_{1\ldots i-1}} P(w_i \mid w_{1\ldots i}) P(w_{1\ldots i-1} \mid C)$$

$$\text{Cost}(w_i \mid C) = \log \frac{1}{P(w_i \mid C)}$$

(Futrell & Levy, 2017)

Application to structural forgetting

- Noisy channel + surprisal = **noisy-context surprisal**: for a noisy input context *C* and next encountered word w_i:

$$P(w_i \mid C) = \sum_{w_{1\ldots i-1}} P(w_i \mid w_{1\ldots i})P(w_{1\ldots i-1} \mid C)$$

$$\text{Cost}(w_i \mid C) = \log \frac{1}{P(w_i \mid C)}$$

(Futrell & Levy, 2017)

Application to structural forgetting

- Noisy channel + surprisal = **noisy-context surprisal**: for a noisy input context C and next encountered word w_i:

$$P(w_i \mid C) = \sum_{w_{1\ldots i-1}} P(w_i \mid w_{1\ldots i}) P(w_{1\ldots i-1} \mid C)$$

$$\text{Cost}(w_i \mid C) = \log \frac{1}{P(w_i \mid C)}$$

- Comparison with humans: is the ungrammatical version of the sentence costlier?

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

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

?

(Futrell & Levy, 2017)

Application to structural forgetting

- Noisy channel + surprisal = **noisy-context surprisal**: for a noisy input context C and next encountered word w_i:

$$P(w_i \mid C) = \sum_{w_{1\ldots i-1}} P(w_i \mid w_{1\ldots i}) P(w_{1\ldots i-1} \mid C)$$

$$\text{Cost}(w_i \mid C) = \log \frac{1}{P(w_i \mid C)}$$

- Comparison with humans: is the ungrammatical version of the sentence costlier?

$\textit{Cost}($NOUN THAT NOUN THAT NOUN VERB VERB$) <$

$\textit{Cost}($NOUN THAT NOUN THAT NOUN VERB VERB VERB$)$

?

(Futrell & Levy, 2017)

Application to structural forgetting

- Noisy channel + surprisal = **noisy-context surprisal**: for a noisy input context C and next encountered word w_i:

$$P(w_i \mid C) = \sum_{w_{1\ldots i-1}} P(w_i \mid w_{1\ldots i}) P(w_{1\ldots i-1} \mid C)$$

$$\text{Cost}(w_i \mid C) = \log \frac{1}{P(w_i \mid C)}$$

- Comparison with humans: is the ungrammatical version of the sentence costlier?

$$\textit{Cost}(2\ \text{VERBS}) < \textit{Cost}(3\ \text{VERBS})$$

?

(Futrell & Levy, 2017)

Noisy-Context Surprisal Account of Structural Forgetting

- This turns out to work 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*

Generates sequences like:

NOUN VERB

NOUN PREP NOUN VERB

NOUN THAT VERB NOUN VERB

NOUN THAT NOUN VERB VERB

NOUN THAT NOUN THAT NOUN...

English: *s*=0.8 (Roland et al., 2007)

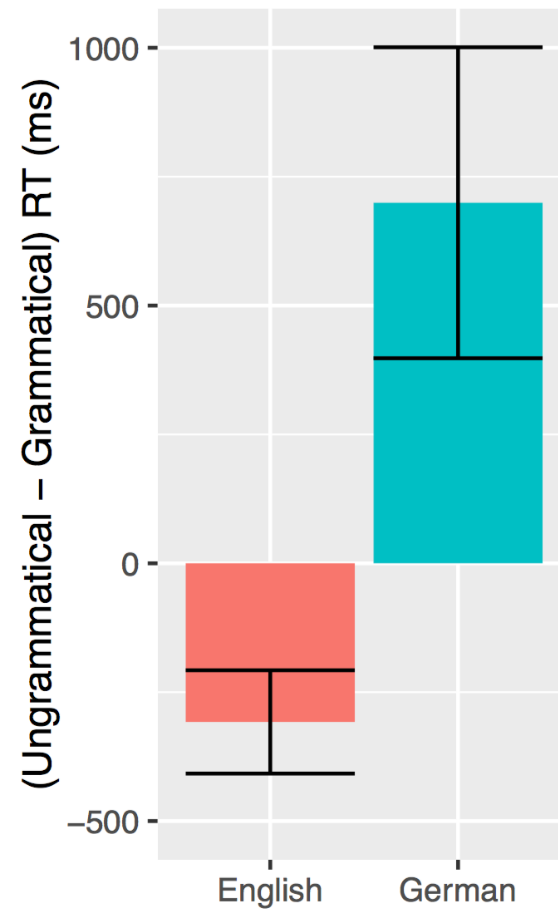
German *s*=0.0 (obligatorily verb-final)

3-verb (grammatical) version preferred

Model behavior

Futrell & Levy (2017)

Human reading time differences



Vasishth et al. (2010)

2-verb (ungrammatical) version preferred

Summary & open questions

- NLP and cognitive science offer each other a great deal
- NLP→cognitive science: formal theory-building for understanding human language processing
- Cognitive science→NLP: desiderata for human-like language processing systems
- Experimental methods can probe human cognitive state during language processing in remarkable detail
- Principles of rational analysis provide us guidance in theory building
- Scientific progress good, but many open questions:
  - How to fully characterize memory constraints in language?
  - Key principles of human conversational interaction?
  - Neural implementation of linguistic computations?
- ***These are great opportunities for everyone here!!!***

References I

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Anderson, J. R. (1990). *The adaptive character of human thought*. Hillsdale, NJ: Lawrence Erlbaum.

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.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

References II

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent Neural Network Grammars. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433.

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*(3), 554–578.

Frazier, L. (1985). Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, 129–189.

References III

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

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., 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*.

References IV

- Gibson, E., & Thomas, J. (1999). The perception of complex ungrammatical sentences as grammatical. *Language & Cognitive Processes*, *14*(3), 225–248.
- Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 10–18).
- 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.
- Henderson, J. (2004). Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04), main volume* (pp. 95–102).

References V

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, *24*(4), 613–632.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of acl*.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G., & Smith, N. A. (2017). What do Recurrent Neural Network Grammars learn about syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

References VI

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., & Blunsom, P. (2018). LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1426–1436). Melbourne, Australia: Association for Computational Linguistics.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

References VII

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.

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., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In *Proceedings of the 22nd conference on Neural Information Processing Systems (NIPS)*.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 1245994.

References VIII

- Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. Kieras & M. A. Just (Eds.), *New methods in reading comprehension*. Hillsdale, NJ: Earlbaum.
- Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.
- 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.
- Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*(3), 550–569.

References IX

Stern, M., Fried, D., & Klein, D. (2017). Effective inference for generative neural parsing. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1695–1700).

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562.

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 X

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.

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cognitive Evaluation
and
Language Evolution and Emergence

Richard Futrell
UC Irvine
rfutrell@uci.edu
@rljfutrell

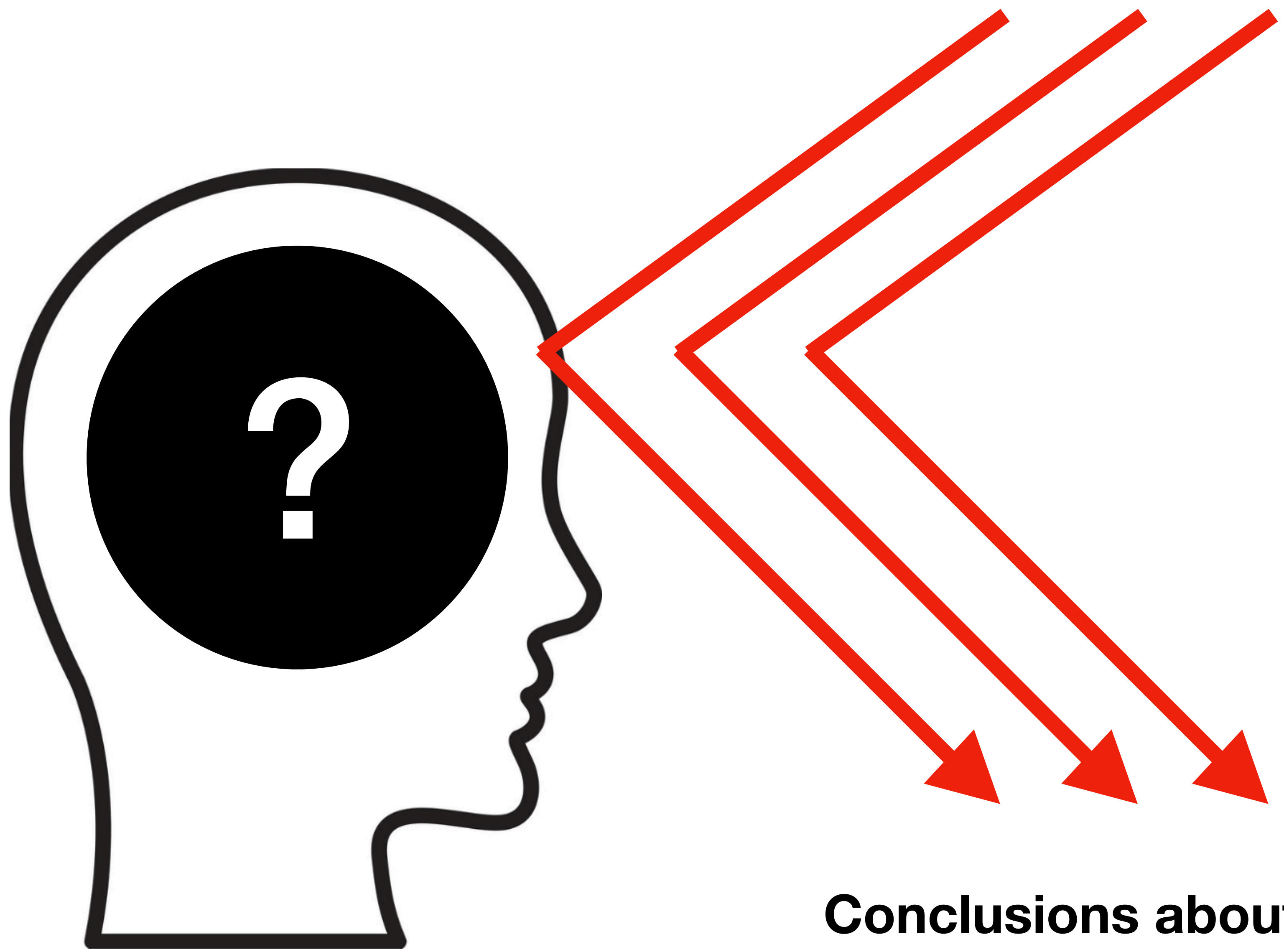
Goals of Part III

- Two sections:
  - **<u>Cognitive Evaluation:</u>**
    - Applying **methods from psycholinguistics and cognitive science** to **analyze neural networks**
    - Characterizing complex human behavior around language as a **target for NLP systems**
  - **<u>Language Evolution and Emergence</u>**
    - A recently-emerging exciting problem in NLP
    - Some highlights from **20 years of research from the field of Language Evolution** about under what circumstances **language-like codes** emerge in **agent-based models**

Cognitive Evaluation

Psycholinguistic Assessment

Battery of behavioral tests



Conclusions about…
form of linguistic knowledge,
data structures used in online processing,
sources of difficulty in production & comprehension
…

What Psycholinguists Do

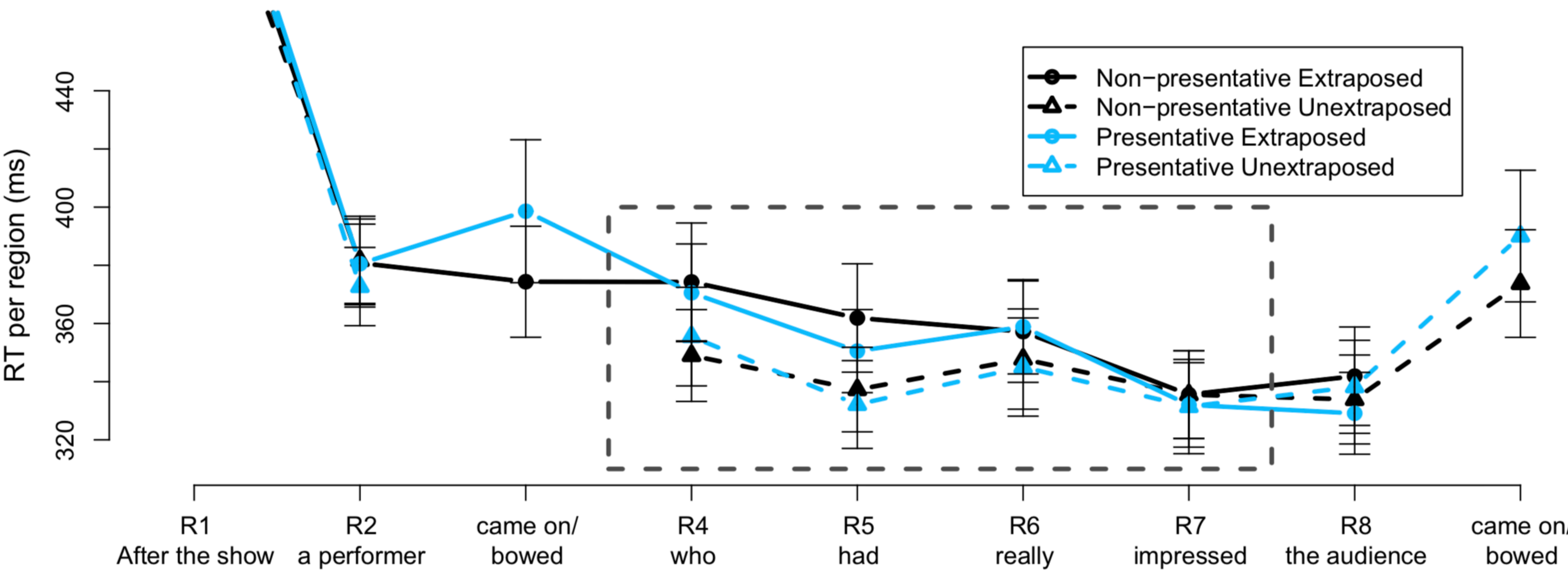
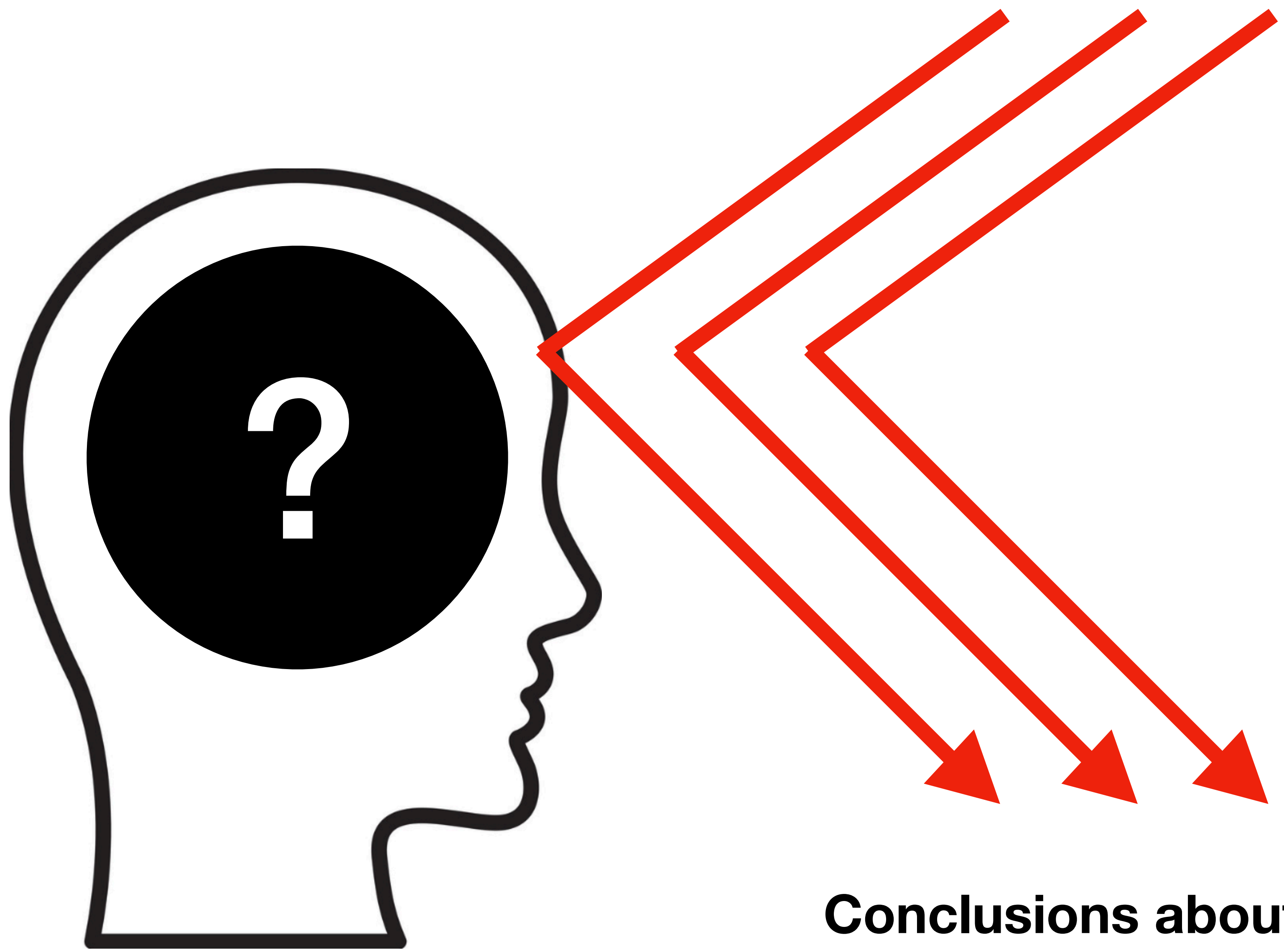


Fig. 2. Reading-time results as a function of region and condition for Experiment 1. Onset of the relative clause (first four words) is boxed.

Levy et al. (2012)

Psycholinguistic Assessment

Battery of behavioral tests



Conclusions about…
form of linguistic knowledge,
data structures used in online processing,
sources of difficulty in production & comprehension
…

Psycholinguistic Assessment

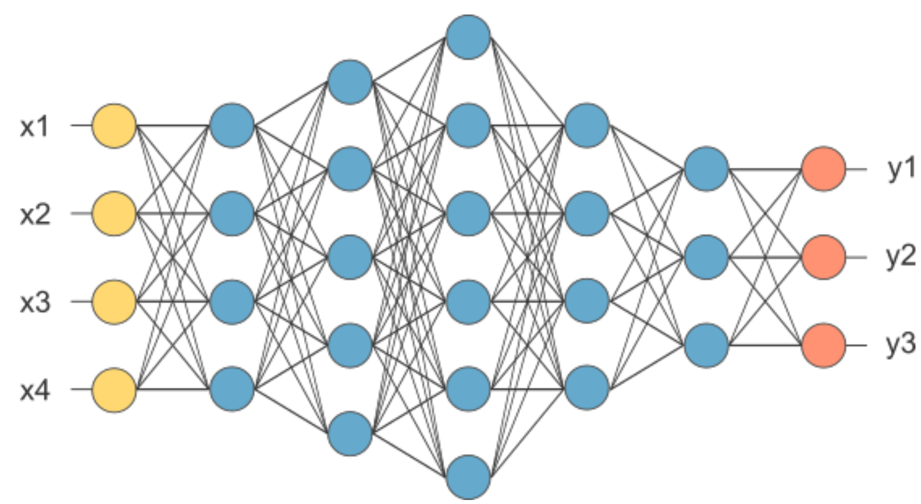
Battery of behavioral tests

Conclusions about…
form of linguistic knowledge,
data structures used in online processing,
sources of difficulty in production & comprehension
...

Probing NN Behavior

(a) *"The keys to the cabinet is on the table"

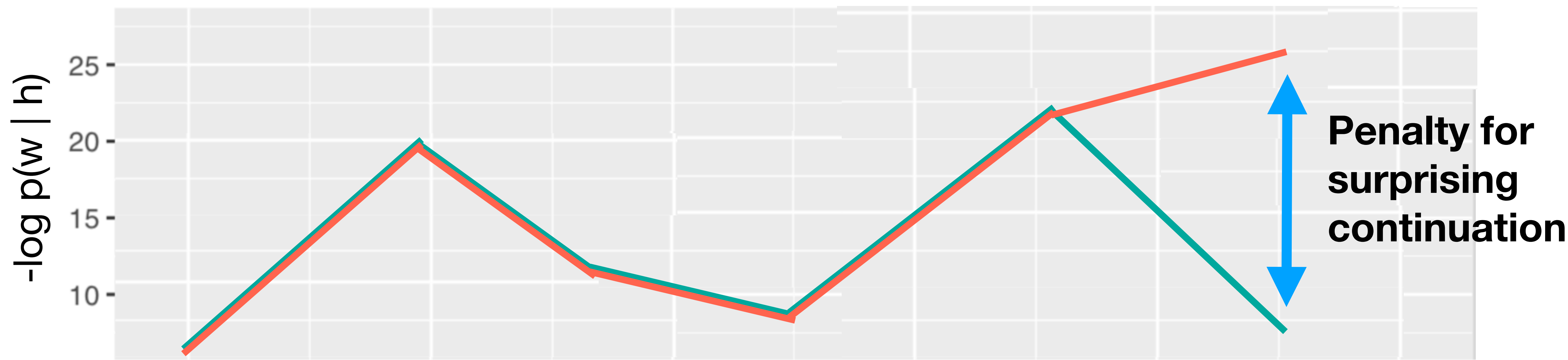
(b) "The keys to the cabinet are on the table"



(a) is SURPRISING! **(b) is UNSURPRISING**

Elman (1991, 1993); Linzen et al. (2016)

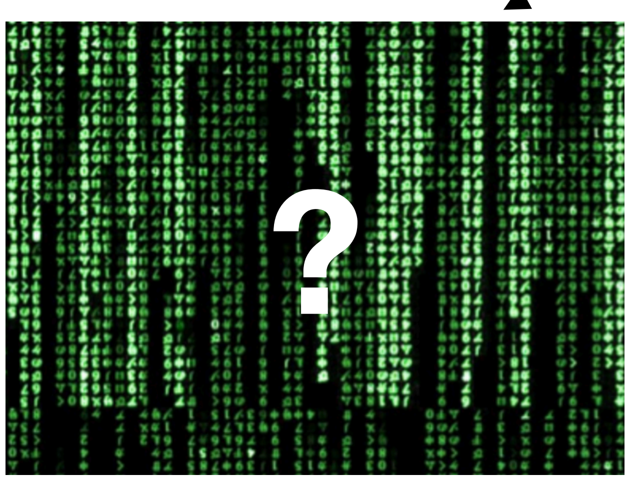
Probing NN Behavior

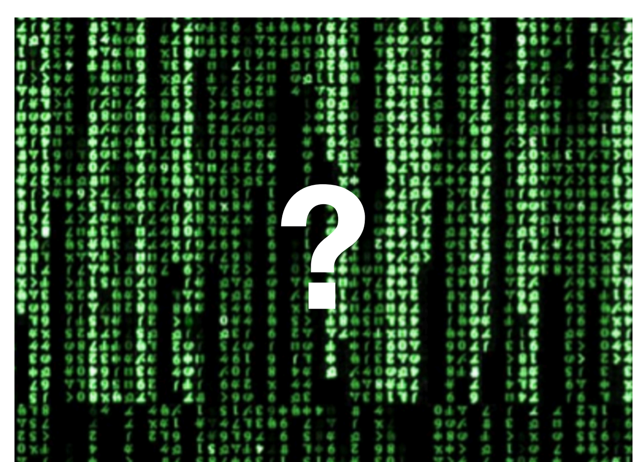


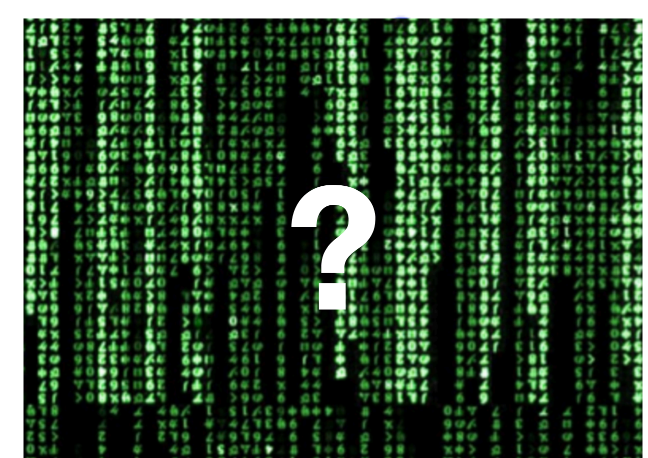
the keys to the cabinet are

the keys to the cabinet is

h_1 h_2 h_3 h_4 h_5







Linzen et al. (2016)

Phenomenon	Do NN Language Models Learn It?
Subject—Verb Agreement	? ✓ (Linzen et al., 2016; Gulordava et al., 2018)
Garden Path Effects	✓ ✓ ✓ (van Schijndel & Linzen, 2018a,b; Futrell et al., 2018, 2019)
Filler-Gap Dependencies	? ✓ ✓ (Chowdhury & Zamparelli, 2018; McCoy et al, 2018; Wilcox et al., 2018, 2019)
Island Constraints	? ✓ (some) (Chowdhury & Zamparelli, 2018; Wilcox et al., 2018)
NPI Licensing	✗ ✗ (Marvin & Linzen, 2018; Futrell et al., 2018)
Anaphor Agreement	✗ ✗ (Marvin & Linzen, 2018; Futrell et al., 2018)

What syntactic structures are easy vs. hard for NN language models?

- They find this contrast *easy*
  (**Filler-Gap Dependencies**: Wilcox et al., 2018, 2019).
  - *I know what the lion standing in the Serengeti devoured _ at sunrise.*
  - **I know what the lion standing in the Serengeti devoured a gazelle at sunrise.*
- They find this contrast *hard*
  (**Reflexive Anaphora**: Marvin & Linzen, 2018; Futrell et al., 2018)
  - *The king standing next to the queen saw himself*
  - **The king standing next to the queen saw herself*
- They **don't generalize in a clear way** across constructions that humans find similar.

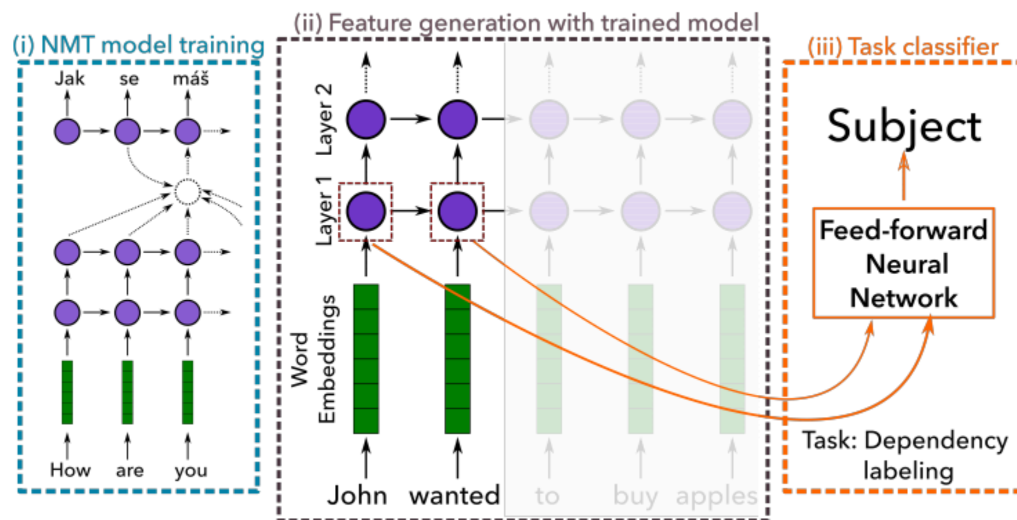
Targeted Evaluation Datasets

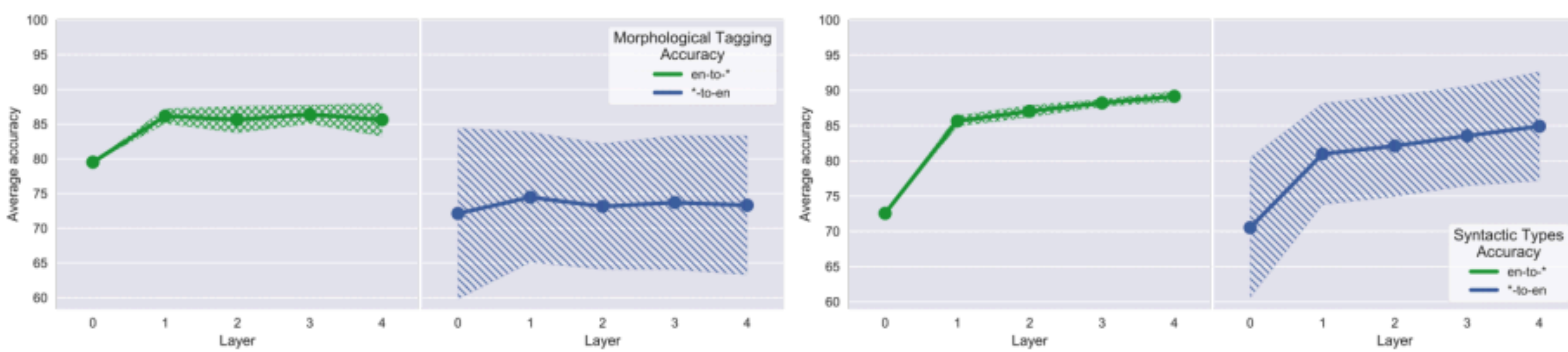
- Marvin & Linzen (2018)
- Used in e.g. Shen et al. (2019) [Ordered Neurons]

	ON-LSTM	LSTM
Short-Term Dependency		
SUBJECT-VERB AGREEMENT:		
Simple	0.99	**1.00**
In a sentential complement	0.95	**0.98**
Short VP coordination	0.89	**0.92**
In an object relative clause	0.84	**0.88**
In an object relative (no *that*)	0.78	**0.81**
REFLEXIVE ANAPHORA:		
Simple	**0.89**	0.82
In a sentential complement	**0.86**	0.80
NEGATIVE POLARITY ITEMS:		
Simple (grammatical vs. intrusive)	0.18	**1.00**
Simple (intrusive vs. ungrammatical)	**0.50**	0.01
Simple (grammatical vs. ungrammatical)	0.07	**0.63**
Long-Term Dependency		
SUBJECT-VERB AGREEMENT:		
Long VP coordination	**0.74**	**0.74**
Across a prepositional phrase	0.67	**0.68**
Across a subject relative clause	**0.66**	0.60
Across an object relative clause	**0.57**	0.52
Across an object relative (no *that*)	**0.54**	0.51
REFLEXIVE ANAPHORA:		
Across a relative clause	0.57	**0.58**
NEGATIVE POLARITY ITEMS:		
Across a relative clause (grammatical vs. intrusive)	0.59	**0.95**
Across a relative clause (intrusive vs. ungrammatical)	**0.20**	0.00
Across a relative clause (grammatical vs. ungrammatical)	**0.11**	0.04

Probing Classifiers

- Alain & Bengio (2016); Belinkov et al. (2018); Hupkes, Veldhoen & Zuidema (2018)





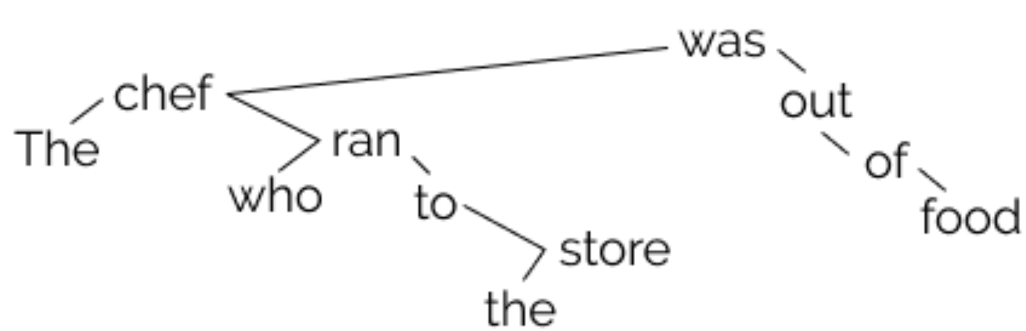
(a) Morphology

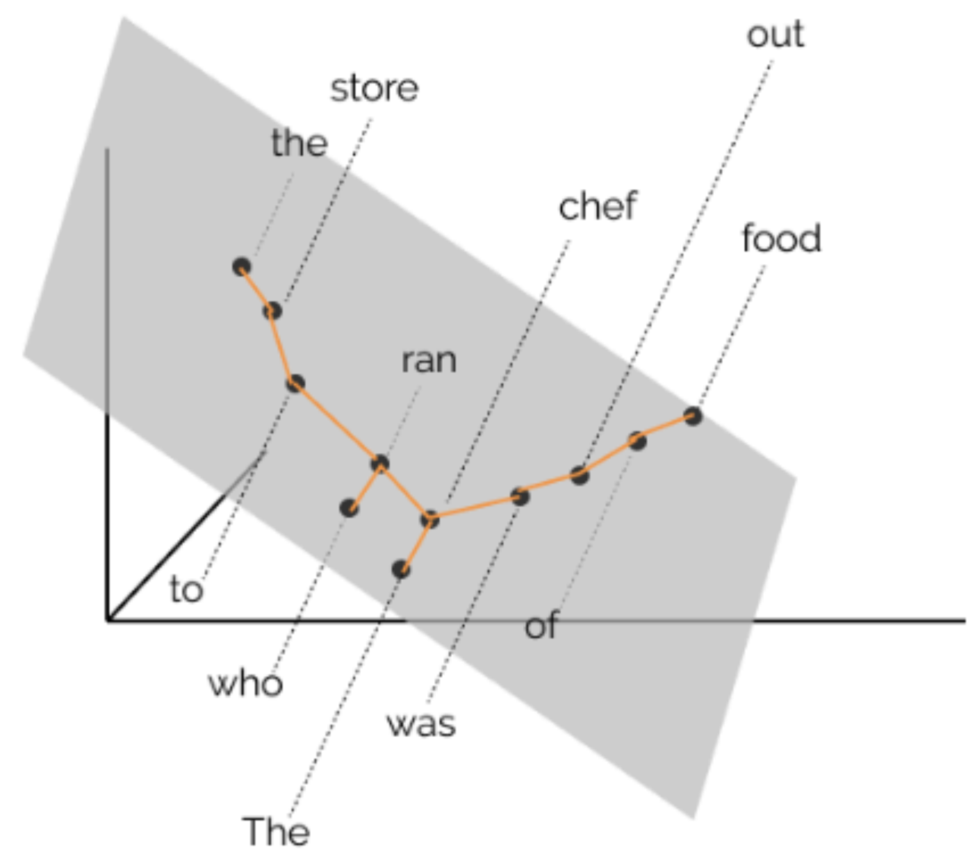
(b) Syntax

Similar to neuroscience methods: Wallis (2018)

Other Methods of Peering In

- Hewitt & Manning (2019): *Structural probe*: Does there exist a linear transformation of the contextual word embedding space such that the distances reflect syntactic parse trees?

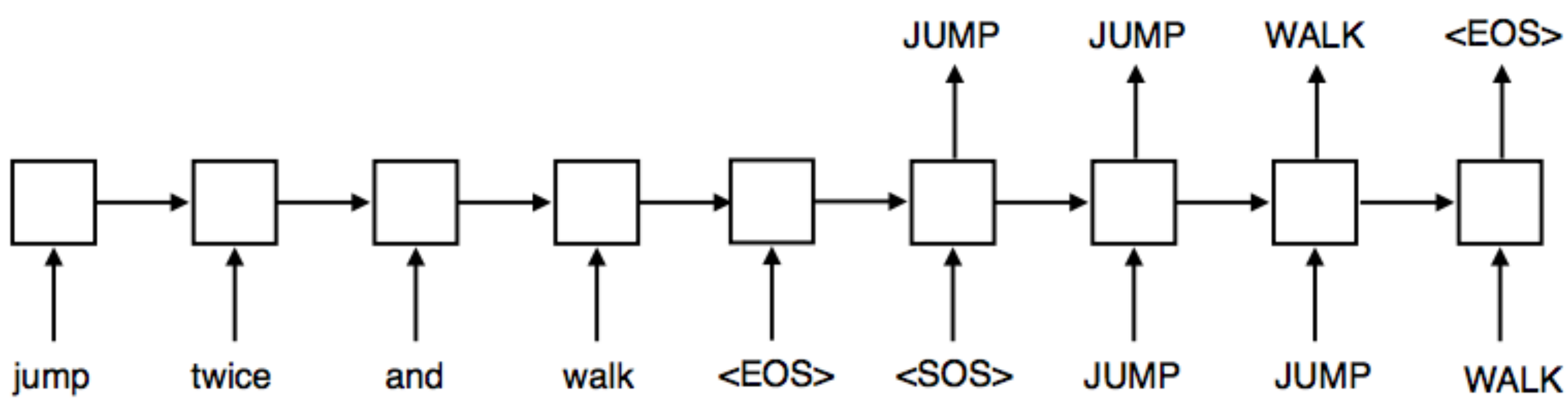




Sequence (to Sequence) Models

- Do generic sequence (to sequence) models show human-like generalization?

jump	⇒	JUMP
jump left	⇒	LTURN JUMP
jump around right	⇒	RTURN JUMP RTURN JUMP RTURN JUMP RTURN JUMP
turn left twice	⇒	LTURN LTURN
jump thrice	⇒	JUMP JUMP JUMP
jump opposite left and walk thrice	⇒	LTURN LTURN JUMP WALK WALK WALK
jump opposite left after walk around left	⇒	LTURN WALK LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP



run => RUN

Lake & Baroni (2018)

Sequence (to Sequence) Models

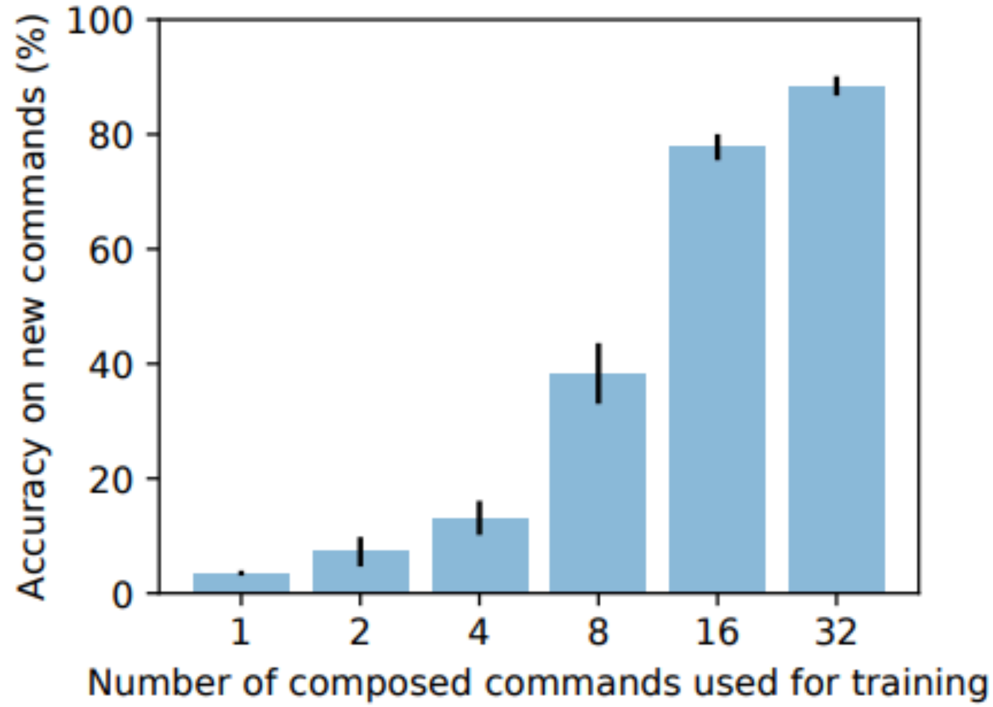


Figure 5. Zero-shot generalization after adding the primitive "jump" and some compositional "jump" commands. The model that performed best in generalizing from primitive "jump" only was retrained with different numbers of composed "jump" commands (x-axis) in the training set, and generalization was measured on new composed "jump" commands (y-axis). Each bar shows the mean over 5 runs with varying training commands along with the corresponding ± 1 SEM.

Lake & Baroni (2018)

Embedding Spaces

- Standard modern approach in NLP is to embed words and sentences into a metric space.
- Are human intuitions about word similarity well-modeled by a (Euclidean) metric space?

Word Similarity

vanish	disappear
behave	obey
belief	impression
muscle	bone
modest	flexible
hole	agreement

- Other human word similarity datasets:
  - Free-association Nelson Norms (Nelson et al., 1998)
  - Small World of Words (smallworldofwords.org)

SimLex

Embedding Spaces

- Standard modern approach in NLP is to embed words and sentences into a metric space.
- Are human intuitions about word similarity well-modeled by a (Euclidean) metric space?

Minimality:

$$\delta(a,b) \geq \delta(a,a) = 0.$$

Symmetry:

$$\delta(a,b) = \delta(b,a).$$

The triangle inequality:

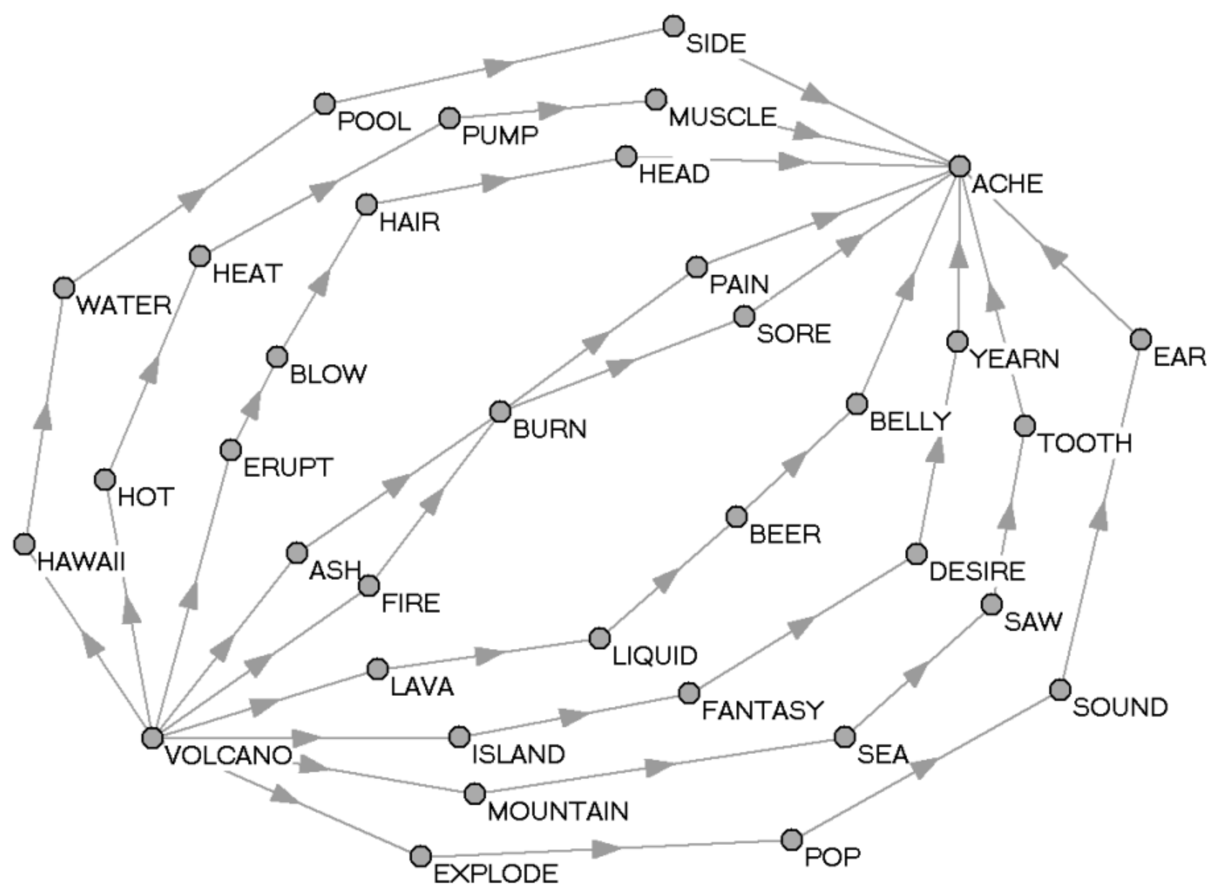
$$\delta(a,b) + \delta(b,c) \geq \delta(a,c).$$

- *keg, beer*
  - *vs.* *beer, keg*
- *cobra, snake*
  - *vs.* *snake, cobra*
- *meow, cat*
  - *vs.* *cat, meow*

Tversky (1977); Griffiths, Steyvers & Tenenbaum (2007)

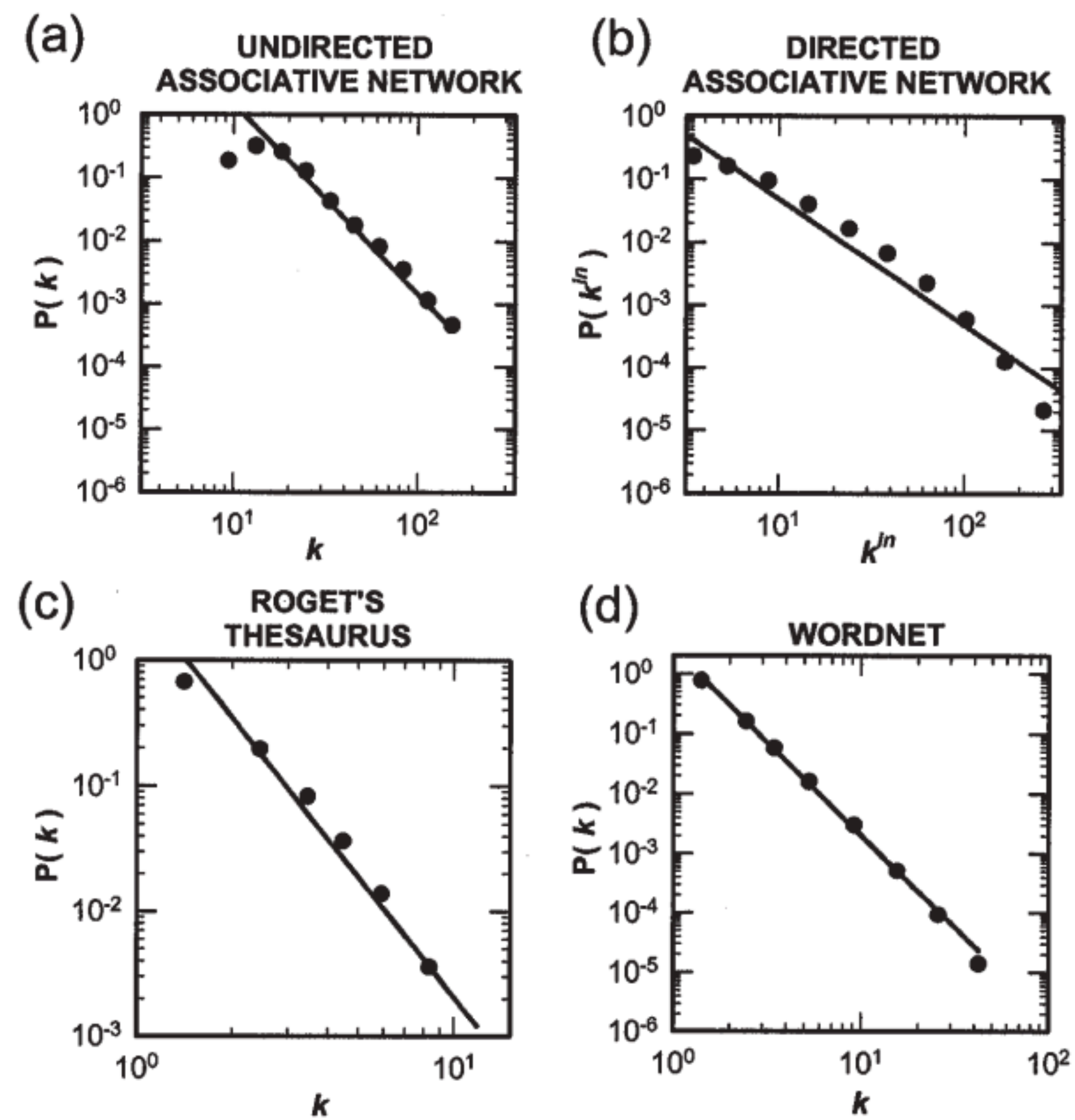
Semantic Networks

- Human word similarity judgments are best modeled using *semantic networks* (Steyvers & Tenenbaum, 2005).



Semantic Networks

- **Degree distributions** in human-derived semantic networks follow a **power law**:



Steyvers & Tenenbaum (2005)

Semantic Networks

- Degree distributions in semantic networks extracted from **distributional embeddings** follow an **exponential law**:

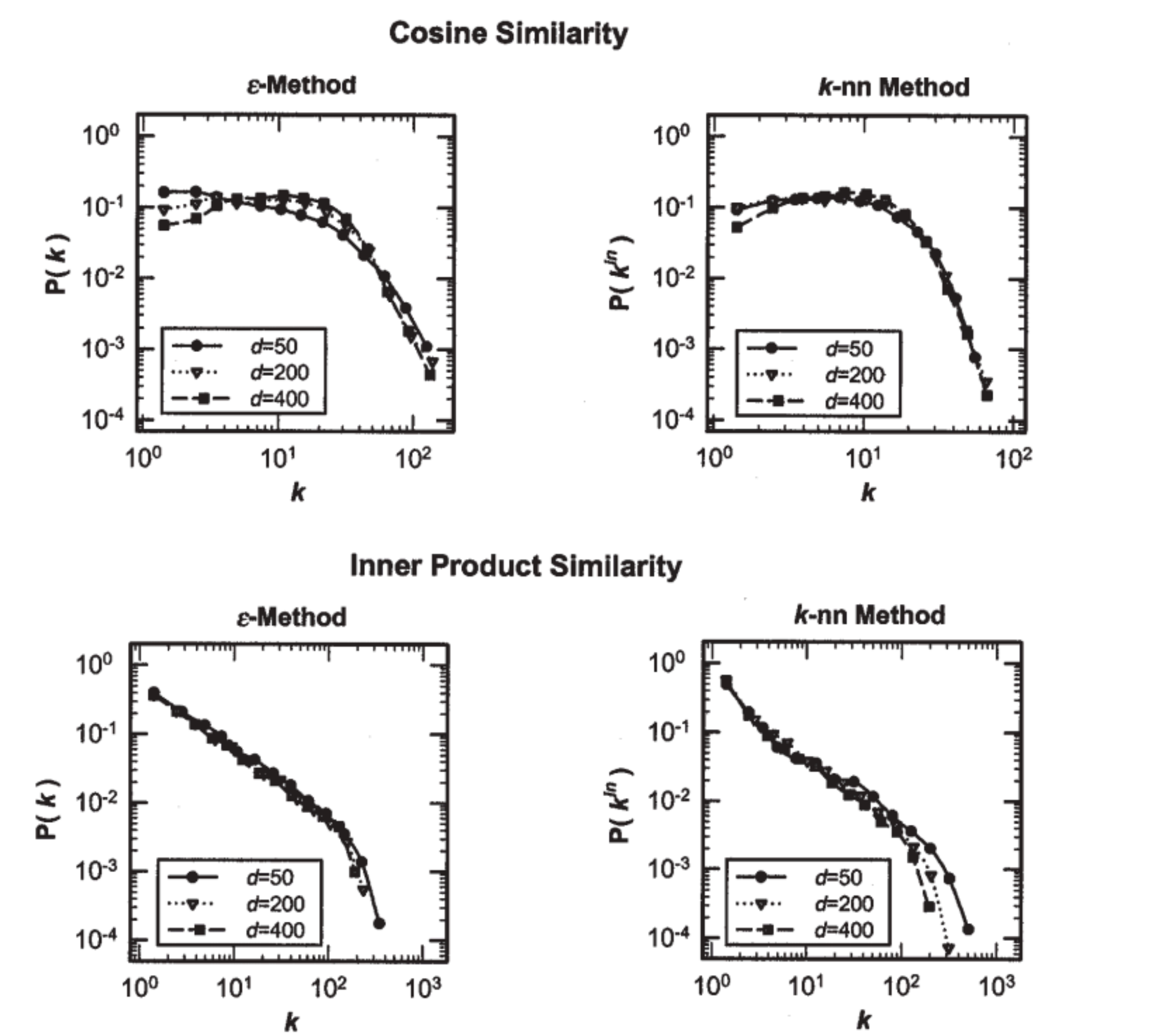


Fig. 8. The degree distributions for networks based on thresholded LSA spaces. For the ε-method, degree distributions of undirected networks are shown. For the *k*-nn method, the in-degree distributions are shown.

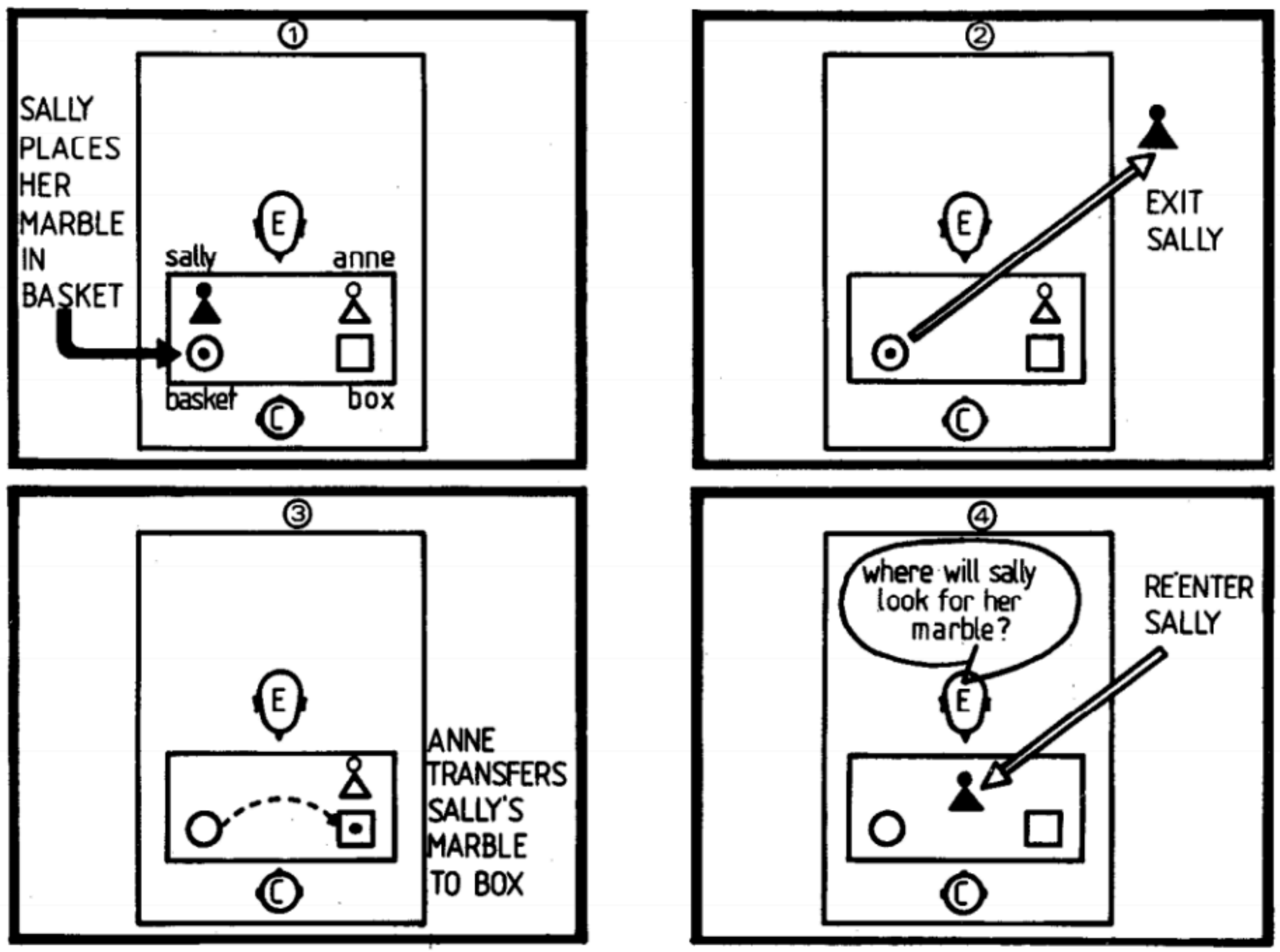
Steyvers & Tenenbaum (2005)

Embedding Spaces

- Distributionally-derived metric spaces do not capture human intuitions about word similarity, nor human free associations between words.
  - Human data violates **symmetry** and the **triangle inequality**, but follows **minimality**.
  - Human data implies a **power-law degree distribution** in semantic networks, but distributional methods give an **exponential degree distribution.**
- **Premetric spaces** (such as defined by KL divergence in information geometry) may be compatible with the human data.
- There is a rich modeling and experimental literature to draw from to define these spaces.

Tversky (1977); Steyvers & Tenenbaum (2005); Griffiths, Steyvers & Tenenbaum (2007)

Theory of Mind



Baron-Cohen et al. (1985)

Theory of Mind as a Question Answering Challenge

Mary went to the bathroom. John moved to the hallway. Mary travelled to the office. Where is Mary? A: office

bAbi (Weston et al., 2006)

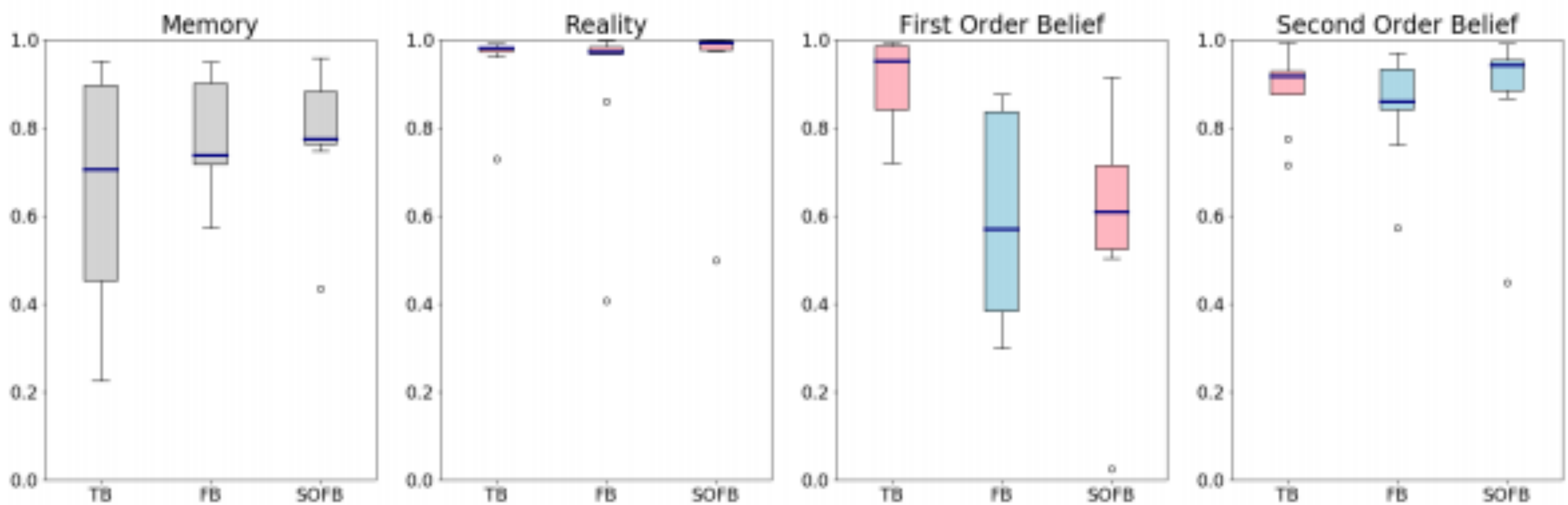
Second-order False Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.
Anne exited the kitchen.
Sally entered the kitchen.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Anne think that Sally searches for the milk?

Nematzadeh et al. (2018)

Question Answering

(d) Multiple Observer Model with memory size 50 evaluated on the *ToM* dataset.



Nematzadeh et al. (2018)

Cognitive Evaluation

- Behavioral work in cognitive science can feed into NLP in two ways:
  - Providing **careful analytical techniques** for evaluating black-box models.
    - Reveals **structural representations** and **inductive biases** in neural models.
  - Providing **challenging datasets and phenomena.**
    - **Compositionality & systematicity**
    - **Non-metric nature of human similarity judgments**
    - **Question answering involving Theory of Mind**
    - Many more!

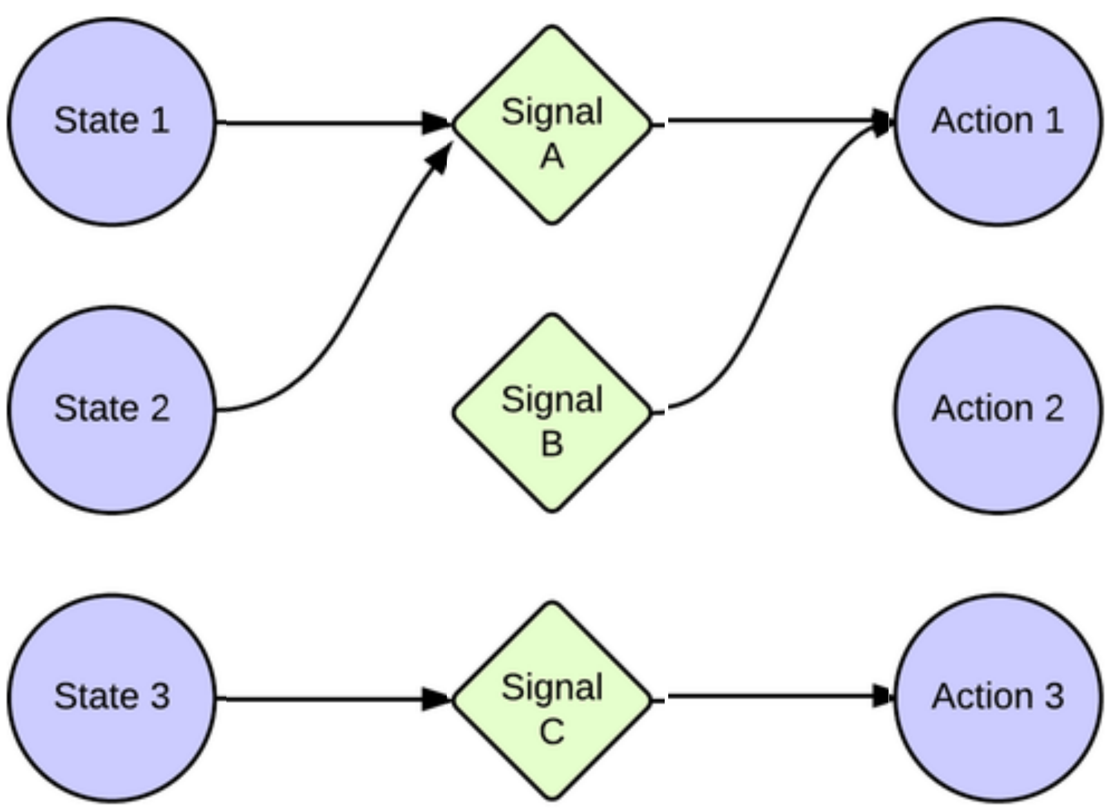
Language Evolution and Emergence

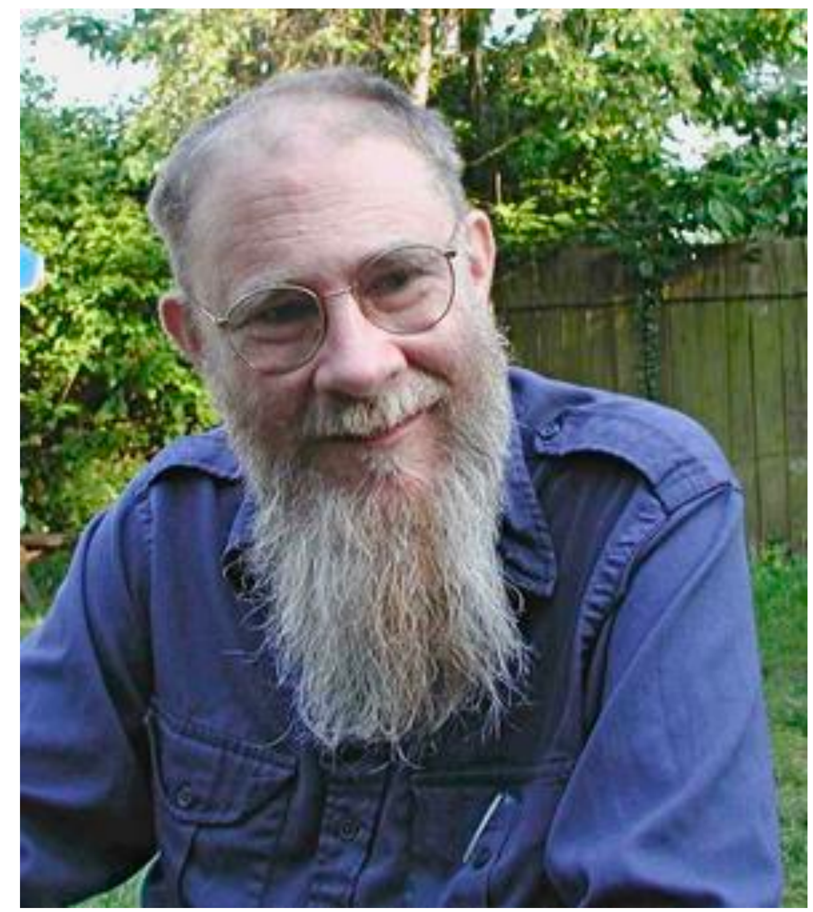
Language Evolution and Emergence

- If you have something like deep reinforcement learning agents trying to cooperate to solve a task, when will they **evolve a language-like code for communication?**
  - Havrylov & Titov (2017); Lazaridou et al. (2017, 2018); Mordatch & Abbeel (2017); Chaabouni et al. (2019); Lee et al. (2018)
- A potential new way to model *what language is.*
- I'll present some high-level takeaways from over 20 years of research in agent-based models of **Evolution of Language.**

Emergence of Symbols

- Simplest setting:
  David Lewis's Signaling Game

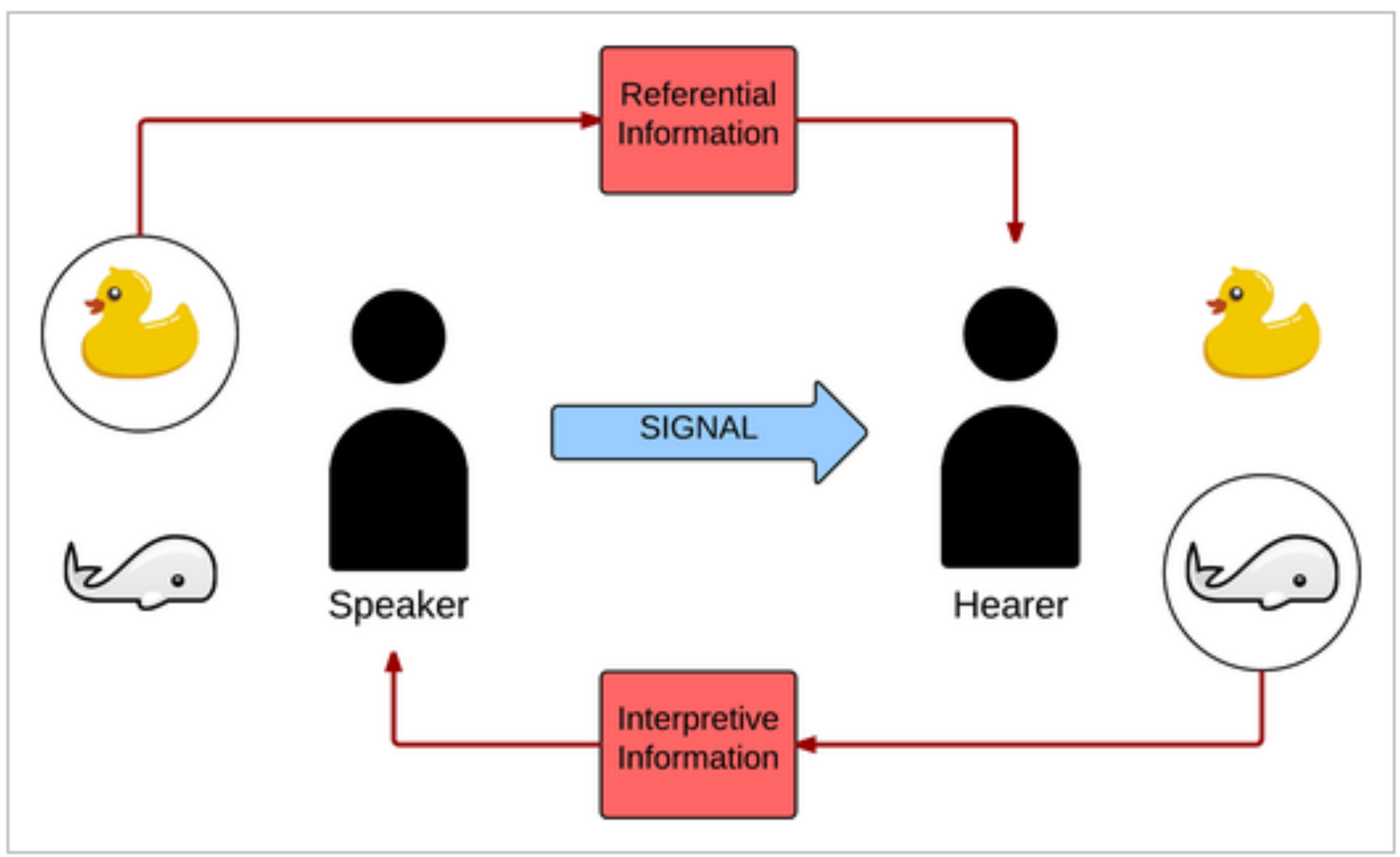




Lewis (1969). *Convention: A Philosophical Study*

Emergence of Symbols

- Three requirements for emergence of **learned signalling**:
  - **Availability of referential-interpretative information**
  - **Bias against ambiguity**
  - **Information loss**



Spike, Stadler, Kirby & Smith (2017)

From Symbols to Linguistic Structure

- Two hallmarks of human language:
  - **Combinatoriality**
  - **Compositionality**
- **Combinatoriality:**
  - A small set of **meaningless units** (phonemes/letters) **combine together** to form a large set of meaningful units (morphemes/words) according to an **arbitrary function.**

/k/ + /æ/ + /t/ = /kæt/, “cat”

From Symbols to Linguistic Structure

- Two hallmarks of human language:
  - **Combinatoriality**
  - **Compositionality**
- **Compositionality:**
  - A large set of **meaningful units** (morphemes/words) **combine together** to form an infinite set of meaningful sentences (Montague, 1970) according to a **simple function.**

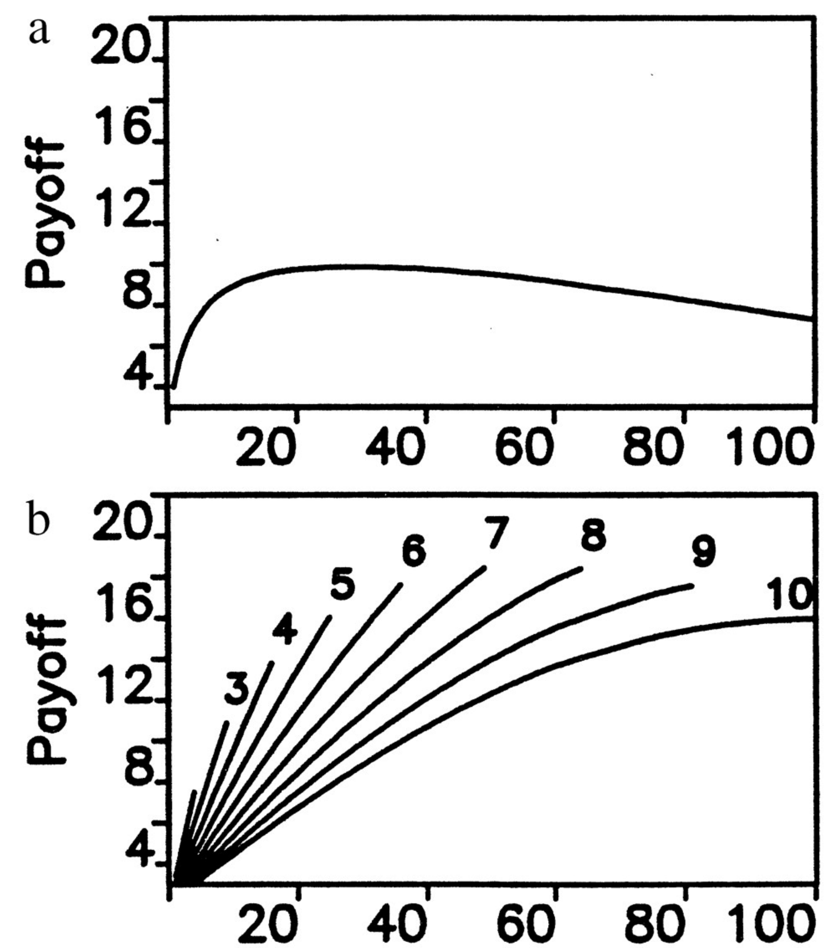
The + cat + meows

Meaning = f(f(the, cat), meows)

Duality of patterning

Emergence of Combinatoriality

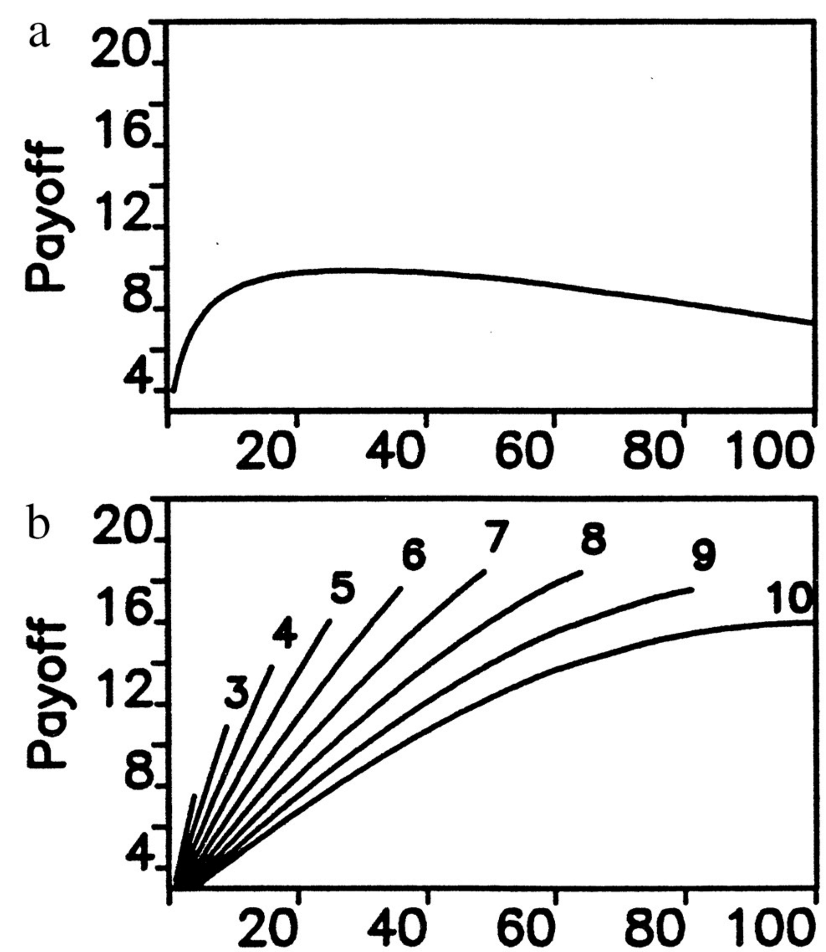
- Nowak & Krakauer (1999)
  - Imagine you are communicating about K objects in a Lewis signaling game.
  - Imagine it is *hard to perceive the difference* between signals.
  - Then it is better for a signal to consist of multiple discriminable parts (for redundancy), rather than each signal consisting of one atomic part.



Verhoef (2012); Tria (2012); Del Giudice (2012); Hofer, Tenenbaum & Levy (2019)

Emergence of Combinatoriality

- Related: Chaabouni et al. (2019) find that emergent languages in deep reinforcement learning agents favor long utterances due to discriminability.



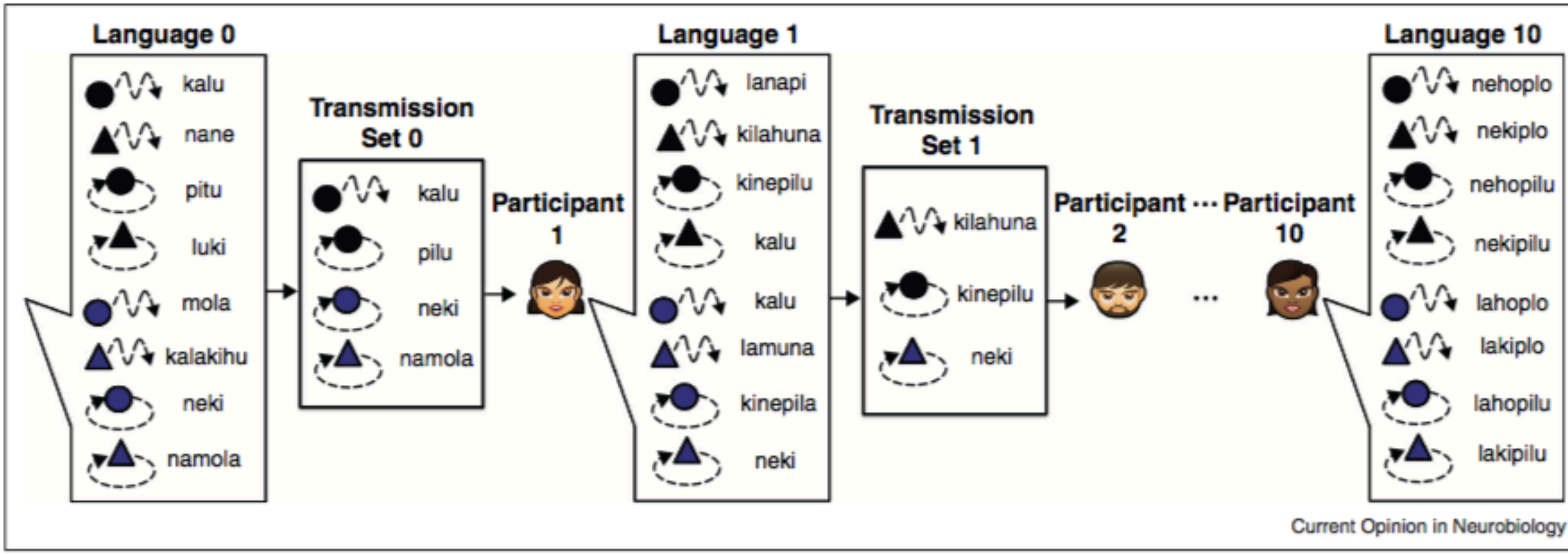
Defining Compositionality

Compositionality In intuitive terms, the representations computed by f are compositional if each $f(x)$ is determined by the structure of $D(x)$. Most discussions of compositionality, following Montague (1970), make this precise by defining a *composition* operation $\theta_a * \theta_b \mapsto \theta$ in the space of representations. Then the model f is compositional if it is a homomorphism from inputs to representations: we require that for any x with $D(x) = \langle D(x_a), D(x_b) \rangle$,

$$f(x) = f(x_a) * f(x_b) \; . \tag{1}$$

Montague (1970); Andreas (2019)

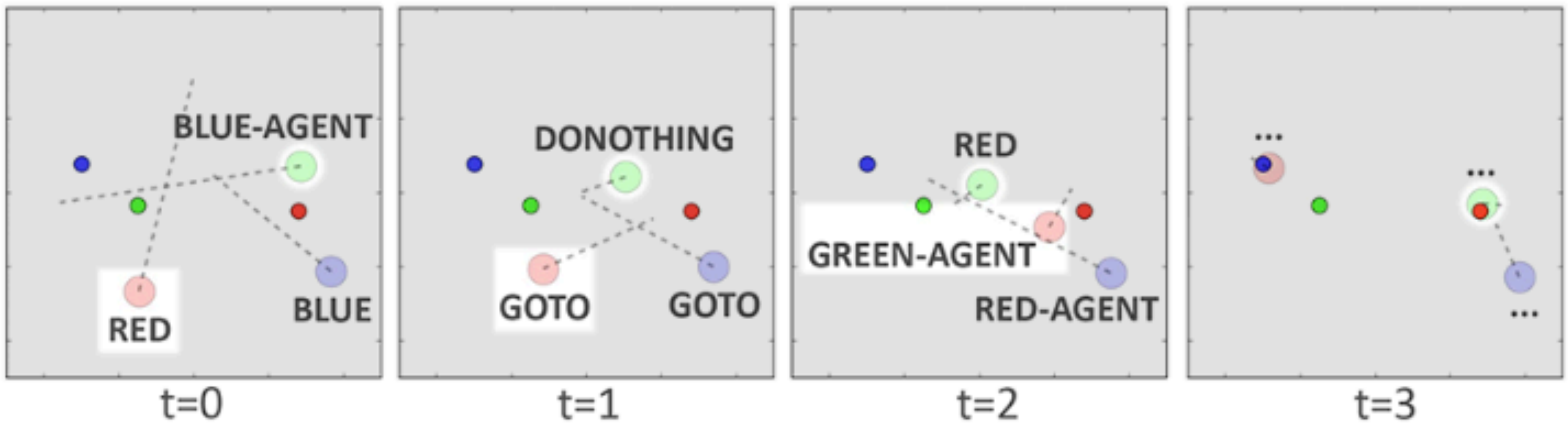
Emergence of Compositionality



- Iterated language learning experiments
- Compositionality emerges from a **transmission bottleneck** — which implements a **simplicity constraint.**
- **Compositionality = Simplicity + Communicativity**

Kirby, Cornish & Smith (2008)

Simple Compositionality in Agent-Based Modeling



In the above step-by-step run, at t=0 the red agent says a word corresponding to the red landmark (center right), then at t=1 says a word that is equivalent to 'Goto', then in t=2 says 'green-agent'. The green-agent hears its instructions and immediately moves to the red landmark.

- An implementation of compositionality = simplicity + communicativity

Abbeel & Mordatch (2017)

High-level Generalizations about Human Language

- **Modeling targets for language emergence experiments** beyond combinatoriality & compositionality.
  - The **set of phonemes** used in any language is much **smaller** than the set of all pronounceable phonemes used in all languages.
  - The set of phonemes in a language has a lot of **repeated substructure** in terms of phonetic features.
  - The set of phonemes in a language has a pressure to be **maximally acoustically distinct.**

High-level Generalizations about Human Language

- Languages usually have on the order of 10^1 **phonemes** and on the order of 10^4 **morphemes**: relatively invariant sequences of phonemes which correspond to atomic components of the meaning of an utterance.
  - A “**hierarachy problem**” for natural language.
  - In contrast, **animal communication systems** usually have 10^1 symbols with no internal structure.
- Morphemes vary in length; frequent/more predictable morphemes are shorter (Zipf, 1949; Piantadosi et al., 2011)
  - Compare Chaabouni et al. (2019)
- Morphemes contain a great deal of repeated substructure in their sequences of phonemes (**phonotactics**).
- Phonotactics is formally characterizable as ***k*-tier-based strictly local languages** with *k*=~*2* (Heinz, 2011)

High-level Generalizations about Human Language

- Utterances consist of sequences of multiple morphemes.
- Utterances vary in length.
- The overall meaning of an utterance is **compositional**: it is a **simple function** of the meanings of the morphemes and their order.
- There are an **unbounded number** of possible utterances.
- Utterances have **tree-like hierarchical structure**
- In these structures, **one word composes typically with one other word** in the computation of the meaning of the utterance (defining the **dependency tree**). This property is called **endocentricity** (Jakobson, 1961).
- The set of possible utterances is characterizable as a **Multiple Context Free Language** (Seki et al., 1991), with **block degree ~2** (Weir, 1988; Kuhlmann, 2013).

Language Evolution

- There is a vast literature! (see evolang.org)
    - **Evolution of Language Conference** every 2 years
- Requirements for *learned signaling*: **referential feedback, ambiguity avoidance, information loss**
- Requirements for *combinatoriality*: **noise in communication**
- Requirements for *compositionality*: **simplicity + communicativity**
- Natural language provides a number of modeling targets!

Wrapping Up

Wrapping Up

- **Cognitive modeling** provides **inspiration**, **challenges**, and **analytical tools** for NLP.
- **Language is a human object**—created by humans, for humans.
  - The human cognitive side is especially important!
- A vast unexplored territory in characterizing **human language learning**, **human language processing**, and **emergence of language**
  - The bottleneck in the field is a lack of computationally-skilled researchers!

Thanks all!