

# Processing models 3 – syntactic parsing

Reading: Crocker: Mechanisms for sentence processing

Jakub Dotlačil

Cognitive Modeling, 2019-2020

# INTRODUCTION

Two extra papers (optional reading):

- Hale, 2001, A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings of the 2nd Meeting of the North American Association for Computational Linguistics*
- Brasoveanu and Dotlačil. An extensible framework for mechanistic processing models: From representational linguistic theories to quantitative model comparison. *Proceedings of 2018 ICCM*.

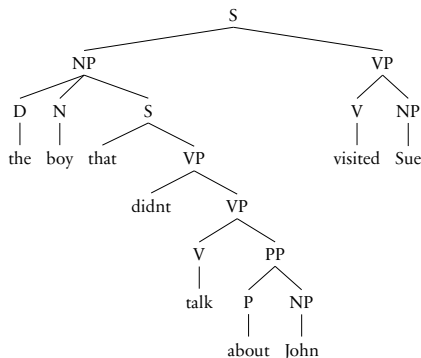
# INTRODUCTION

- Human-like machines should be able to communicate with us
- Communication should be in natural language and should be human-like

# INTRODUCTION: CONSTITUENCY

One dominant property of language: constituency  
Constituency models how strings of words are formed into meaning blocks

(1) The boy that didn't talk about John visited Sue.



e.g., *visited Sue* forms a constituent, VP

e.g., *didn't talk about John* forms a constituent, VP, hence, this is the negated event (negation does not affect *visited Sue*)

# INTRODUCTION

Parsing - the process of analyzing a string of words into constituents, conforming to the rules of grammar

- Constituents = syntactic structure

# PARSING AND TIME

- Parsing develops in time (is incremental)
- Parsing is eager

Before

the king



rides

his

beautiful

white

horse

is

always

groomed.



# Problematic sentence

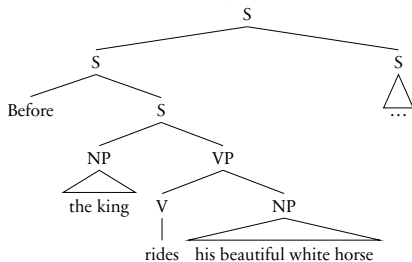
(2) Before the king rides his beautiful white horse is always groomed.

- *his beautiful white horse* – first interpreted as the object of *ride*
- later it turns out that *his beautiful white horse* is the subject of the matrix clause

# Locally ambiguous sentence

(3) Before the king rides his beautiful white horse...

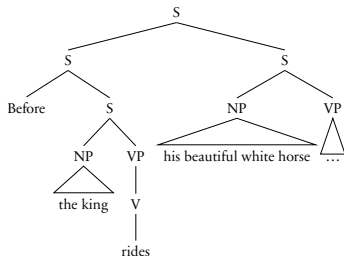
Option 1: *his beautiful white horse* – an object of *ride* (preferred)



# Locally ambiguous sentence

(4) Before the king rides his beautiful white horse...

Option 2: *his beautiful white horse* – the subject of the matrix clause  
(dispreferred)



## Locally ambiguous sentence

- (5) Before the king rides his beautiful white horse is always groomed.

## Locally ambiguous sentence

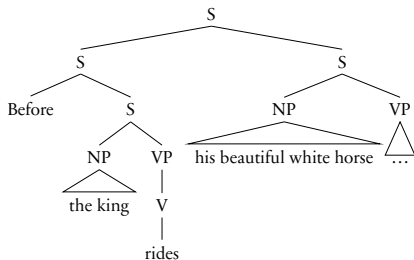
(5) Before the king rides his beautiful white horse is always groomed.

☞ Option 2 correct

# Locally ambiguous sentence

- (5) Before the king rides his beautiful white horse is always groomed.

☞ Option 2 correct



# Syntactic parsers are incremental and eager

Garden-path effects (observed cognitive difficulties when the dispreferred option turns out correct):

- 1 Before the king rides his horse is groomed.
- 2 The horse raced past the barn fell.

👉 syntactic parsing is incremental and eager

# Syntactic parsers are incremental and eager

- Readers construct syntactic structure incrementally and eagerly
- If the structure does not adhere to their expectations, they have cognitive difficulties
- Cognitive difficulties translate to increased reading times, a.o.



Modelling reading patterns:

- Abstract models
- Mechanistic models of reading

Modelling reading patterns in abstract

Hale, 2001, A Probabilistic Earley Parser as a Psycholinguistic Model.  
*Proceedings of the 2nd Meeting of the North American Association for  
Computational Linguistics*

# ABSTRACT MODELS, II

## Surprisal theory

Hale, 2001

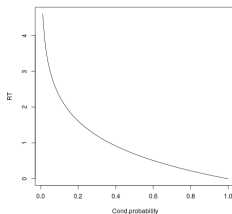
- reading patterns correlated with:  
 $P(w_i|w_0, \dots, w_{i-1})$ , where  $w_i$  – the current word
- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots, w_{i-1}, w_i)}{P(w_0, \dots, w_{i-1})}$
- $P(w_i|w_0, \dots, w_{i-1})$  correlates with reading patterns  
 $RT \approx \log(\frac{1}{P(w_i|w_0, \dots, w_{i-1})})$

# ABSTRACT MODELS, II

## Surprisal theory

Hale, 2001

- reading patterns correlated with:  
 $P(w_i|w_0, \dots, w_{i-1})$ , where  $w_i$  – the current word
- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots, w_{i-1}, w_i)}{P(w_0, \dots, w_{i-1})}$
- $P(w_i|w_0, \dots, w_{i-1})$  correlates with reading patterns  
 $RT \approx \log(\frac{1}{P(w_i|w_0, \dots, w_{i-1})})$



# ABSTRACT MODELS, III

## Surprisal theory

Hale, 2001

- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots, w_{i-1}, w_i)}{P(w_0, \dots, w_{i-1})}$
- Let's set  $i = 2$
- $P(w_2|w_0, w_1)$   
 $RT \approx \log(\frac{1}{P(w_2|w_0, w_1)})$

(6) The book fell.

$$RT(\text{fell}) \approx \log(\frac{1}{P(\text{fell}|\text{the}, \text{book})})$$

(7) The book sleeps.

$$RT(\text{sleeps}) \approx \log(\frac{1}{P(\text{sleeps}|\text{the}, \text{book})})$$

$\frac{P(\text{fell}|\text{the}, \text{book})}{P(\text{sleeps}|\text{the}, \text{book})} = 100$ , hence *sleeps* – cognitive difficulties (increased RTs) in (7)

# ABSTRACT MODELS, IV

## Surprisal theory

Hale, 2001

- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots, w_{i-1}, w_i)}{P(w_0, \dots, w_{i-1})}$

- Let's set  $i = 2$

- $P(w_2|w_0, w_1)$   
 $RT \approx \log(\frac{1}{P(w_2|w_0, w_1)})$

(8) The horse raced past the barn fell.

$$RT(\text{fell}) \approx \log(\frac{1}{P(\text{fell}|\text{the}, \text{barn})})$$

# ABSTRACT MODELS, IV

## Surprisal theory

Hale, 2001

- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots, w_{i-1}, w_i)}{P(w_0, \dots, w_{i-1})}$

- Let's set  $i = 2$

- $P(w_2|w_0, w_1)$

$$\text{RT} \approx \log\left(\frac{1}{P(w_2|w_0, w_1)}\right)$$

(8) The horse raced past the barn fell.

$$\text{RT}(\text{fell}) \approx \log\left(\frac{1}{P(\text{fell}|\text{the}, \text{barn})}\right)$$

(9) The horse that raced past the barn fell.

$$\text{RT}(\text{fell}) \approx \log\left(\frac{1}{P(\text{fell}|\text{the}, \text{barn})}\right)$$

## Surprisal theory enriched with syntax

Hale, 2001

- reading patterns correlated with:  
 $P(w_i | w_0, \dots, w_{i-1})$ , where  $w_i$  – the current word



## Surprisal theory enriched with syntax

Hale, 2001

- reading patterns correlated with:  
 $P(w_i | w_0, \dots, w_{i-1})$ , where  $w_i$  – the current word
- $P(w_0, \dots, w_{i-1})$  calculated as the sum of all joint probabilities of all the syntactic structures compatible with  $w_0, \dots, w_{i-1}$
- $P(w_0, \dots, w_{i-1}, w_i)$  calculated as another sum

## Surprisal theory enriched with syntax

Hale, 2001

- reading patterns correlated with:  
 $P(w_i|w_0, \dots, w_{i-1})$ , where  $w_i$  – the current word
- $P(w_0, \dots w_{i-1})$  calculated as the sum of all joint probabilities of all the syntactic structures compatible with  $w_0, \dots w_{i-1}$
- $P(w_0, \dots w_{i-1}, w_i)$  calculated as another sum
- $P(w_i|w_0, \dots, w_{i-1}) = \frac{P(w_0, \dots w_{i-1}, w_i)}{P(w_0, \dots w_{i-1})}$

$$\text{RT} \approx \log\left(\frac{1}{P(w_i|w_0, \dots, w_{i-1})}\right)$$

$$\text{RT} \approx \log\left(\frac{P(w_0, \dots w_{i-1})}{P(w_0, \dots w_{i-1}, w_i)}\right)$$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

- ➊  $S \rightarrow NP VP . (1.0)$
- ➋  $NP \rightarrow DT NN (1.0)$
- ➌  $DT \rightarrow the (0.5)$
- ➍  $DT \rightarrow a (0.5)$
- ➎  $NN \rightarrow boy$
- ➏  $VP \rightarrow V PP (0.3)$
- ➐  $VP \rightarrow V (0.7)$
- ➑  $V \rightarrow fell (0.5)$
- ➒  $V \rightarrow talked (0.5)$

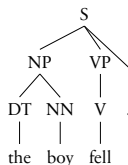
## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

- 1  $S \rightarrow NP VP . (1.0)$
- 2  $NP \rightarrow DT NN (1.0)$
- 3  $DT \rightarrow the (0.5)$
- 4  $DT \rightarrow a (0.5)$
- 5  $NN \rightarrow boy$
- 6  $VP \rightarrow V PP (0.3)$
- 7  $VP \rightarrow V (0.7)$
- 8  $V \rightarrow fell (0.5)$
- 9  $V \rightarrow talked (0.5)$

The boy fell.



## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

The boy fell.

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

- $P(\#) = 1$

④  $DT \rightarrow a (0.5)$

- stack: [S]

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

The boy fell.

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

- $P(\#) = 1$

④  $DT \rightarrow a (0.5)$

- stack: [S]

⑤  $NN \rightarrow boy$

- expand: [NP] [VP] [.] (1.0)

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

The boy fell.

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

- $P(\#) = 1$

④  $DT \rightarrow a (0.5)$

- stack: [S]

⑤  $NN \rightarrow boy$

- expand: [NP] [VP] [.] (1.0)

⑥  $VP \rightarrow V PP (0.3)$

- expand: [DT] [NN] [VP] [.] (1.0)

⑦  $VP \rightarrow V (0.7)$

- expand: [the] [NN] [VP] [.] (0.5)

⑧  $V \rightarrow fell (0.5)$

- $P(\#, the) = 0.5$

⑨  $V \rightarrow talked (0.5)$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

The boy fell.

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#) = 1$

- stack: [S]

- expand: [NP] [VP] [.] (1.0)

- expand: [DT] [NN] [VP] [.] (1.0)

- expand: [the] [NN] [VP] [.] (0.5)

- $P(\#, the) = 0.5$

- $\log_2(\frac{1}{0.5}) = 1$



## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$       The **boy** fell.

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

•  $P(\#, the) = 0.5$

⑤  $NN \rightarrow boy$

• stack: [the] [NN] [VP] [. ] (0.5)

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$       The **boy** fell.

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the) = 0.5$

- stack: [the] [NN] [VP] [. ] (0.5)

- scan: [NN] [VP] [. ]

- expand: [boy] [VP] [. ] (1.0)

- $P(\#, the, boy) = 0.5$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$       The **boy** fell.

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the) = 0.5$

- stack: [the] [NN] [VP] [.] (0.5)

- scan: [NN] [VP] [.]

- expand: [boy] [VP] [.] (1.0)

- $P(\#, the, boy) = 0.5$

- $\log_2\left(\frac{0.5}{0.5}\right) = 0$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

The boy fell.

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy) = 0.5$

- stack: [boy] [VP] [.] (0.5)

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

The boy fell.

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

•  $P(\#, the, boy) = 0.5$

• stack: [boy] [VP] [.] (0.5)

• scan: [VP] [.] (1.0)

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

The boy fell.

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy) = 0.5$

- stack: [boy] [VP] [.] (0.5)

- scan: [VP] [.] (1.0)

- expand (Op 1): [V] [PP] [.] (0.3)

- expand (Op 1): [fell] [PP] [.] (0.5)

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

The boy fell.

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy) = 0.5$

- stack: [boy] [VP] [.] (0.5)

- scan: [VP] [.] (1.0)

- expand (Op 1): [V] [PP] [.] (0.3)

- expand (Op 1): [fell] [PP] [.] (0.5)

- expand (Op 2): [V] [.] (0.7)

- expand (Op 2): [fell] [.] (0.5)

- $P(\#, the, boy, fell) = 0.25$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

The boy fell.

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy) = 0.5$

- stack: [boy] [VP] [.] (0.5)

- scan: [VP] [.] (1.0)

- expand (Op 1): [V] [PP] [.] (0.3)

- expand (Op 1): [fell] [PP] [.] (0.5)

- expand (Op 2): [V] [.] (0.7)

- expand (Op 2): [fell] [.] (0.5)

- $P(\#, the, boy, fell) = 0.25$

- $\log_2\left(\frac{0.5}{0.25}\right) = 1.0$



## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$       The boy fell.

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy, fell) = 0.25$

- stack (Op 1): [fell] [PP] [.] (0.075)

- scan (Op 1): [PP] [.] (0.075)

- scan (Op 2): [.] (0.175)

- $P(\#, the, boy, fell, .) = 0.175$

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

①  $S \rightarrow NP VP . (1.0)$

②  $NP \rightarrow DT NN (1.0)$       The boy fell.

③  $DT \rightarrow the (0.5)$

④  $DT \rightarrow a (0.5)$

⑤  $NN \rightarrow boy$

⑥  $VP \rightarrow V PP (0.3)$

⑦  $VP \rightarrow V (0.7)$

⑧  $V \rightarrow fell (0.5)$

⑨  $V \rightarrow talked (0.5)$

- $P(\#, the, boy, fell) = 0.25$

- stack (Op 1): [fell] [PP] [. ] (0.075)

- scan (Op 1): [PP] [. ] (0.075)

- scan (Op 2): [. ] (0.175)

- $P(\#, the, boy, fell, .) = 0.175$

- $\log_2\left(\frac{0.25}{0.175}\right) = 0.5$

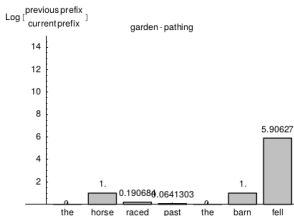
# ABSTRACT MODELS, V

## Surprisal theory

Hale, 2001

$$RT \approx \log\left(\frac{P(w_0, \dots, w_{i-1})}{P(w_0, \dots, w_{i-1}, w_i)}\right)$$

1.0	S	→	NP VP .
0.876404494831	NP	→	DT NN
0.123595505169	NP	→	NP VP
1.0	PP	→	IN NP
0.171428571172	VP	→	VBD PP
0.752380952552	VP	→	VCN PP
0.0761904762759	VP	→	VBD
1.0	DT	→	<i>the</i>
0.5	NN	→	<i>horse</i>
0.5	NN	→	<i>barn</i>
0.5	VBD	→	<i>fell</i>
0.5	VBD	→	<i>raced</i>
1.0	VBN	→	<i>raced</i>
1.0	IN	→	<i>past</i>



## ABSTRACT MODELS, VI, SUMMARY

- Surprisal theory – cognitive difficulties on word  $w$  correlate with conditional probabilities of  $w$  given the preceding words

# ABSTRACT MODELS, VI, SUMMARY

- Surprisal theory – cognitive difficulties on word  $w$  correlate with conditional probabilities of  $w$  given the preceding words
- Surprisal theory can be combined with a parsing model (top-down parser) to generate predictions for arbitrary grammatical sentences

## ABSTRACT MODELS, VI, CRITICISM

Hale, 2014:

*[I used to] explain garden path effects in terms of probabilistic grammars. This stage of my own development seems to have paralleled the one John Anderson was in during the late 1980s, at the height of his fascination with rational analysis. Since that time, I have retraced Anderson's steps and returned to algorithmic models. I did this because I wondered, what would I say to a person who walked in my door asking how people understand English sentences? An earnest questioner would not be satisfied with an explanation that starts with "well, people know these conditional probabilities and these equations show that the optimal thing to do is...." Real people want to know how it works. For them, a good explanation is a causal explanation.*

1 Introduction

2 Abstract models of parsing – parsing as rational analysis

3 Mechanistic models of reading

# MECHANISTIC MODELS OF READING

Brasoveanu and Dotlačil. An extensible framework for mechanistic processing models: From representational linguistic theories to quantitative model comparison. *Proceedings of 2018 ICCM*.

- parsing itself embedded in an independently motivated, general cognitive architecture
- Soar

Hale, 2014; Young and Lewis, 1999

- Adaptive Control of Thought-Rational (ACT-R)

Engelmann et al., 2013; Lewis and Vasishth, 2005; Nicenboim and Vasishth, 2018; Rij, 2012; Taatgen and Anderson, 2002; Vasishth et al., 2008



# PREVIOUS RELATED RESEARCH IN ACT-R

## ACT-R parsers

Lewis and Vasishth, 2005

See also Brasoveanu and Dotlačil, 2015; Engelmann et al., 2016; Rij, 2012; Vogelzang et al., 2017

# PREVIOUS RELATED RESEARCH IN ACT-R

## ACT-R parsers

Lewis and Vasishth, 2005

See also Brasoveanu and Dotlačil, 2015; Engelmann et al., 2016; Rij, 2012; Vogelzang et al., 2017

- parsers are built as part of Adaptive Control of Thought-Rational (ACT-R)

Anderson and Lebiere, 1998; Anderson and Schooler, 1991

- incremental construction of syntactic structures
- parsers capture retrieval interference, garden path

# PREVIOUS RELATED RESEARCH IN ACT-R

## ACT-R parsers

Lewis and Vasishth, 2005

See also Brasoveanu and Dotlačil, 2015; Engelmann et al., 2016; Rij, 2012; Vogelzang et al., 2017

- parsers are built as part of Adaptive Control of Thought-Rational (ACT-R)

Anderson and Lebiere, 1998; Anderson and Schooler, 1991

- incremental construction of syntactic structures
- parsers capture retrieval interference, garden path

## Limitations:

- parsers manually built (parsing rules are hand-coded)

# DATA-DRIVEN PARSING IN ACT-R

- No hand-coding of parsing rules
- Reading times are by-product of parsing itself (unlike, e.g., Hale 2001, cf. Hale 2014)
- see also Demberg et al., 2013

# TRANSITION-BASED (DETERMINISTIC) PARSING

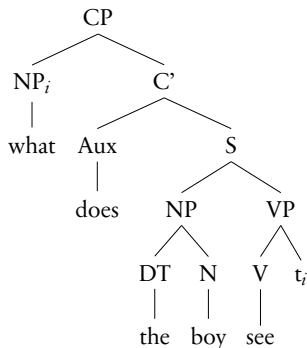
- Only one path explored (deterministic parser, see Crocker)
- Shift-reduce, stack-based parsing algorithm (see Crocker)
- Context-free grammar parser
- The parser includes gaps (traces)

Nivre, 2004

# TRANSITION-BASED (DETERMINISTIC) PARSING

- Only one path explored (deterministic parser, see Crocker)
- Shift-reduce, stack-based parsing algorithm (see Crocker)
- Context-free grammar parser
- The parser includes gaps (traces)

Nivre, 2004 (10)      What does the boy see?



# PARSER'S PROPERTIES

- A stack  $\mathcal{S}$  of trees built so far
- A queue of  $\mathcal{W}$  of upcoming words with their syntactic category
- Parser's job: function from stacks, queues to **actions**
- **actions** are:

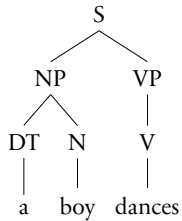
# PARSER'S PROPERTIES

- A stack  $\mathcal{S}$  of trees built so far
- A queue of  $\mathcal{W}$  of upcoming words with their syntactic category
- Parser's job: function from stacks, queues to **actions**
- **actions** are:
  - shift (put word with its category from  $\mathcal{W}$  to  $\mathcal{S}$ )
  - reduce top of  $\mathcal{S}$  (+label of reduce)
  - postulate gap and resolve it to antecedent



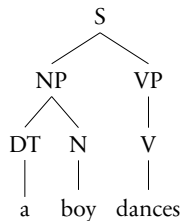
## PARSER – TRAINING EXAMPLE

(11) A boy dances.



# PARSER – TRAINING EXAMPLE

(11) A boy dances.



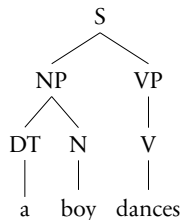
$\mathcal{W} = [\langle a, DT \rangle]$

① shift

$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

# PARSER – TRAINING EXAMPLE

(11) A boy dances.



$\mathcal{W} = [\langle a, DT \rangle]$

① shift

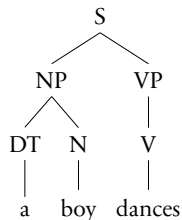
② shift

$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

$\mathcal{S} = [\langle a, DT \rangle, \langle boy, N \rangle], \mathcal{W} = [\langle dances, V \rangle]$

# PARSER – TRAINING EXAMPLE

(11) A boy dances.



$\mathcal{W} = [\langle a, DT \rangle]$

① shift

② shift

③ reduce

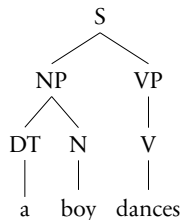
$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

$\mathcal{S} = [\langle a, DT \rangle, \langle boy, N \rangle], \mathcal{W} = [\langle dances, V \rangle]$

$\mathcal{S} = [\langle boy, NP \rangle], \mathcal{W} = [\langle dances, V \rangle]$

# PARSER – TRAINING EXAMPLE

(11) A boy dances.



$\mathcal{W} = [\langle a, DT \rangle]$

① shift

$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

② shift

$\mathcal{S} = [\langle a, DT \rangle, \langle boy, N \rangle], \mathcal{W} = [\langle dances, V \rangle]$

③ reduce

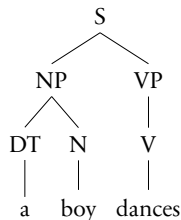
$\mathcal{S} = [\langle boy, NP \rangle], \mathcal{W} = [\langle dances, V \rangle]$

④ shift

$\mathcal{S} = [\langle boy, NP \rangle, \langle dances, V \rangle]$

# PARSER – TRAINING EXAMPLE

(11) A boy dances.



$\mathcal{W} = [\langle a, DT \rangle]$

① shift

② shift

③ reduce

④ shift

⑤ reduce

$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

$\mathcal{S} = [\langle a, DT \rangle, \langle boy, N \rangle], \mathcal{W} = [\langle dances, V \rangle]$

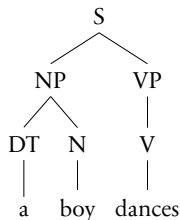
$\mathcal{S} = [\langle boy, NP \rangle], \mathcal{W} = [\langle dances, V \rangle]$

$\mathcal{S} = [\langle boy, NP \rangle, \langle dances, V \rangle]$

$\mathcal{S} = [\langle boy, NP \rangle, \langle dances, VP \rangle]$

# PARSER – TRAINING EXAMPLE

(11) A boy dances.



$\mathcal{W} = [\langle a, DT \rangle]$

① shift

$\mathcal{S} = [\langle a, DT \rangle], \mathcal{W} = [\langle boy, N \rangle]$

② shift

$\mathcal{S} = [\langle a, DT \rangle, \langle boy, N \rangle], \mathcal{W} = [\langle dances, V \rangle]$

③ reduce

$\mathcal{S} = [\langle boy, NP \rangle], \mathcal{W} = [\langle dances, V \rangle]$

④ shift

$\mathcal{S} = [\langle boy, NP \rangle, \langle dances, V \rangle]$

⑤ reduce

$\mathcal{S} = [\langle boy, NP \rangle, \langle dances, VP \rangle]$

⑥ reduce

$\mathcal{S} = [\langle dances, S \rangle]$

(12) A boy dances.

$\mathcal{S} = [\langle \text{boy}, \text{D} \rangle, \langle \text{boy}, \text{N} \rangle], \mathcal{W} = [\langle \text{dances}, \text{V} \rangle]$

- How will we find out what **action** to take? (Classification problem)
- ACT-R memory search as the classifier



# MEMORY AND ACTIVATION

any past **action** is a chunk and is stored in memory

- retrieval of a chunk from memory dependent on the activation of that chunk Anderson, 1990

# MEMORY AND ACTIVATION

any past **action** is a chunk and is stored in memory

- retrieval of a chunk from memory dependent on the activation of that chunk Anderson, 1990

# MEMORY AND ACTIVATION

any past **action** is a chunk and is stored in memory

- retrieval of a chunk from memory dependent on the activation of that chunk Anderson, 1990

(13)     Activation:  $\boxed{A_i} = B_i + SA_i$

- $B_i$ : base-level activation of chunk  $i$
- $SA_i$ : spreading activation of chunk  $i$
- $A_i$ : activation

# MEMORY AND ACTIVATION

any past **action** is a chunk and is stored in memory

- retrieval of a chunk from memory dependent on the activation of that chunk Anderson, 1990

(13)      Activation:  $\boxed{A_i} = B_i + SA_i$

- $B_i$ : base-level activation of chunk  $i$
- $SA_i$ : spreading activation of chunk  $i$
- $A_i$ : activation
- If chunk  $X$  has higher activation than chunk  $Y$ , then:
  - $X$  is recalled,  $Y$  is not
  - $X$  has higher chance of not being forgotten compared to  $Y$
  - $X$  is recalled faster than  $Y$  would be

# ACTIVATIONS

$$\boxed{A_i} = B_i + SA_i$$

# ACTIVATIONS

$$\boxed{A_i} = B_i + SA_i$$

$$\boxed{B_i = \log \left( t_i^{-\frac{1}{2}} \right)}$$

$t_i$ : time since the use of chunk  $i$

# ACTIVATIONS

$$\boxed{A_i} = B_i + SA_i$$

$$\boxed{SA_i = S_{1i} + S_{2i} + \dots}$$

$$\boxed{B_i = \log \left( t_i^{-\frac{1}{2}} \right)}$$

$S_{1i}$ : activation spreading from value 1

$S_{2i}$ : activation spreading from value 1

$t_i$ : time since the use of chunk  $i$

# ACTIVATIONS

$$\boxed{A_i} = B_i + SA_i$$

$$\boxed{SA_i = S_{1i} + S_{2i} + \dots}$$

$$\boxed{B_i = \log \left( t_i^{-\frac{1}{2}} \right)}$$

$t_i$ : time since the use of chunk  $i$

$S_{1i}$ : activation spreading from value 1

$S_{2i}$ : activation spreading from value 1

$$S_{1i} = \begin{cases} 0 & \text{if 1 not} \\ & \text{present in } i \\ S - \log(fan_1) & \text{otherwise} \end{cases}$$



# ACTIVATIONS

$$\boxed{A_i} = B_i + SA_i$$

$$\boxed{SA_i = S_{1i} + S_{2i} + \dots}$$

$$\boxed{B_i = \log \left( t_i^{-\frac{1}{2}} \right)}$$

$t_i$ : time since the use of chunk  $i$

$S_{1i}$ : activation spreading from value 1

$S_{2i}$ : activation spreading from value 1

$$S_{1i} = \begin{cases} 0 & \text{if 1 not} \\ & \text{present in } i \\ S - \log(fan_1) & \text{otherwise} \end{cases}$$

$fan_1$ : how often 1 is in memory

## PROPERTIES OF MEMORY – SUMMARY

- retrieval of a chunk from memory dependent on base activation
- base activation: activation of an element due to past experience

## PROPERTIES OF MEMORY – SUMMARY

- retrieval of a chunk from memory dependent on base activation
- base activation: activation of an element due to past experience
- retrieval of a chunk from memory also dependent on spreading activation
- spreading activation: activation of an element due to context in which it is recalled

# PROPERTIES OF MEMORY – SUMMARY

- retrieval of a chunk from memory dependent on base activation
- base activation: activation of an element due to past experience
- retrieval of a chunk from memory also dependent on spreading activation
- spreading activation: activation of an element due to context in which it is recalled

$$(14) \quad \text{Activation: } \boxed{A_i} = B_i + SA_i$$

$$(15) \quad B_i = \log \left( t_i^{-\frac{1}{2}} \right)$$

$$(16) \quad SA_i = S_{1i} + S_{2i} + \dots, \text{ where } S_{ji} = S - \log(fan_j)$$

$$(17) \quad T_i = Fe^{-fA_i} \text{ (time to retrieve a rule)}$$

(18) A boy dances.

$\mathcal{S} = [\langle a, D \rangle, \langle \text{boy}, N \rangle], \mathcal{W} = [\langle \text{dances}, V \rangle]$

① How will we find out what **action** to select?

- Retrieval from memory

Actions

$\langle \text{reduce}, NP \rangle$

$\langle \text{shift} \rangle$

Context

$\langle a, D \rangle, \langle \text{boy}, N \rangle, \langle \text{dances}, V \rangle$

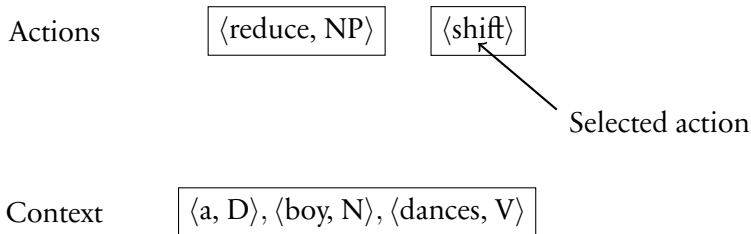
# PARSER – TEST PHASE

(18) A boy dances.

$\mathcal{S} = [\langle a, D \rangle, \langle \text{boy}, N \rangle], \mathcal{W} = [\langle \text{dances}, V \rangle]$

① How will we find out what **action** to select?

- Retrieval from memory



(18) A boy dances.

$\mathcal{S} = [\langle a, D \rangle, \langle \text{boy}, N \rangle], \mathcal{W} = [\langle \text{dances}, V \rangle]$

① How will we find out what **action** to select?

- Retrieval from memory

Actions

$\langle \text{kids}, N \rangle, \langle \text{sleeps}, V \rangle, \text{reduce}, NP$

$\langle \text{shift} \rangle$

Context

$\langle a, D \rangle, \langle \text{boy}, N \rangle, \langle \text{dances}, V \rangle$

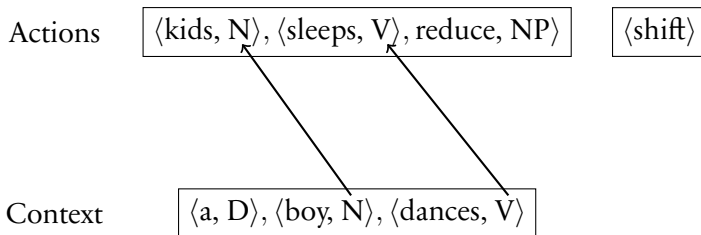
# PARSER – TEST PHASE

(18) A boy dances.

$\mathcal{S} = [\langle a, D \rangle, \langle \text{boy}, N \rangle], \mathcal{W} = [\langle \text{dances}, V \rangle]$

① How will we find out what **action** to select?

- Retrieval from memory





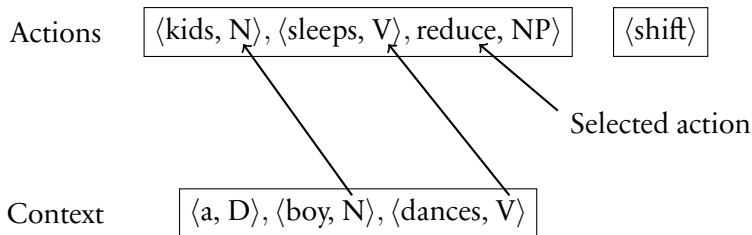
# PARSER – TEST PHASE

(18) A boy dances.

$\mathcal{S} = [\langle a, D \rangle, \langle \text{boy}, N \rangle], \mathcal{W} = [\langle \text{dances}, V \rangle]$

① How will we find out what **action** to select?

- Retrieval from memory



# PARSER'S PROPERTIES

What action should parser take: based on the search in memory and the following context:

- 2 elements in  $\mathcal{W}$  (2 upcoming words and their category)  
(but see later)
- labels of top 4 elements in  $\mathcal{S}$
- head (and its category) of top 4 elements in  $\mathcal{S}$
- left and right children in top 2 elements in  $\mathcal{S}$

## TRAINING AND TESTING THE PARSER

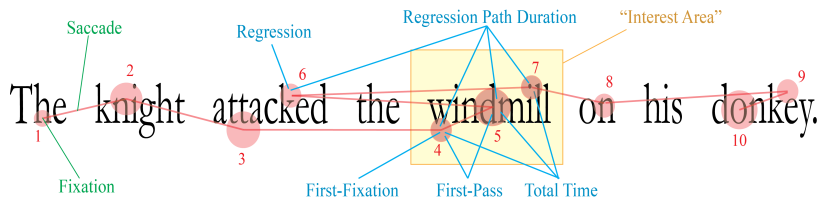
- The parser was trained on Penn TreeBank, sections up to 21.
- Reading times are a by-product of parsing.
- The parser is *serial, incremental, eager*

# NOTES ON THE PARSER

- The parser can be combined with other mechanistic models developed in the same cognitive architecture (reading and eye movements, e.g., E-Z Reader, memory recall...)

# What do eyes do during reading?

- Reading - fixations and saccades



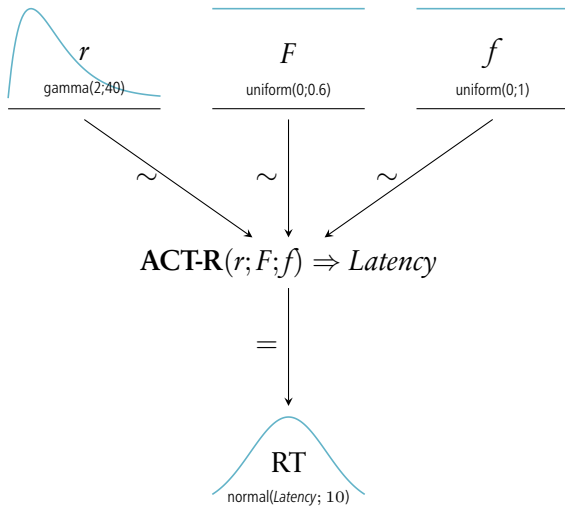
# APPLICATION I: EYE TRACKING (FRANK ET AL., 2013)

- Eye-tracking reader  
Word length as the only parameter modulating visual encoding time
- Lexical retrieval (each word a chunk, its activation calculated based on frequency)

Brasoveanu and Dotlačil, 2019

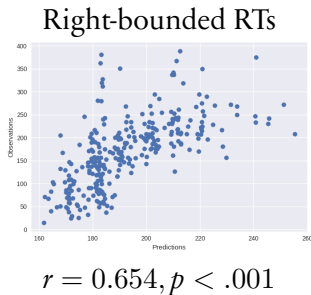
- RT is a function of activation and visual encoding time so our reader predicts RTs
- However, RT is arrived at by fitting free parameters. We need to do that, as well.

# BAYESIAN MODEL STRUCTURE TO TRAIN ACT-R FREE PARAMETERS



# FIT TO WITHHELD DATA

- Eye-tracking reader
- Lexical retrieval based on ACT-R memory retrieval
- Data-driven parser
- RT - function of activation and visual encoding time





# FIT TO WITHHELD DATA

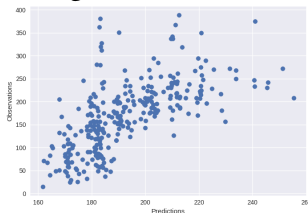
- Eye-tracking reader
- Lexical retrieval based on ACT-R memory retrieval
- Data-driven parser
- RT - function of activation and visual encoding time

$$\log(freq) = -13, p < .001$$

$$length = -1.8, p < .001$$

$$activation\_of\_syn\_rules = -1.7, p = .006$$

Right-bounded RTs



$$r = 0.654, p < .001$$

## APPLICATION II: SELF-PACED READING

Brasoveanu and Dotlačil. An extensible framework for mechanistic processing models: From representational linguistic theories to quantitative model comparison. *Proceedings of 2018 ICCM*.

- Grodner and Gibson (2005, Exp. 1): self-paced reading, matrix subject is modified by a subject or object-extracted relative clause (RC)

(19) The  
reporter who sent the photographer to the editor hoped for  
a story.

(20) The  
reporter who the photographer sent to the editor hoped for  
a story.

9 ROIs: word 2 through word 10 (underlined above)

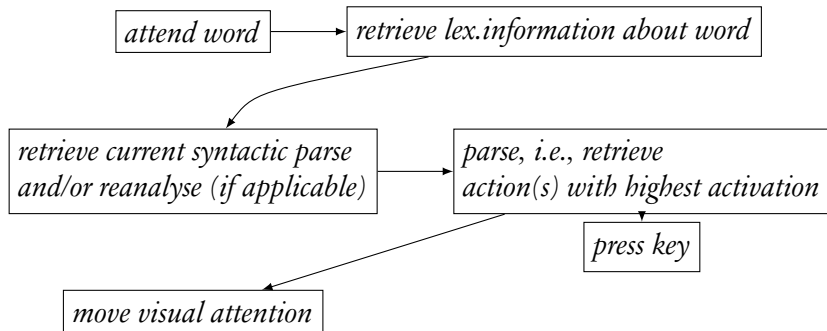
# READER IN ACT-R

Building on Lewis and Vasishth, 2005

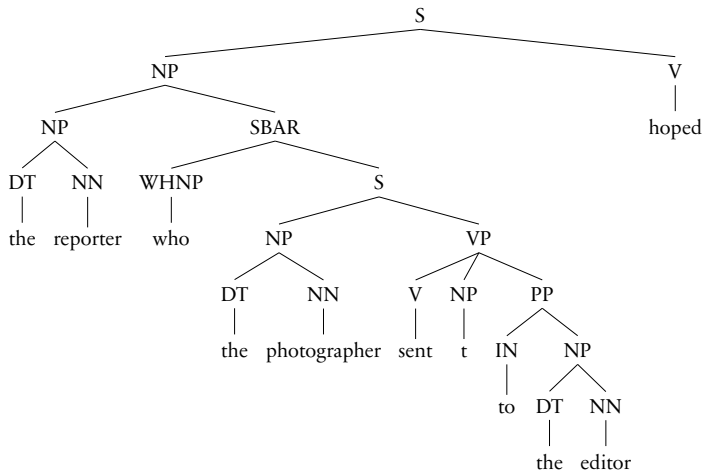
## Reader components:

- Just as Model 1, but one extra component:
- motor module – EPIC Kieras and Meyer, 1996; Meyer and Kieras, 1997
- Just as Model 1:
- In self-paced reading, readers do not see more than the actual word.
- To model this, we consider another parser that has no information about upcoming words ( $\mathcal{W} = \emptyset$ )
- We let it parse; the second parser (with the upcoming information) follows the blind parser and corrects it when needed

# FLOW CHART OF PARSING PROCESS PER WORD

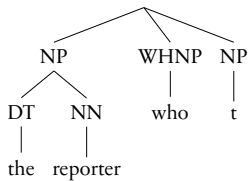


# PARSING – FINAL STRUCTURE



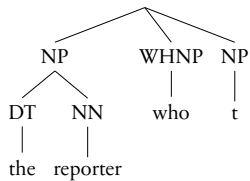
# PARSING – STEPWISE REANALYSIS

Blind parser

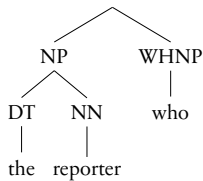


# PARSING – STEPWISE REANALYSIS

Blind parser

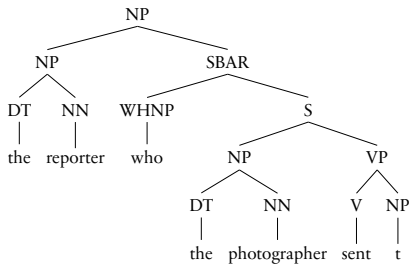


Correcting parser



# PARSING – STEPWISE REANALYSIS

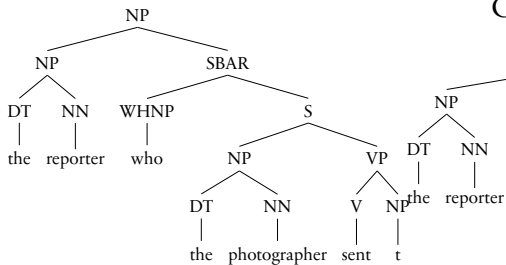
## Blind parser



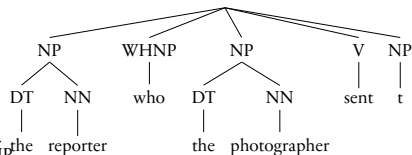


# PARSING – STEPWISE REANALYSIS

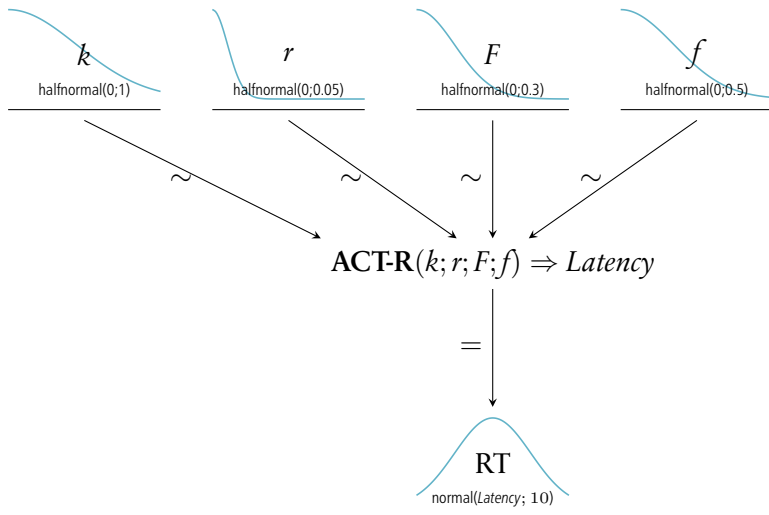
## Blind parser



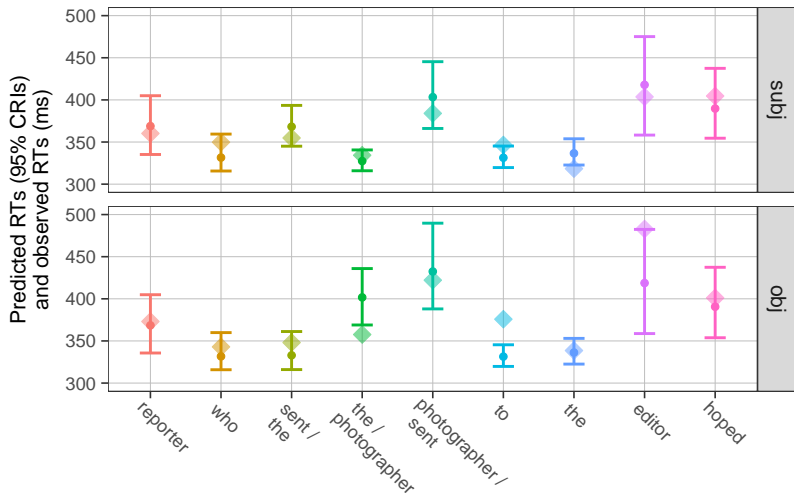
## Correcting parser



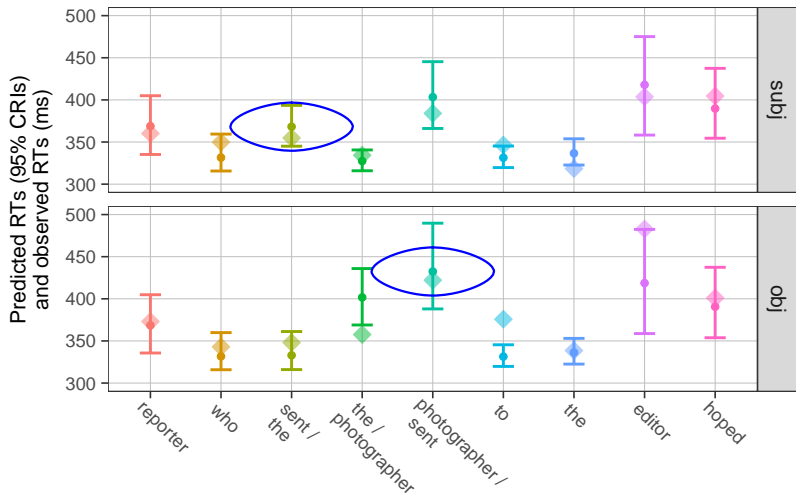
# BAYESIAN MODEL STRUCTURE



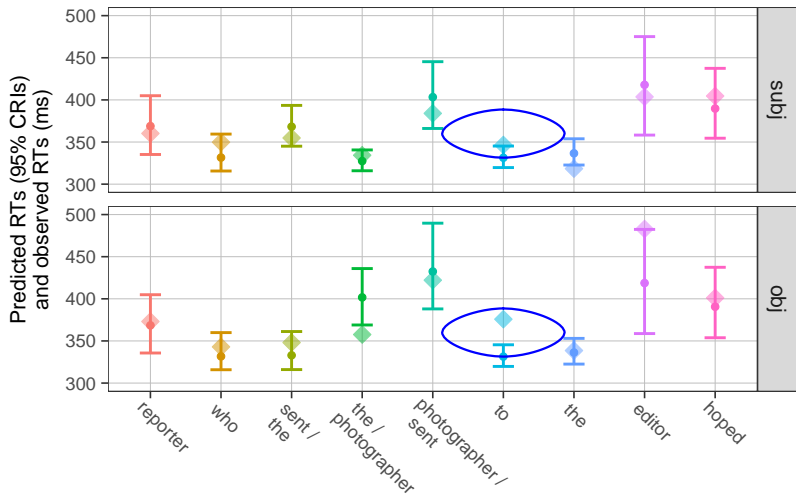
# POSTERIOR PREDICTIONS (MODEL 1)



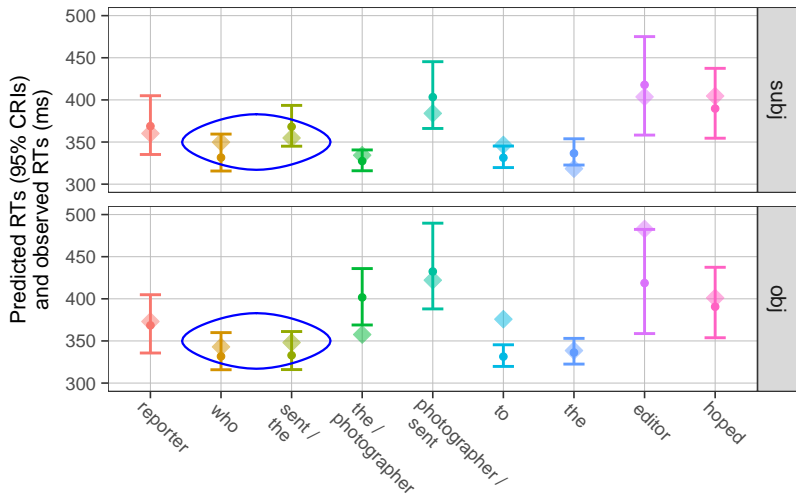
# POSTERIOR PREDICTIONS (MODEL 1)



# POSTERIOR PREDICTIONS (MODEL 1)



# POSTERIOR PREDICTIONS (MODEL 1)

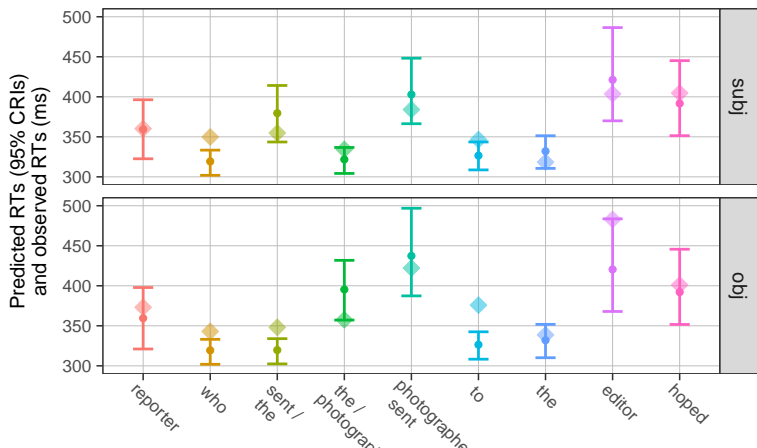


## MODEL 2: NO POSTULATED SUBJECT GAPS

The model does not eagerly predict the interpretation of wh-words as subjects

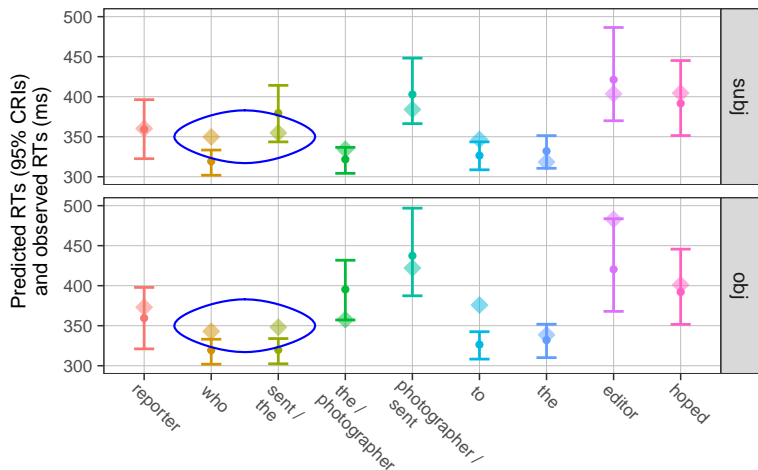
## MODEL 2: NO POSTULATED SUBJECT GAPS

The model does not eagerly predict the interpretation of wh-words as subjects



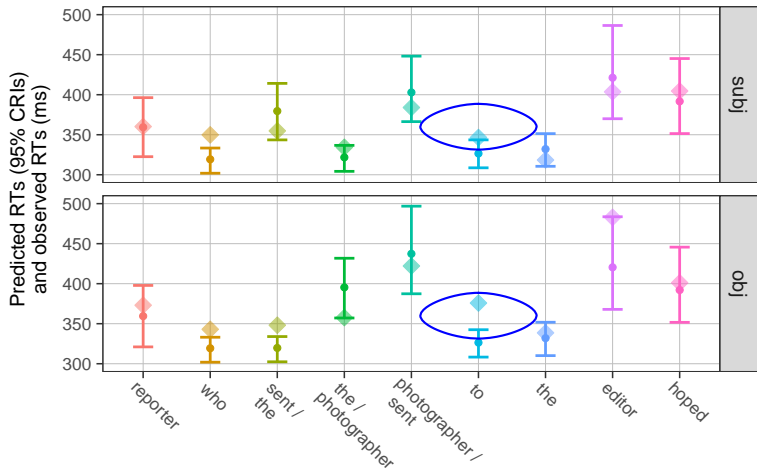


## MODEL 2: NO POSTULATED SUBJECT GAPS

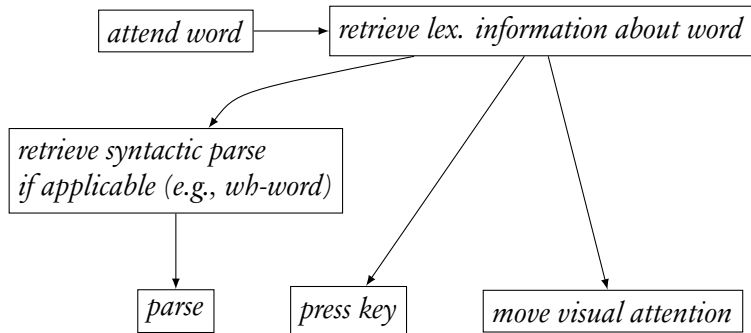


wh-word and following word not modeled well; Model 1 better

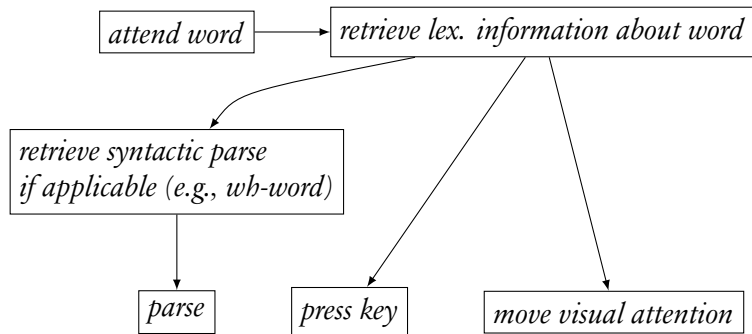
## MODEL 2: NO POSTULATED SUBJECT GAPS



## MODEL 3: PARALLEL READER

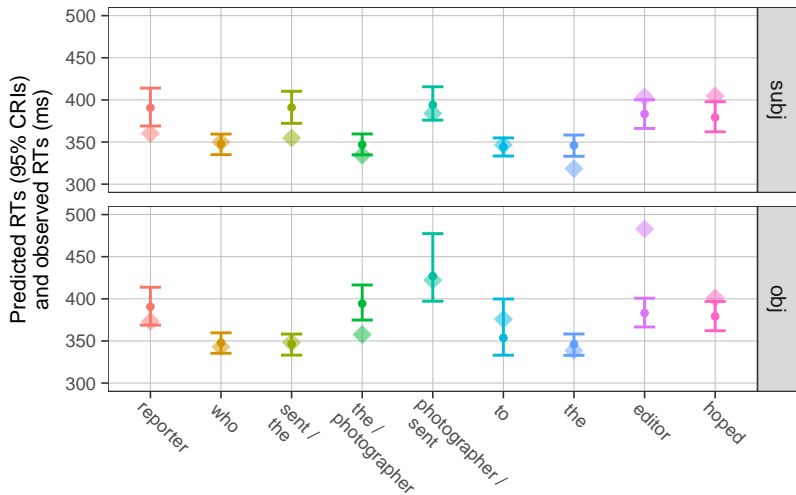


## MODEL 3: PARALLEL READER

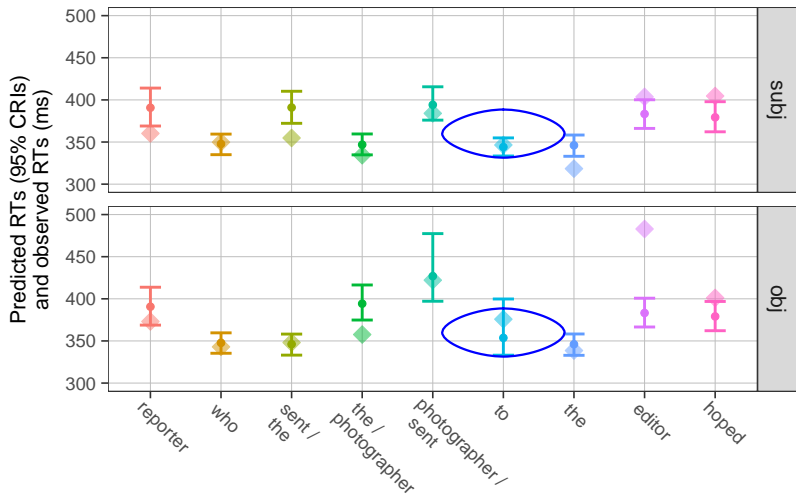


- Model 1 completes all available parsing before key press (serial)
- Model 3: first lexical retrieval, then structure building & key press in parallel
- Outcome: spillover on word after object gap captured

# MODEL 3: SPILLOVER AFTER OBJECT GAP CAPTURED



# MODEL 3: SPILLOVER AFTER OBJECT GAP CAPTURED



# Model comparison

For statistics, see Brasoveanu and Dotlačil, 2018





- Model 1 is better than Model 2
- Model 3 is better than Model 1

# PARSING AND COGNITIVE MODELING




- Abstract models (probabilistic, focus on the ‘why’ question)
- Mechanistic models (algorithmic, focus on the ‘how’ question, but also ‘why’)
- A big advantage of mechanistic models is that they can postulate and test very specific, detailed predictions (cf. 3 models above)
- An advantage of abstract models is an ease of implementation and comprehensiveness







# References I

- 
- Anderson, John R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 
- Anderson, John R. and Christian Lebiere (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 
- Anderson, John R. and Lael J. Schooler (1991). “Reflections of the Environment in Memory”. In: *Psychological Science* 2.6, pp. 396–408.
- 
- Brasoveanu, Adrian and Jakub Dotlačil (2015). “Incremental and predictive interpretation: Experimental evidence and possible accounts”. In: *Proceedings of Semantics and Linguistic Theory (SALT)* 25, pp. 57–81. DOI: 10.3765/salt.v25i0.3047.






# References II

-  Brasoveanu, Adrian and Jakub Dotlačil (2018). “An extensible framework for mechanistic processing models: From representational linguistic theories to quantitative model comparison”. In: *Proceedings of the 2018 International Conference on Cognitive Modelling*.
-  – (2019). “Quantitative Comparison for Generative Theories: Embedding Competence Linguistic Theories in Cognitive Architectures and Bayesian Models”. In: *Proceedings of the 2018 Berkeley Linguistic Society 44*.
-  Demberg, Vera et al. (2013). “Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar”. In: *Computational Linguistics 39.4*, pp. 1025–1066.





# References III

-  Engelmann, Felix et al. (2013). “A Framework for Modeling the Interaction of Syntactic Processing and Eye Movement Control”. In: *Topics in Cognitive Science* 5.3, pp. 452–474. DOI: 10.1111/tops.12026.
-  Engelmann, Felix et al. (2016). “The determinants of retrieval interference in dependency resolution: Review and computational modeling”. Submitted to *Journal of Memory and Language*.
-  Frank, Stefan L et al. (2013). “Reading time data for evaluating broad-coverage models of English sentence processing”. In: *Behavior Research Methods* 45.4, pp. 1182–1190.
-  Grodner, Daniel and Edward Gibson (2005). “Consequences of the Serial Nature of Linguistic Input for Sentential Complexity”. In: *Cognitive Science* 29, pp. 261–291.

# References IV

-  Hale, John (2001). “A Probabilistic Earley Parser as a Psycholinguistic Model”. In: *Proceedings of the 2nd Meeting of the North American Association for Computational Linguistics*, pp. 159–166.
-  Hale, John T. (2014). *Automaton Theories of Human Sentence Comprehension*. Stanford: CSLI Publications.
-  Kieras, David E and David E Meyer (1996). “The EPIC architecture: Principles of operation”. Unpublished manuscript from <ftp://ftp.eecs.umich.edu/people/kieras/EPICarch.ps>.
-  Lewis, Richard and Shravan Vasishth (2005). “An activation-based model of sentence processing as skilled memory retrieval”. In: *Cognitive Science* 29, pp. 1–45.
-  Meyer, David E and David E Kieras (1997). “A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms.” In: *Psychological review* 104.1, p. 3.

# References V

-  Nicenboim, Bruno and Shravan Vasishth (2018). “Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling” In: *Journal of Memory and Language* 99, pp. 1–34. DOI: <https://doi.org/10.1016/j.jml.2017.08.004>.
-  Nivre, Joakim (2004). “Incrementality in deterministic dependency parsing” In: *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*. Association for Computational Linguistics, pp. 50–57.
-  Rij, Jacolien van (2012). *Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults*. Groningen.
-  Taatgen, Niels A. and John R. Anderson (2002). “Why do children learn to say “broke”? A model of learning the past tense without feedback” In: *Cognition* 86.2, pp. 123–155.

# References VI



Vasishth, Shravan et al. (2008). “Processing Polarity: How the Ungrammatical Intrudes on the Grammatical”. In: *Cognitive Science* 32, pp. 685–712.



Vogelzang, Margreet et al. (2017). “Toward Cognitively Constrained Models of Language Processing: A Review”. In: *Frontiers in Communication* 2, pp. 1–11. DOI: 10.3389/fcomm.2017.00011.



Young, Richard M. and Richard L. Lewis (1999). “The Soar cognitive architecture and human working memory”. In: ed. by Akira Miyake and Priti Shah, pp. 224–256.