

Surprising Mechanisms

Marten van Schijndel
Department of Linguistics, Cornell University
January 20, 2023

Background

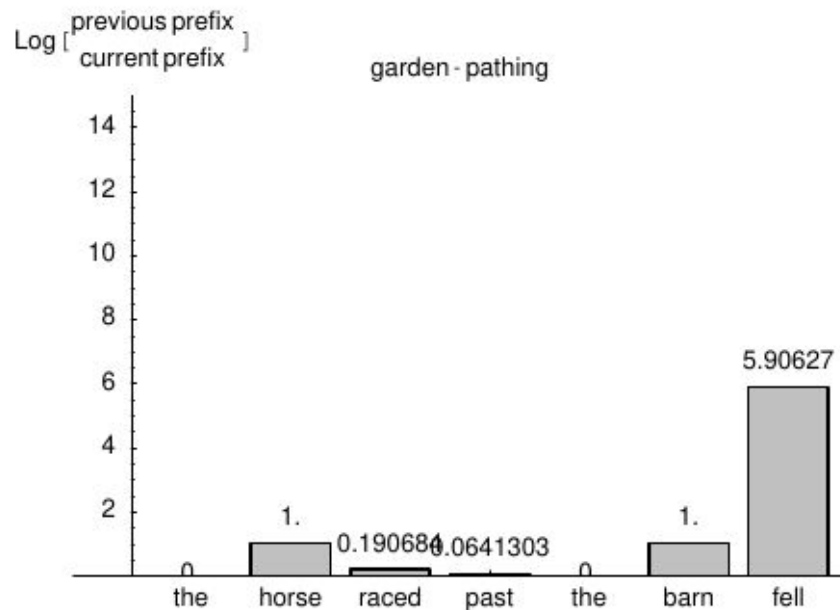
$$\text{Surprisal}(w_t) = -\log P(w_t \mid w_1 \dots w_{t-1})$$

Surprisal reflects the contextual (im)probability of an event

Terminology: Surprisal = information content = information load = (un)predictability

Surprisal predicts linguistic disambiguation

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	VCN	→	<i>raced</i>
1.0	IN	→	<i>past</i>



Surprisal predicts human behavior

Speake

Word surprisal predicts N400 amplitude during reading

Stefan L. Frank^{1,2}

Leun J. Otten³

Giulia Galli³

Gabriella Vigliocco²



ELSEVIER

Synta

Using surprisal and fMRI to map the neural bases of broad and local contextual prediction during natural language comprehension

Shohini Bhattasali and Philip Resnik

Linguistics/UMIACS

University of Maryland

College Park, MD

{shohini, resnik}@umd.edu

Data fr
theorie

Vera Demberg

{vera, asayeed, philipg, nikolaos}@coli.uni-saarland.de

But *how* does surprisal influence behavior?



Tal Linzen

COGNITIVE SCIENCE

A Multidisciplinary Journal




Cognitive Science 45 (2021) e12988

© 2021 Cognitive Science Society LLC

ISSN: 1551-6709 online

DOI: 10.1111/cogs.12988

Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty

Marten van Schijndel, PhD,^a  Tal Linzen, PhD^b

^a*Department of Linguistics, Cornell University*

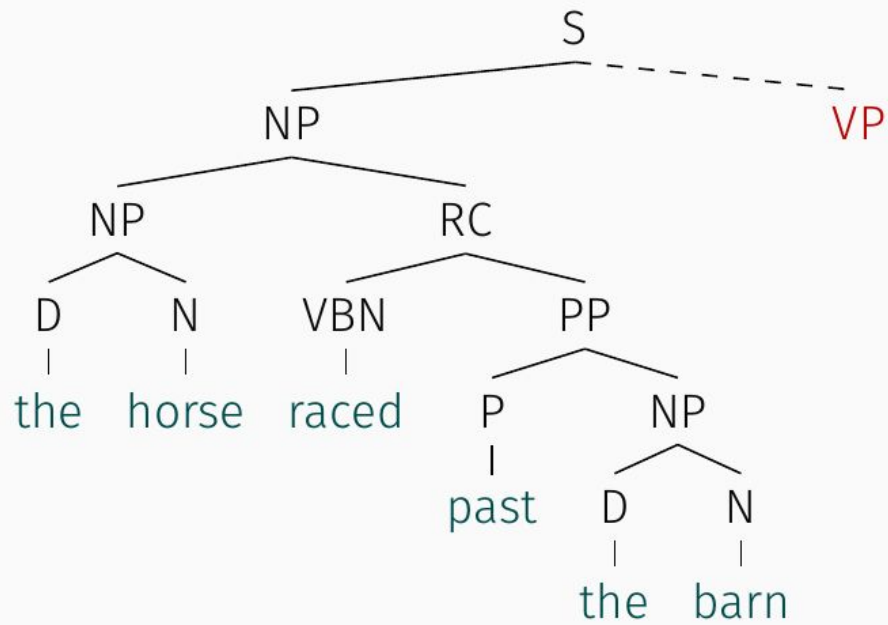
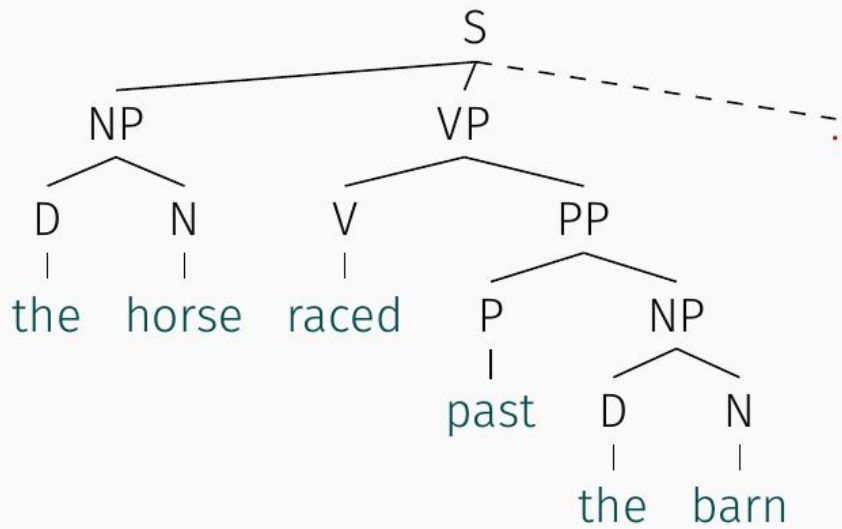
^b*Department of Linguistics and Center for Data Science, New York University*

The horse raced past the barn fell

Bever, 1970, *Cognition and the Development of Language*

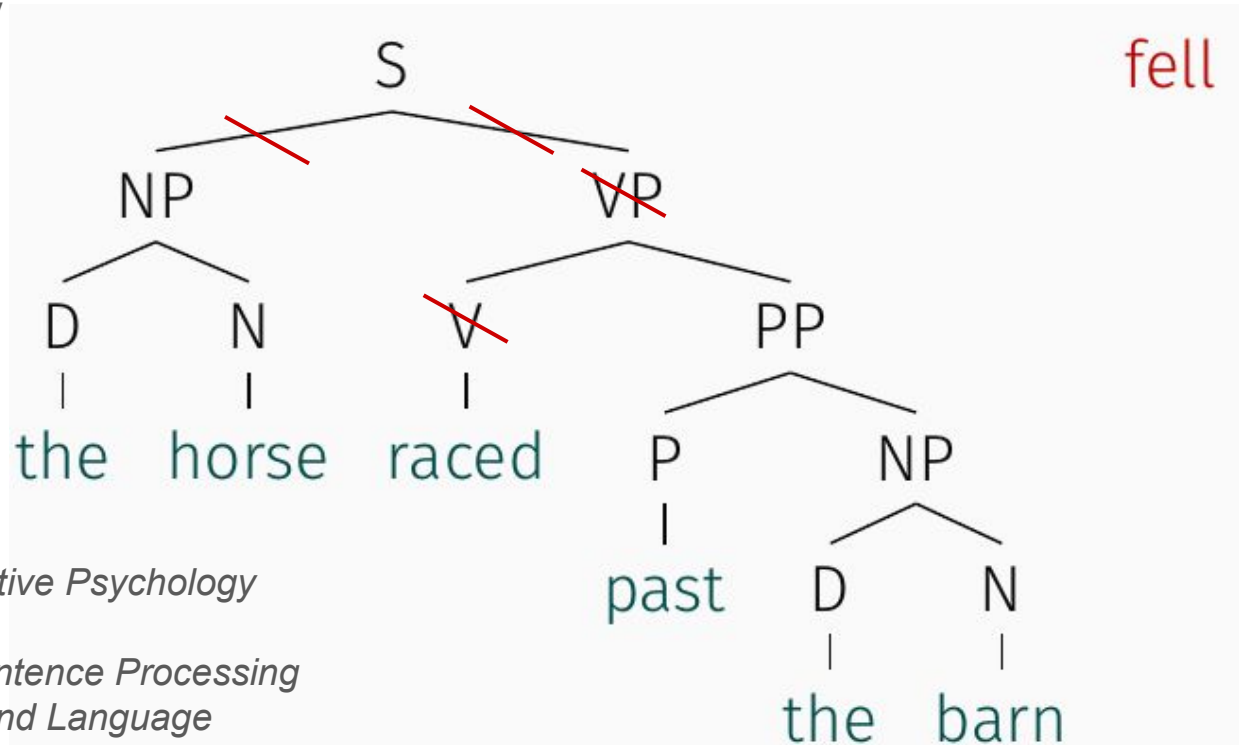
The horse which was raced past the barn fell

Bever, 1970, *Cognition and the Development of Language*



Theorized disambiguation mechanisms

H1: Serial tree surgery



Frazier & Rayner, 1982, *Cognitive Psychology*

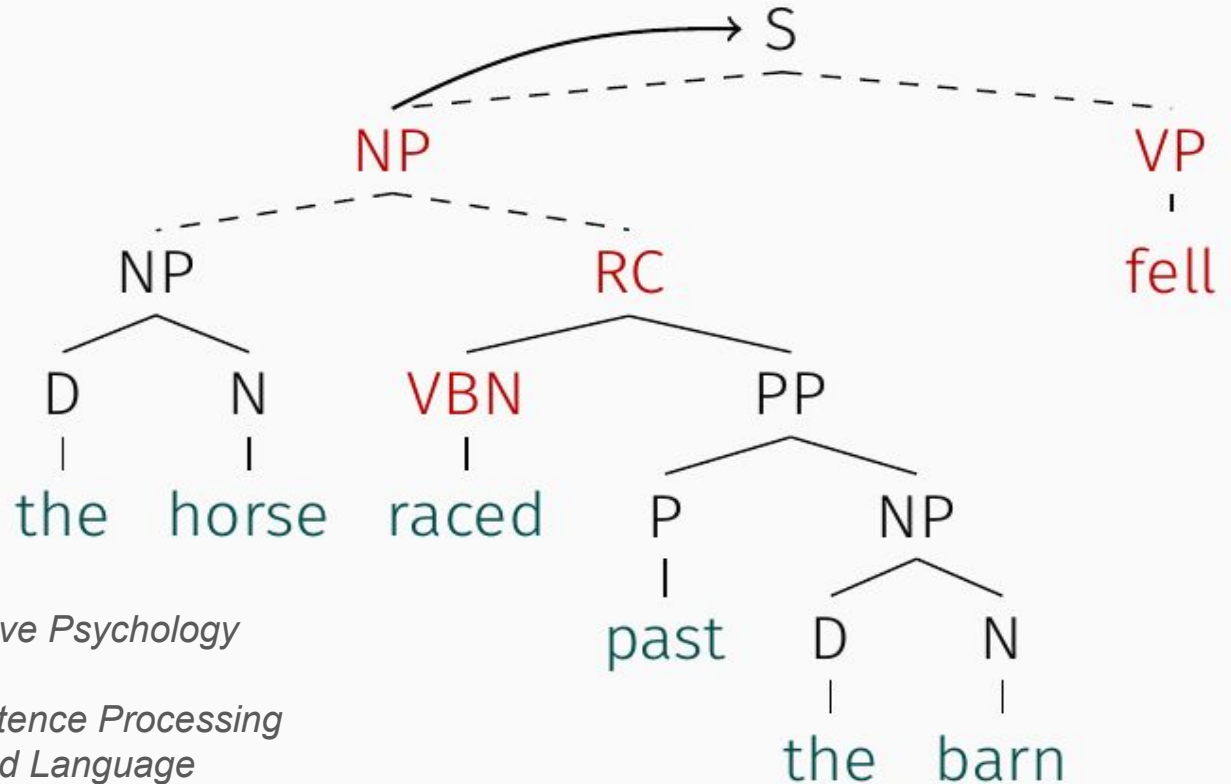
Pritchett, 1988, *Language*

Lewis, 1998, *Reanalysis in Sentence Processing*

Sturt et al., 1999, *J. Memory and Language*

Theorized disambiguation mechanisms

H1: Serial tree surgery



Frazier & Rayner, 1982, *Cognitive Psychology*

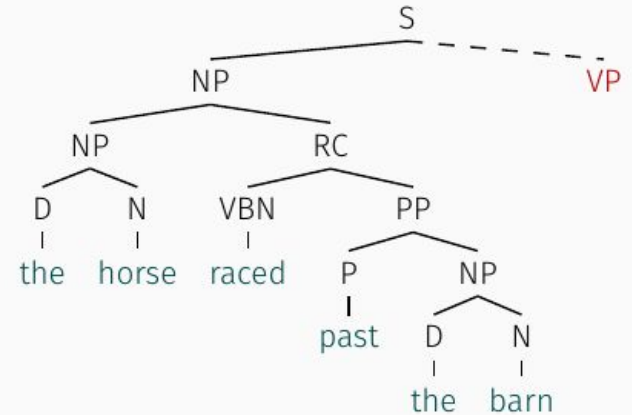
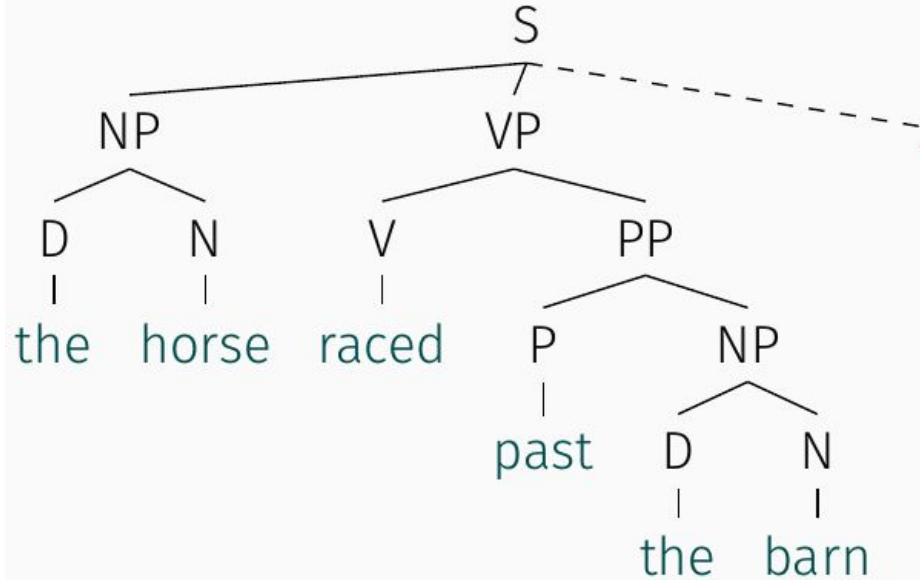
Pritchett, 1988, *Language*

Lewis, 1998, *Reanalysis in Sentence Processing*

Sturt et al., 1999, *J. Memory and Language*

Theorized disambiguation mechanisms

H2: Parallel reranking



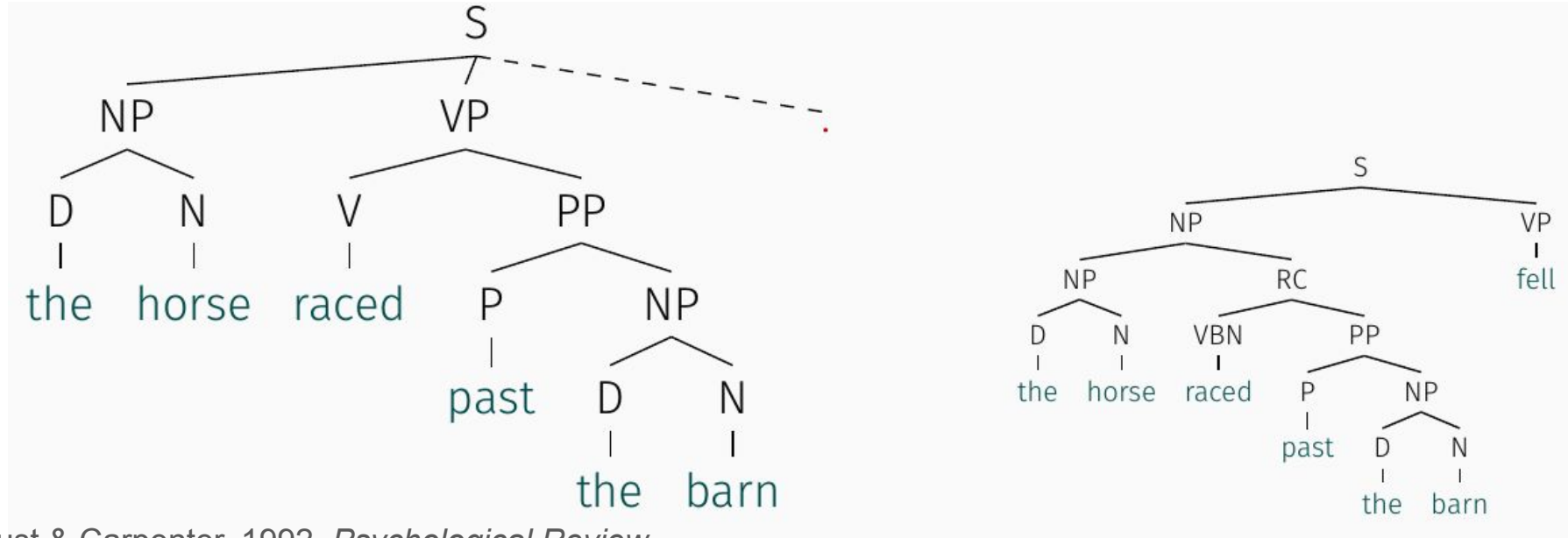
Just & Carpenter, 1992, *Psychological Review*

Hale, 2001, *NAACL*

Levy, 2013, *Sentence Processing*

Theorized disambiguation mechanisms

H2: Parallel reranking



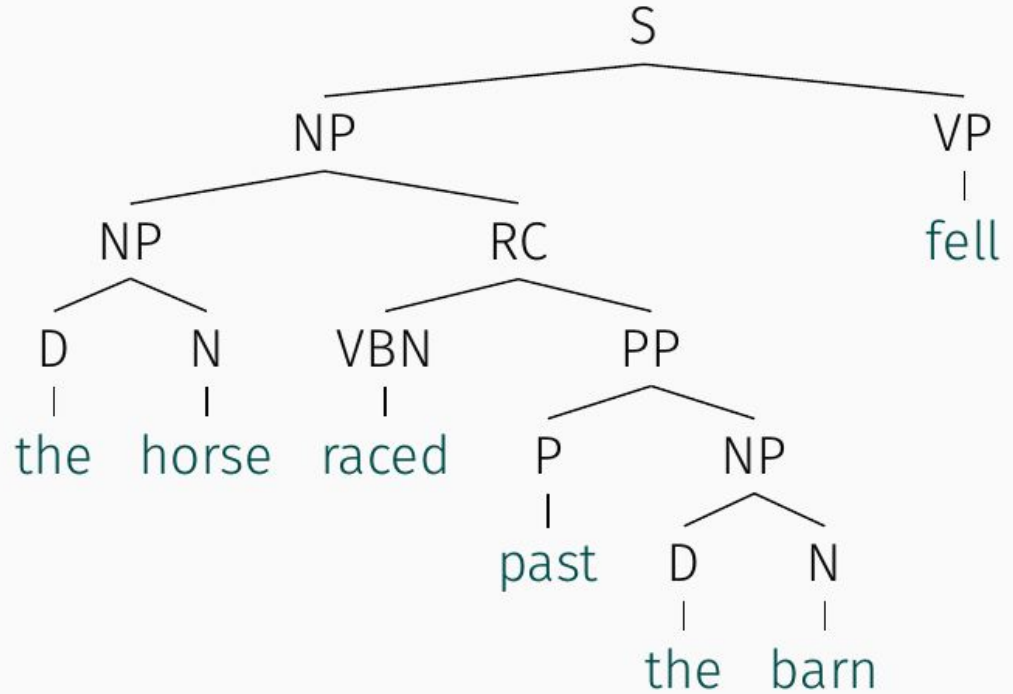
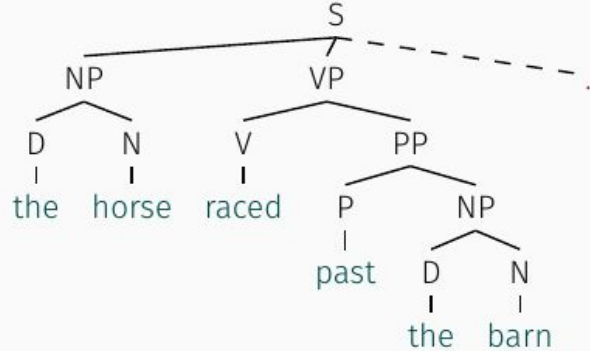
Just & Carpenter, 1992, *Psychological Review*

Hale, 2001, *NAACL*

Levy, 2013, *Sentence Processing*

Theorized disambiguation mechanisms

H2: Parallel reranking



Just & Carpenter, 1992, *Psychological Review*

Hale, 2001, *NAACL*

Levy, 2013, *Sentence Processing*

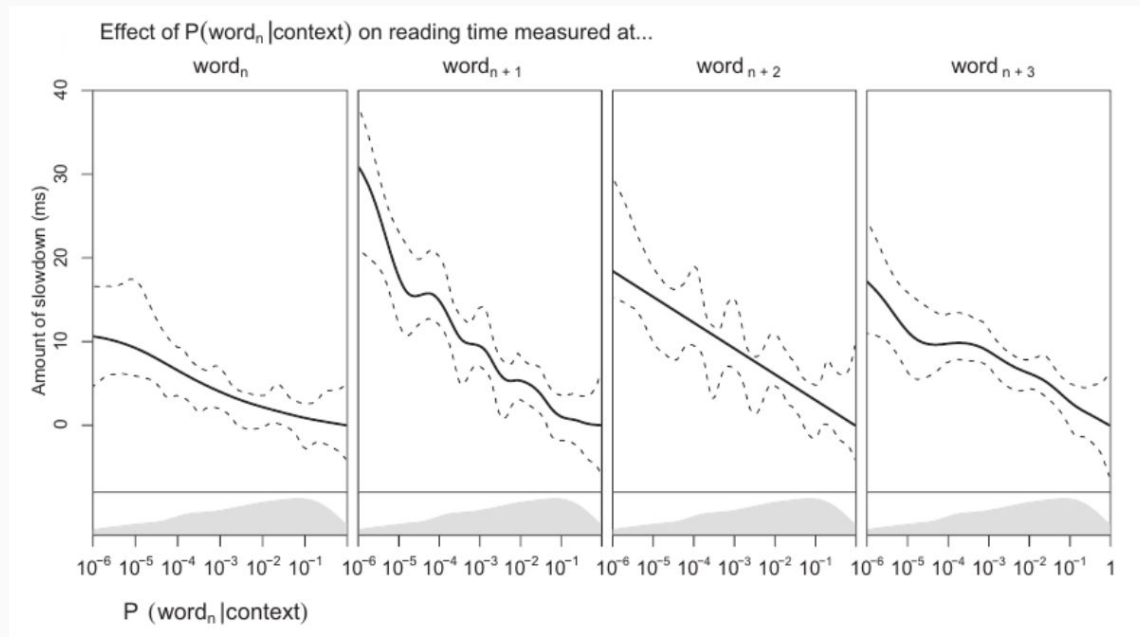
NNs can predict garden path *existence*

van Schijndel & Linzen, 2018, *Proc CogSci*
Futrell et al., 2019, *Proc NAACL*
Frank & Hoeks, 2019, *Proc CogSci*
Davis & van Schijndel, 2020, *Proc CogSci*

NNs can predict garden path *existence*

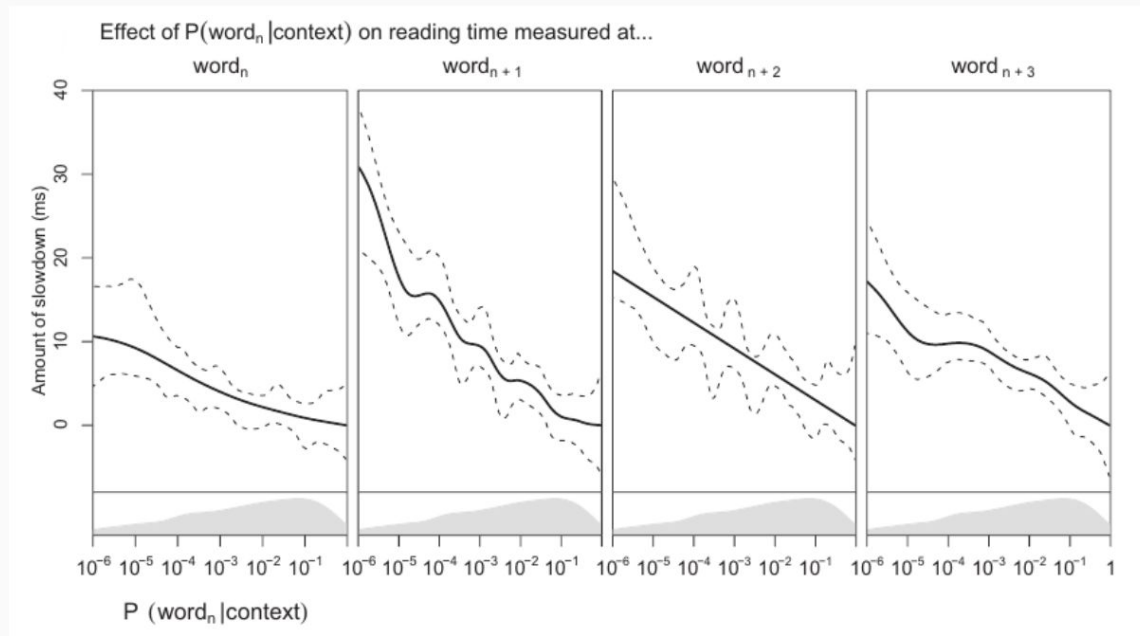
Look beyond garden path *existence* to garden path *magnitude*

Surprisal is linearly related to reading times!



Smith and Levy, 2013, *Cognition*

Surprisal is linearly related to reading times!



$$RT(w_i) = \delta_0 S(w_i) + \delta_{-1} S(w_{i-1}) + \delta_{-2} S(w_{i-2}) + \delta_{-3} S(w_{i-3})$$

Smith and Levy, 2013, *Cognition*

WikiRNN:

Gulordava et al. (2018) LSTM

Data: Wikipedia (80M words)

SoapRNN:

2-layer LSTM (Same parameters as WikiRNN)

Data: Corpus of American Soap Operas (80M words; Davies, 2011)

Mapping probs to reading times

Reading Time Data (SPR; Prasad and Linzen, 2019)

- 80 simple sentences (fillers)
- 224 participants
- 1000 words / participant

Linear Mixed Regression

$$\text{time} \sim \text{text position} + \text{word length} \times \text{frequency} + \dots + \text{predictability}_t$$

Smith & Levy, 2013:

$$\delta_0 = 0.53 \quad \delta_{-1} = 1.53 \quad \delta_{-2} = 0.92 \quad \delta_{-3} = 0.84$$

WikiRNN using Prasad & Linzen, 2019:

$$(\delta_0 = 0.04) \quad \delta_{-1} = 1.10 \quad \delta_{-2} = 0.37 \quad \delta_{-3} = 0.39$$

SoapRNN using Prasad & Linzen, 2019:

$$(\delta_0 = -0.04) \quad \delta_{-1} = 0.83 \quad \delta_{-2} = 0.91 \quad \delta_{-3} = 0.44$$

Three Garden Paths

NP/S: The woman saw { the doctor wore a hat.
that the doctor wore a hat.

Three Garden Paths

NP/S: The woman saw { the doctor wore a hat.
that the doctor wore a hat.

NP/Z: When the woman { visited her nephew laughed loudly.
visited, her nephew laughed loudly.

MV/RR: The horse { raced past the barn fell.
which was raced past the barn fell.

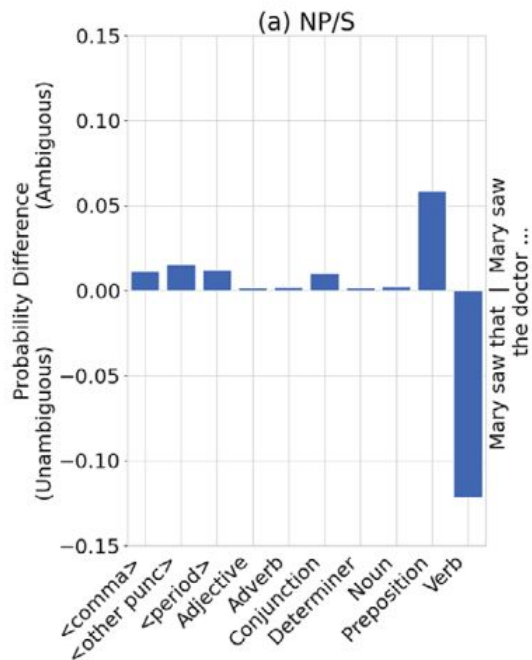
The horse raced past the barn fell

The horse *which was* raced past the barn fell

Bever, 1970, *Cognition and the Development of Language*

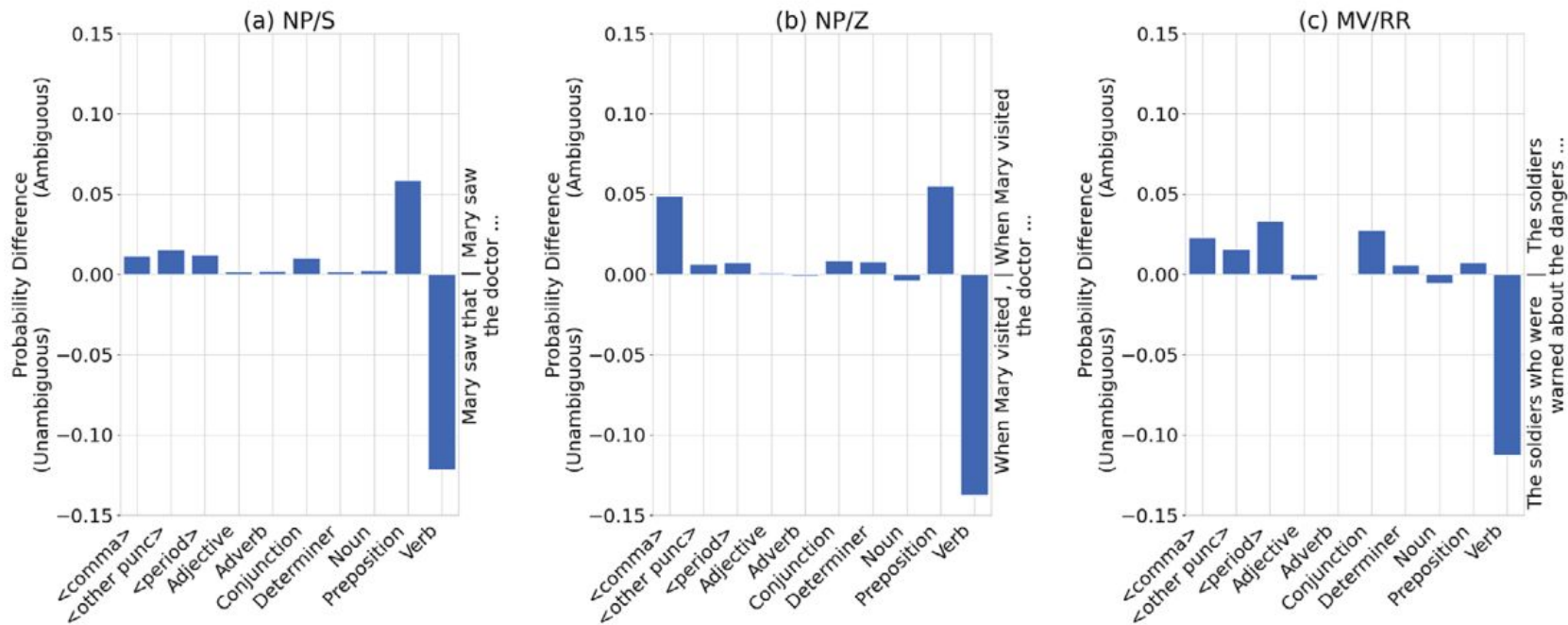
NNs have human-like garden path interpretations

RNN garden path part-of-speech predictions



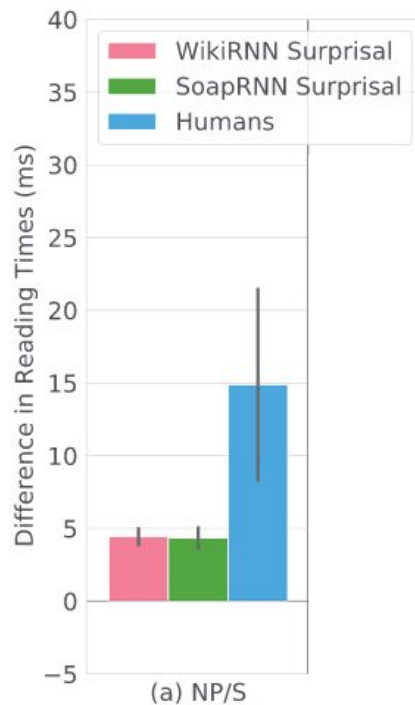
NNs have human-like garden path interpretations

RNN garden path part-of-speech predictions



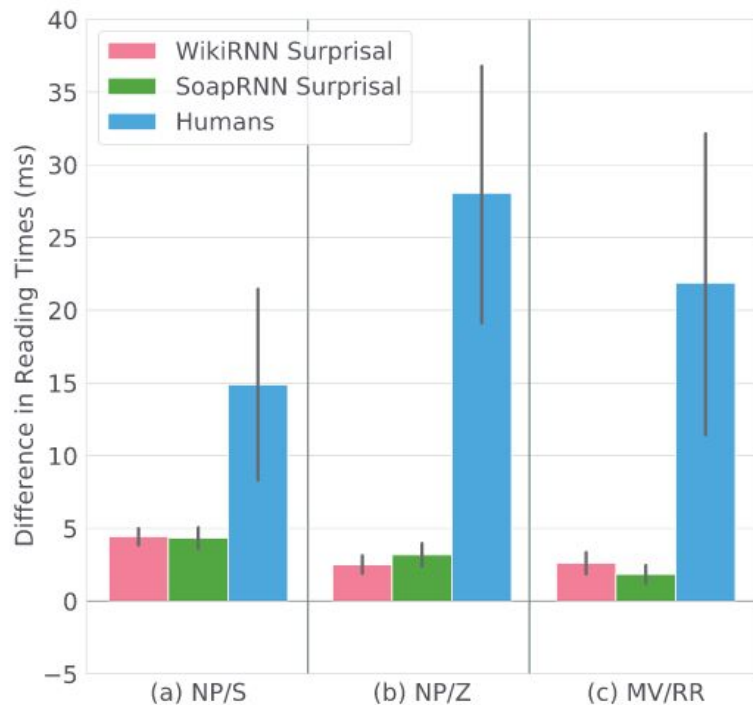
Surprisal is unable to predict effect magnitude

Predicted/empirical mean garden path effects



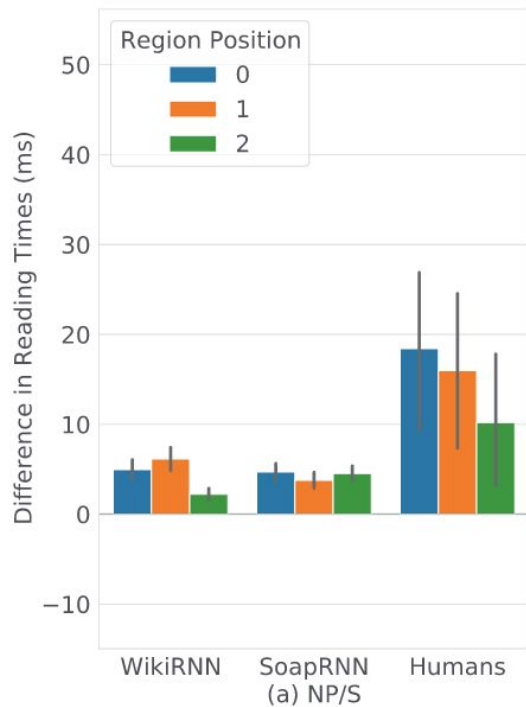
Surprisal is unable to predict effect **magnitude**

Predicted/empirical mean garden path effects



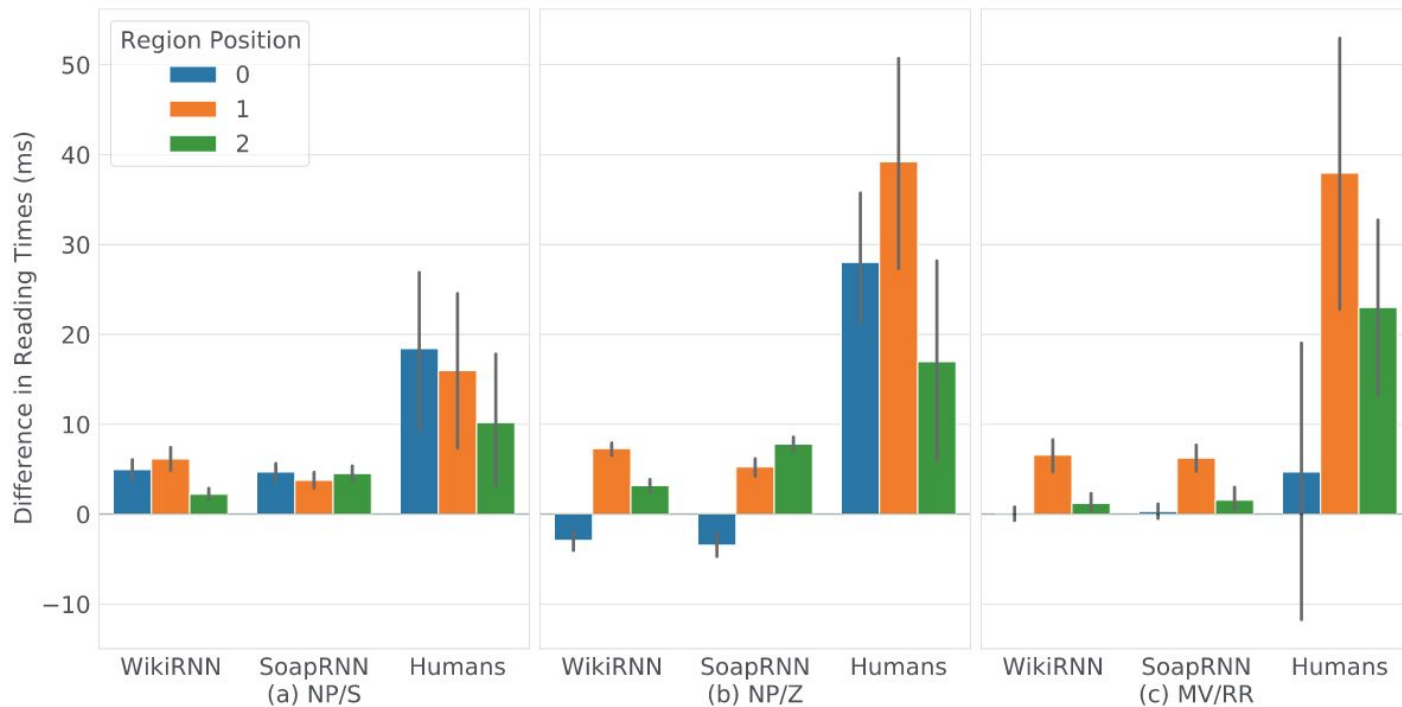
Each construction produces different behavior

Predicted/empirical word-by-word garden path effects



Each construction produces different behavior

Predicted/empirical word-by-word garden path effects



Paper Conclusions

- Neural networks capture expected garden path interpretations
- Conversion rates are fairly **similar**, but all **underestimate** human responses
- Different garden paths exhibit different timecourses
- Suggests human responses influenced by factors **beyond predictability**



Deb Bhattacharya

Code-switching in online posts reveals evidence for audience design

Debasmita Bhattacharya

Department of Computer Science

Columbia University

db3526@columbia.edu

Marten van Schijndel

Department of Linguistics

Cornell University

mv443@cornell.edu

Under Review

What is code-switching

Matrix Language Embedded Language

暑期 短租 还 available 哦。
summer short-rental still available *excl.*
The summer rental is still available.

Surprisal influences code-switching

Surprising continuations are more likely to be code-switched

Why would this be?

Hypothesized mechanisms for surprisal influence

H1: Speaker-driven code-switching



Hypothesized mechanisms for surprisal influence

H2: Audience-driven code-switching



Hypothesized mechanisms for surprisal influence

H1: Speaker-driven code-switching

Prediction: English surprisal $<$ Chinese surprisal

H2: Audience-driven code-switching

Prediction: Chinese surprisal \leq English surprisal

Code-switching data

暑期 短租 还 available 哦。
summer short-rental still available *excl.*
The summer rental is still available.

Code-switching data

(1) CS: 暑期 短租 还 available 哦。
summer short-rental still available *excl.*
The summer rental is still available.

Key:

CS-1

(2) CS: 暑期 短租 还 有空的 哦。
summer short-rental still available *excl.*
The summer rental is still available.

(3) Non-CS: 附近 有 很多 餐馆。
nearby has many restaurant.
There are many restaurants nearby.

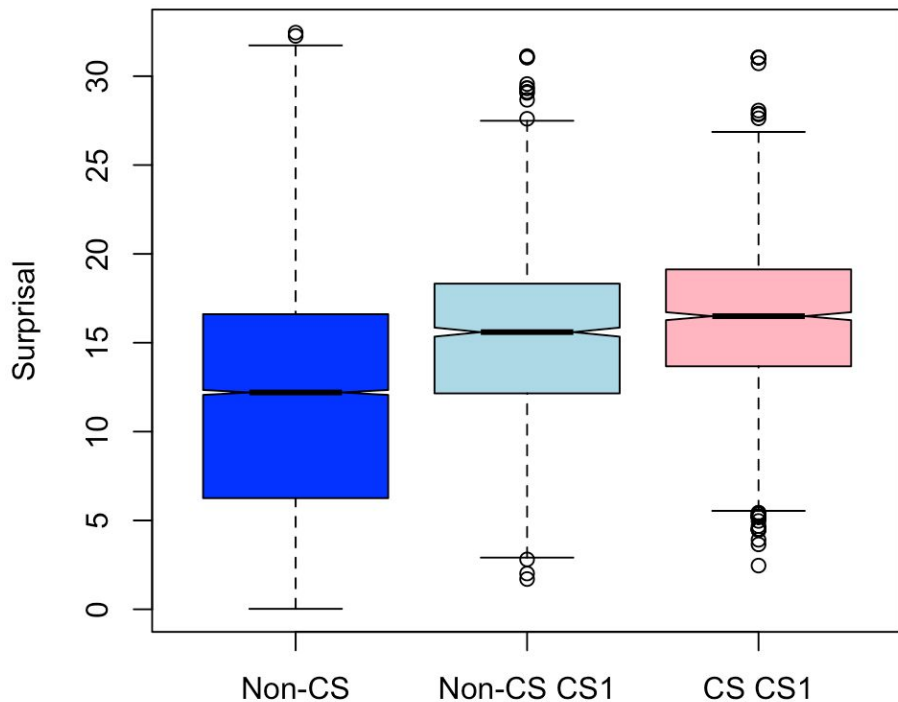
Key:

CS-1

Non-CS

Replication: Surprisal is correlated with code-switching

Unigram surprisal (frequency), 5-gram surprisal: Chinese Wikipedia (35 million tokens)



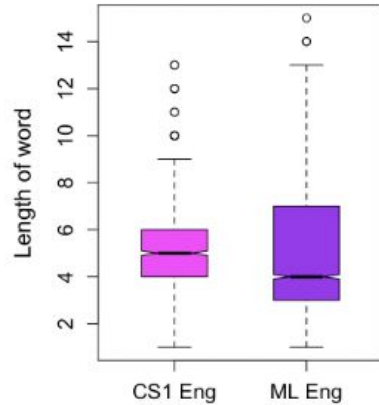
Factor	coef	std err	t
Intercept	0.5223	0.006	94.419
POS=verb	0.0048	0.010	0.481
POS=other	-0.0609	0.009	-6.852
Frequency	0.8935	0.007	119.696
Word length	-0.0431	0.008	-5.716
Sentence length	0.0460	0.007	6.660
Surprisal	0.0605	0.008	7.251

Table 1: Summary of the logistic regression model for CS1 (coded 1) versus random Non-CS1 (coded 0).

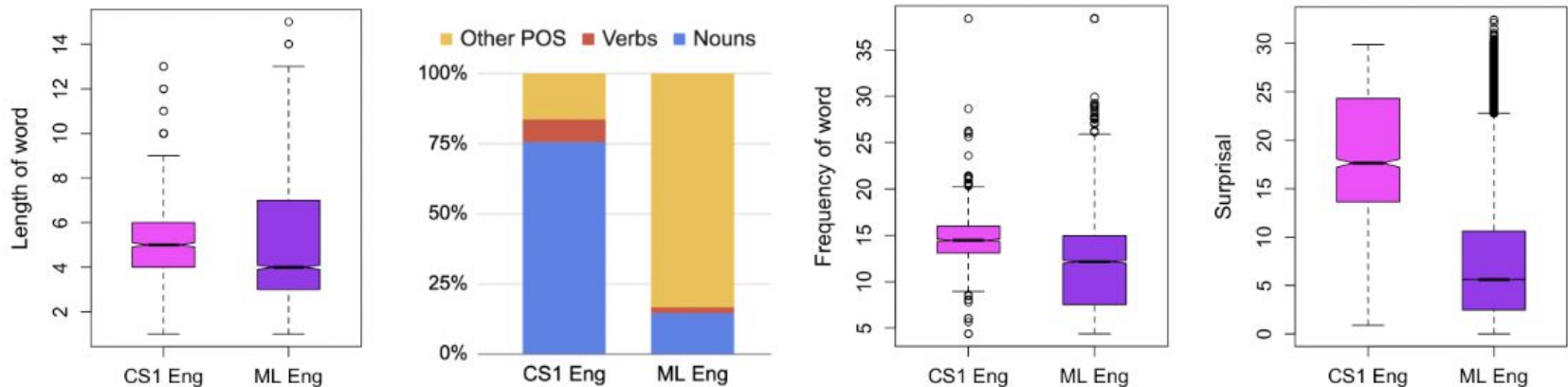
Corpus expansion

We generated English machine translations	(1) CS:	暑期 短租 还 <u>available</u> 哦。 summer short-rental still <u>available</u> <i>excl.</i> <i>The summer rental is still available.</i>	Key: <u>CS-1</u>
	(2) CS:	暑期 短租 还 <u>有空的</u> 哦。 summer short-rental still <u>available</u> <i>excl.</i> <i>The summer rental is still available.</i>	
	(3) Non-CS:	附近 <u>有</u> <u>很多</u> <u>餐馆</u> 。 nearby has many restaurant. <i>There are many restaurants nearby.</i>	Key: <u>CS-1</u> <u>Non-CS</u>

CS1 English is more complex than monolingual English

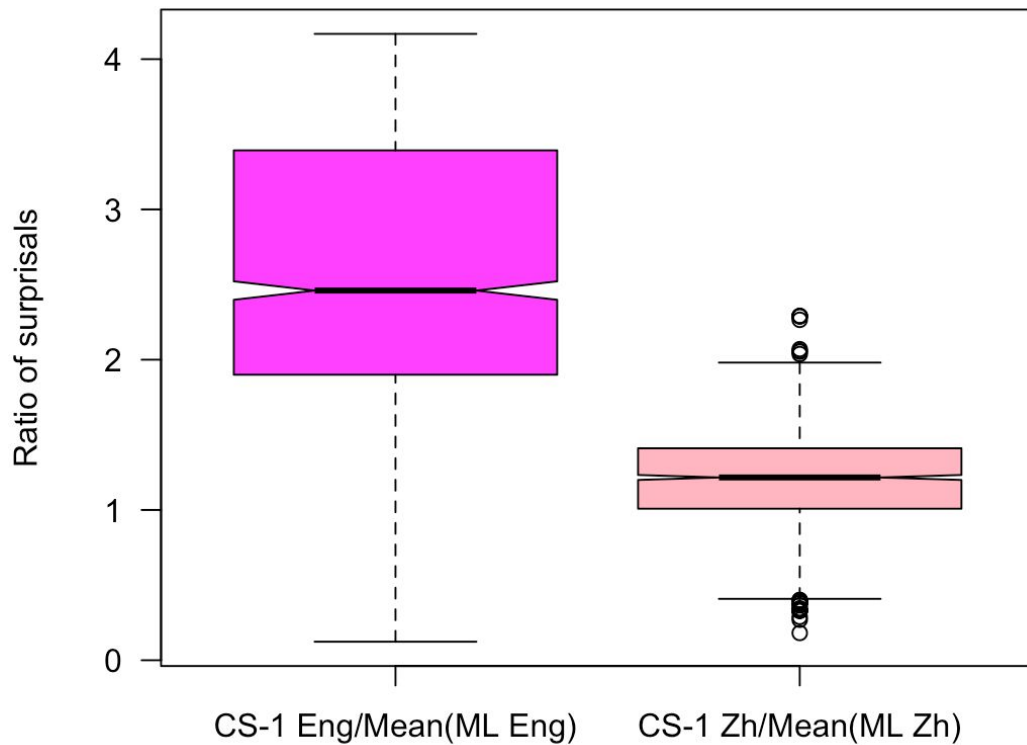


CS1 English is more complex than monolingual English



What is the **relative complexity** of CS1 English compared with CS1 Chinese?

CS1 English is more complex than CS1 Chinese



Paper Conclusions

Surprisal has an **audience-driven** influence on code-switching

- Code-switching is correlated with high surprisal, but code-switches tend to be **more complex** than monolingual speech
- Suggests speakers use code-switching to **signal complexity** for listeners, rather than necessarily finding it more salient for themselves

Talk conclusions

- Surprisal underestimates human behavioral responses
- There are additional repair mechanisms beyond re-ranking
- At areas of high surprisal, code-switching is used to signal to the audience about the area of high complexity

Thanks!



Tal Linzen



Deb Bhattacharya



C.Psyd



Cornell NLP