Language is not language processing

Marten van Schijndel

May 2020

Department of Linguistics, Cornell University

CL/NLP often aim to create models of language comprehension
(NLI, parsing, information extraction, etc)

Often, language models are trained on large amounts of text
And these are the starting point for more complex models
Or they are used for cognitive modeling

Model biases may not align with human comprehension biases

$\rightarrow$ Models may not learn human comprehension during training

All language data comes from production not comprehension
(though annotations provide comprehension cues)

$\rightarrow$ Comprehension signal may not be present in the produced data

In this talk, I explore these two possible problems with our current modeling paradigm

Overview

Part 0: Background

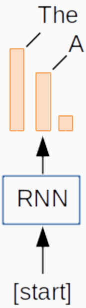
Part 1: Magnitude probing

Part 2: World knowledge probing

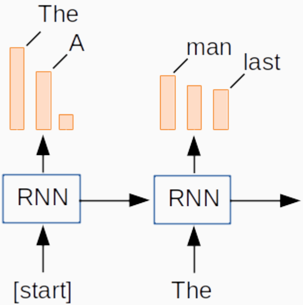
Part 3: Production / comprehension mismatch

Neural networks have proven especially successful at finding *linguistically accurate* language processing solutions.

NNs are often trained on a word prediction task



NNs are often trained on a word prediction task



Why word prediction?

We can measure how unexpected a word is with **surprisal**

$$\text{Surprisal}(w_i) = -\log \text{P}(w_i \mid w_{1..i-1}) \tag{1}$$

Shannon, 1948, *Bell Systems Technical Journal*
Hale, 2001, *Proc. North American Assoc. Comp. Ling.*
Levy, 2008, *Cognition*

Why word prediction?

Surprisal indicates what the model finds unexpected/unnatural which can then be mapped onto human behavioral and neural measurements

- acceptability/grammaticality
- reading/reaction times
- neural activation

So. Many. People.

This is kind of crazy!

We know frequency/predictability affect human language processing

However, many plausible explanations of human responses involve experience beyond language statistics

E.g., can language models learn intention from text alone?
There may be some weak signal, but …

Part 1: Magnitude probing

van Schijndel & Linzen, 2018, *Proc. CogSci*
van Schijndel & Linzen, in prep

Humans experience a visceral response upon encountering garden path constructions

NNs model average stats and therefore average frequency responses.

Garden path *responses* exist in the tail.

They exist in the tail because:

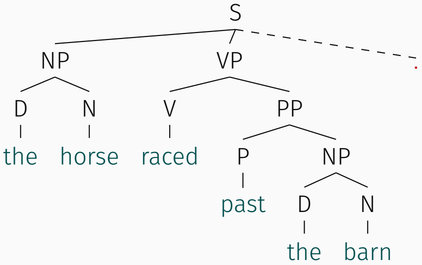
1. the statistics are in the tail (predictability)
   OR
2. the response is unusual (reanalysis)

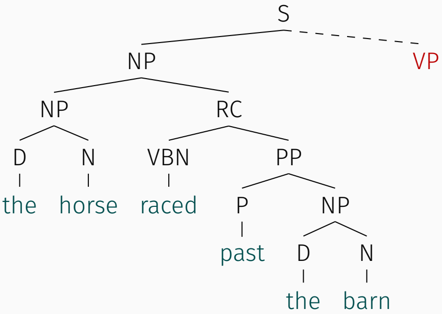
The horse raced past the barn fell .

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

The horse that was raced past the barn fell .

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





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

While human responses are framed in terms of explicit syntactic frequencies,
RNNs can predict garden path responses without explicit syntactic training.

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

Do RNNs process garden paths similar to humans?

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

Models

WikiRNN:
Gulordava et al., (2018) LSTM
Data: Wikipedia (80M words)

SoapRNN:
2-layer LSTM (Same training parameters as above)
Data: Corpus of American Soap Operas (80M words; Davies, 2011)

Three garden paths

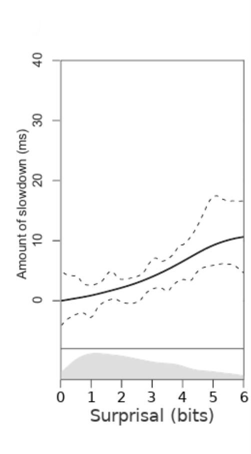
NP/S: The woman saw { the doctor wore a hat. / that the doctor wore a hat. }

- NP/S: The woman saw the doctor wore a hat.
- NP/S: The woman saw 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. }

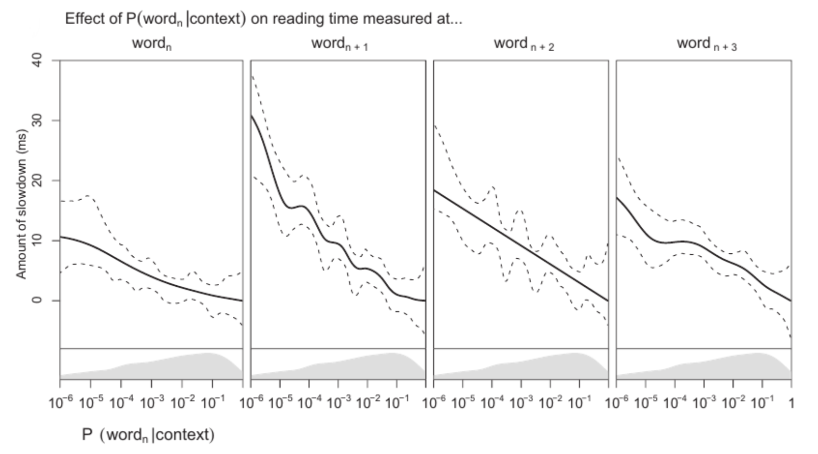
Surprisal-to-ms conversion



$$\text{RT}(w_t) = \alpha \text{S}(w_t) \tag{2}$$

Smith & Levy, 2013, *Cognition*

Probability-to-ms Conversion



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

Smith & Levy, 2013, *Cognition*

Probabilities

- Kneser-Ney trigram probabilities
- Estimated from British National Corpus (100M words)

Reading Time Data (SPR; ignoring ET)

- Brown corpus
- 35 participants
- 5000 words / participant

Generalized Additive Mixed Model

- *mgcv* package
- Factors: text position, word length $\times$ log-frequency, participant

Smith & Levy, 2013, *Cognition*

Probabilities

- LSTM LM probabilities
- Estimated from Wikipedia/Soaps (80M words)

Reading Time Data (SPR)

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

Linear Mixed Model

- *lme4* package
- Factors: text position, word length $\times$ log-frequency, participant entropy, entropy reduction

Smith & Levy, 2013:

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

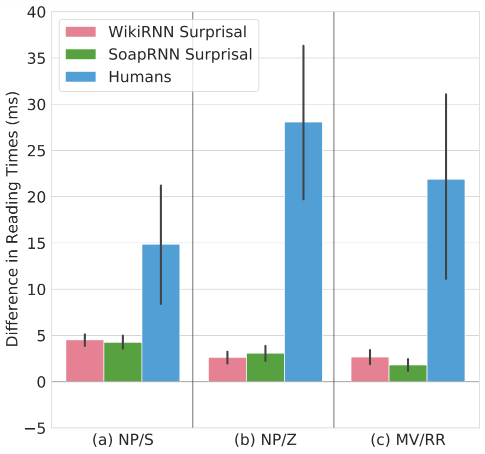
WikiRNN using Prasad & Linzen, 2019:

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

SoapRNN using Prasad & Linzen, 2019:

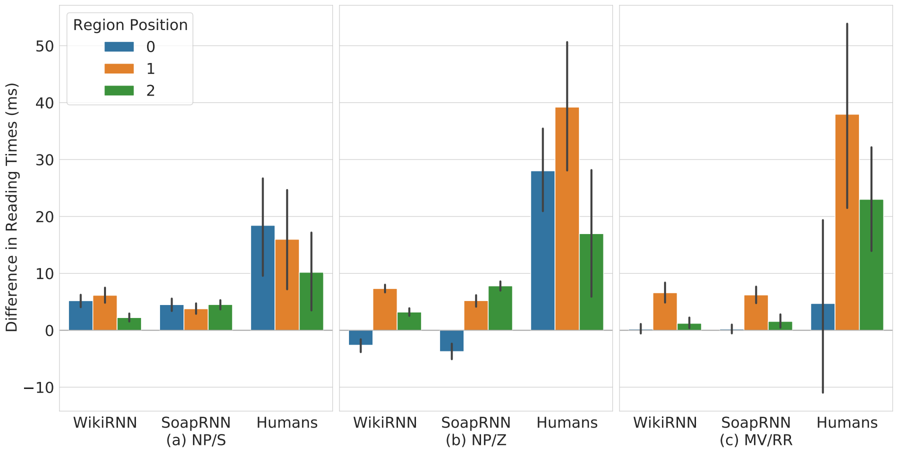
$(\delta_0 = -0.04)$ $\delta_{-1} = 0.83$ $\delta_{-2} = 0.91$ $\delta_{-3} = 0.44$

RNN garden path prediction



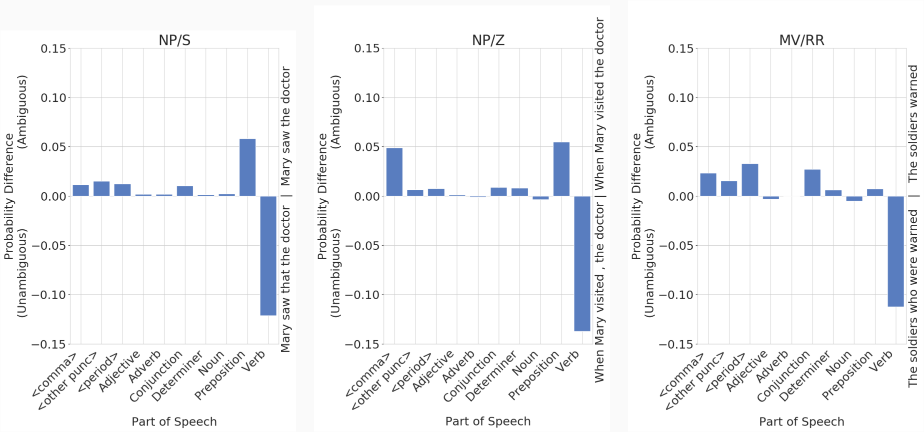
Instead of region response, examine word-by-word response

Word-by-word garden path prediction



Do RNNs garden path in a reasonable way?

Parts-of-speech predictions



Conclusion

- Conversion rates are relatively similar, but all underestimate human effect
- Suggests human processing involves mechanisms outside occurrence statistics

  (We will come back to this in Part 3)

But how well can human responses be explained by text statistics?

We know that RNNs track syntactic and semantic statistics.

What about event representations?

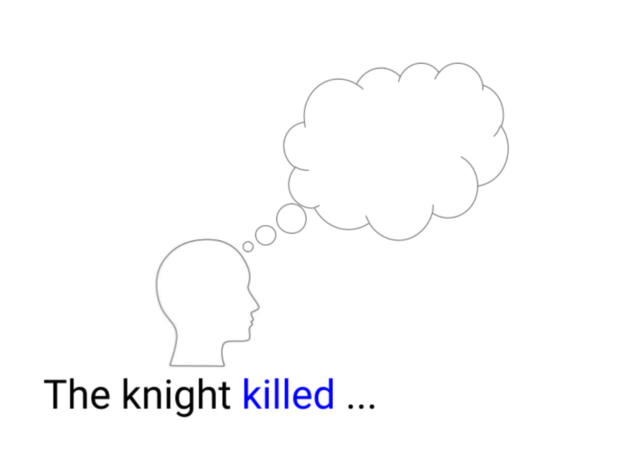
Part 2: World knowledge probing

Davis & van Schijndel, 2020, *Proc. CogSci*
Davis & van Schijndel, 2020, *Proc. CUNY*

(1) a. Context - Several horses were being raced.
    b. Target - The horse raced past the barn fell.

Knowledge of the situation mitigates the garden path

Context: One knight exists

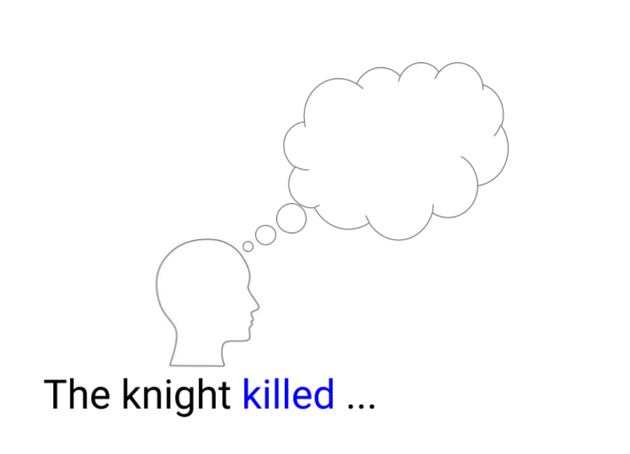


Spivey-Knowlton et al., 1993, *Canadian Journal of Experimental Psychology*

Context: One knight exists

Spivey-Knowlton et al., 1993, *Canadian Journal of Experimental Psychology*

Context: Two knights exist



Spivey-Knowlton et al., 1993, *Canadian Journal of Experimental Psychology*

Context: Two knights exist

Spivey-Knowlton et al., 1993, *Canadian Journal of Experimental Psychology*

(2) a. **Context**

(i) 1NP - A knight and his squire were attacking a dragon. With its breath of fire, the dragon killed the knight but not the squire.

(ii) 2NP - Two knights were attacking a dragon. With its breath of fire, the dragon killed one of the knights but not the other.

b. **Target**

(i) Reduced - The knight killed by the dragon fell to the ground with a thud.

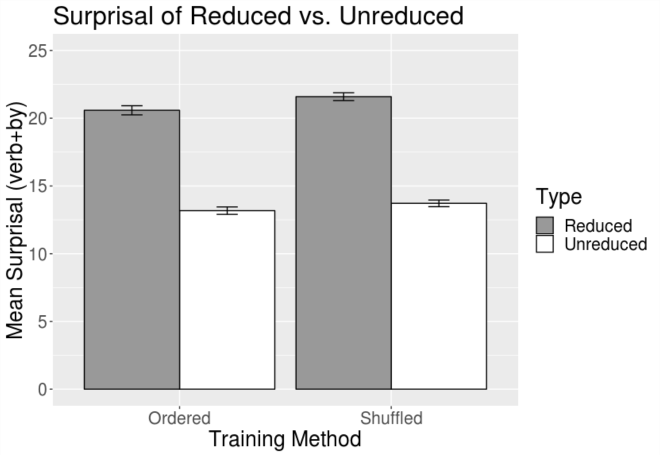
(ii) Unreduced - The knight who was killed by the dragon fell to the ground with a thud.

Spivey-Knowlton et al., 1993, *Canadian Journal of Experimental Psychology*

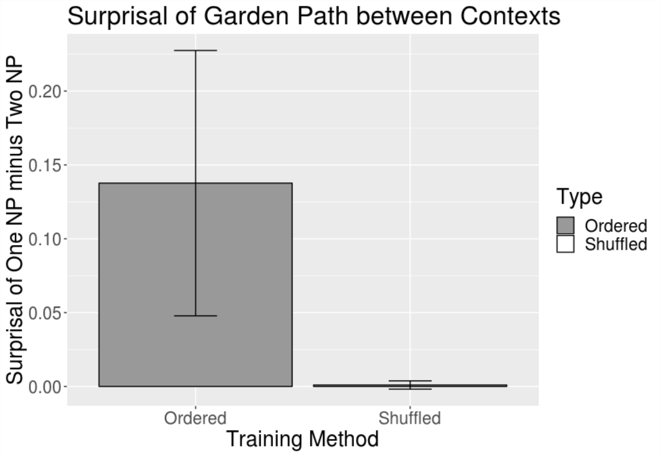
- Models: 5 LSTMs with shuffled context
  5 similar models but with intact context
  trained with different random seeds on 80M Wikipedia
- Test data: Spivey-Knowlton et al. (1993)
  Trueswell & Tanenhaus (1991)

We sum the surprisal of *verb+by*

All models predict garden path effect



Reference mitigates garden path



In humans, temporal context also mitigates garden paths

Trueswell & Tanenhaus, 1991, *Language and Cognitive Processes*

(3) a. **Context**

(i) Past - Several students were sitting together taking an exam in a large lecture hall earlier today. A proctor noticed one of the students cheating.

(ii) Future - Several students will be sitting together taking an exam in a large lecture hall later today. A proctor will notice one of the students cheating.

b. **Target**

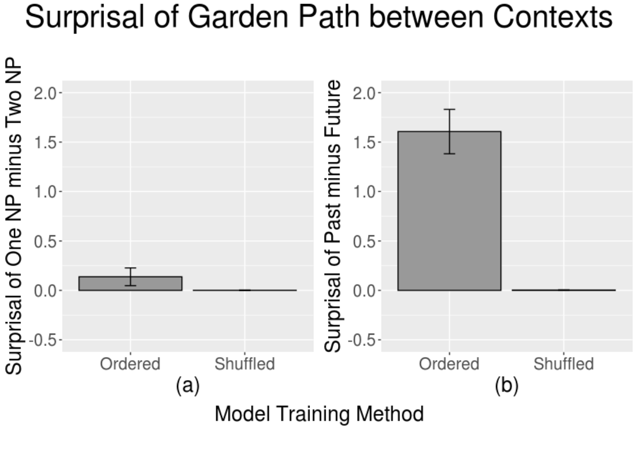
(i) Reduced - The student spotted by the proctor received/will receive a warning.

(ii) Unreduced - The student who was spotted by the proctor received/will receive a warning.

Trueswell & Tanenhaus, 1991, *Language and Cognitive Processes*


Surprisal of Garden Path between Contexts
Surprisal of Past minus Future
1.5
1.0
0.5
0.0
Ordered
Shuffled
Training Method
Type
Ordered
Shuffled

RNNs predict larger temporal mitigation



Conclusions

- Models learn tense information robustly
- Referential context and definiteness are less robust
- RNNs learn enough about discourse to mitigate garden paths (only when trained with intact discourse)
- Event knowledge is encoded in text.
  Understandable since we talk about the world, but still crazy

The problem with garden paths:

The human response correlates with the occurrence statistics

Is there a case where the learned occurrence statistics don't reflect the observed response?

Part 3: Production / comprehension mismatch

Davis & van Schijndel, 2020, *Proc. ACL*

RNNs have an observed recency bias

Idea: Maybe that prevents them from learning known human biases

Recency confounds attachment height

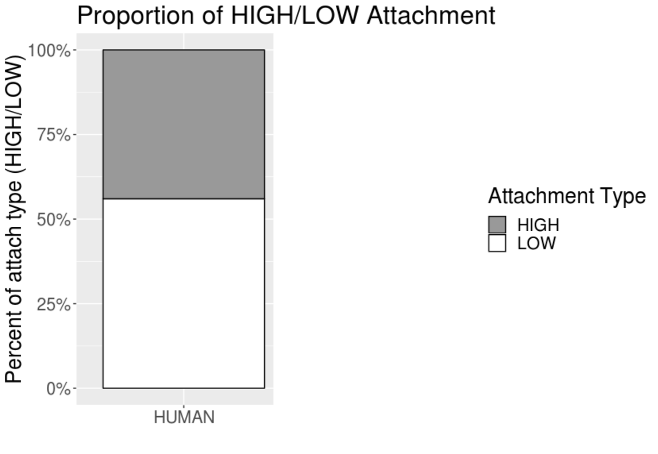
Ravfogel et al., 2019, *Proc. NAACL*

(4) a. Andrew had dinner yesterday with the nephew of the teachers that was divorced.

b. Andrew had dinner yesterday with the nephews of the teacher that was divorced.

Fernández, 2003, *Bilingual Sentence Processing*

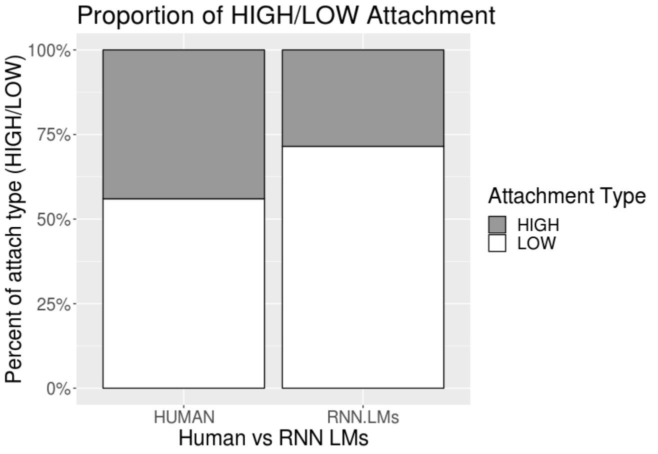
Humans attach LOW/LOCAL/RECENT



Fernández, 2003, *Bilingual Sentence Processing*

- Models: 5 Gulordava et al. (2018) LSTMs trained with different random seeds
- Test data: Fernández (2003), Carreiras & Clifton Jr. (1993), POS templates

English models attach LOW/LOCAL/RECENT



Local

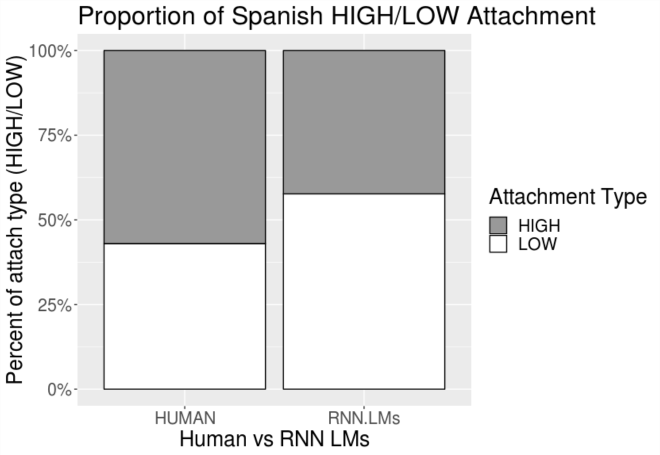
Non-local

Afrikaans	Japanese
Arabic	Norwegian
Croatian	Persian
Danish	Polish
Dutch	B. Portuguese
English	Romanian
French	Russian
German	Spanish
Greek	Swedish
Italian	Thai

Brysbaert & Mitchell, 1996/2008, *Quarterly Journal of Experimental Psychology Section A*

- Models: 5 Gulordava et al. (2018) LSTMs
  trained with different random seeds on 80M tokens of Spanish Wikipedia
- Test data: Fernández (2003),
  Carreiras & Clifton Jr. (1993),
  POS templates

Spanish models attach LOW/LOCAL/RECENT



Maybe the recency bias prevents them from learning HIGH attachment?

Experiment:
Manipulate attachment preference in synthetic training corpus

(5) a. D N (P D N) (Aux) V (D N) (P D N)
b. D N Aux V D N ‘of’ D N ‘that’ ‘was/were’ V

(6) a. The nephew near the children was seen by the players next to the lawyer.
b. The gymnast has met the hostage of the women that was eating.

- Models: 5 2-layer unidirectional LSTMs trained with different random seeds
- Training data: Synthetic corpus
- Test data: 300 ambiguous synthetic RCs

HIGH attachment is easy to learn!

1. Training: All RCs attach HIGH unambiguously
   Vary number of RCs
   Result: 20/120k produces HIGH bias at test
2. Training: 10% of data has unambigous RCs
   Vary HIGH proportion
   Result: $\geq$ 50% HIGH produces HIGH bias at test

If HIGH is easy to learn, why don’t the Spanish models learn it?

- Wikipedia: LOW is 69% more common
- Newswire (AnCora; UD): LOW is 21% more common

Note that it's still possible they contain HIGH bias, just not in RCs

Scheepers, 2003, *Cognition*

- Supports idea that production and comprehension have different distributions
  e.g., Kehler and Rohde (2015, 2018)
- RNNs won't learn human comprehension from text alone
- Provides explanation why increasing training data ceases to help
- Provides explanation for why training on cognitive signals improve model accuracy
  (Klerke et al., 2016; Barrett et al., 2018)

Thanks!

Presentations at CUNY 2020, CogSci 2020, and ACL 2020!

CUNY 2020
Recurrent neural networks use discourse context in human-like garden path alleviation

CogSci 2020
Interaction with context during recurrent neural network sentence processing

ACL 2020
Recurrent neural network language models always learn English-like relative clause attachment