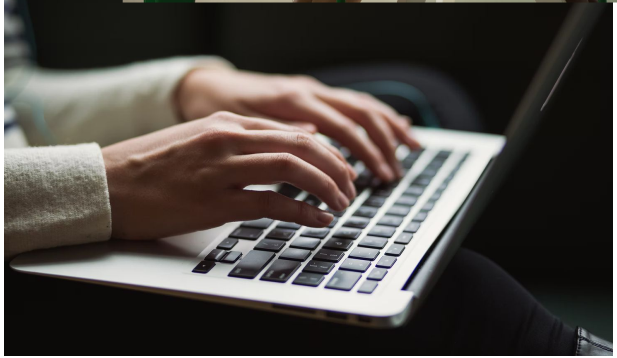
Language Statistics

won’t solve

Language Processing

Marten van Schijndel
Department of Linguistics, Cornell University
December 3, 2021

What is “Language Processing”?



What is “Language Processing”?

Marr’s Levels

- Computational:
  Most NLP applications (sentiment analysis, machine translation, summarization, etc)

- Algorithmic / Representational:
  Some parsing, NN interpretability, computational psycholinguistics

- ~~Implementational~~

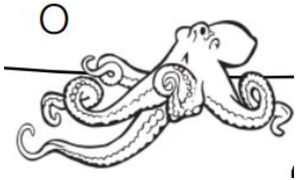
Two kinds of statistical learning naysayers

Generative Linguists

- Poverty of the stimulus
- Language requires special innate cognitive biases

Multimodality Proponents

- Can’t learn meaning from form (Bender & Koller, 2020)
- Need to be embodied physically and socially (Bisk et al., 2020)

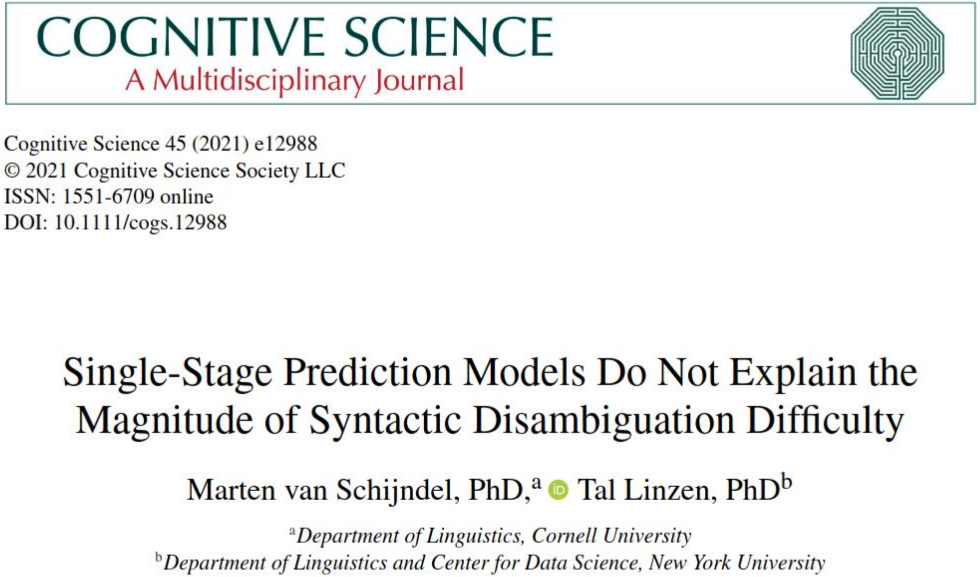


talk tldr:

Check your data

Algorithmic level requires more than Language stats

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]

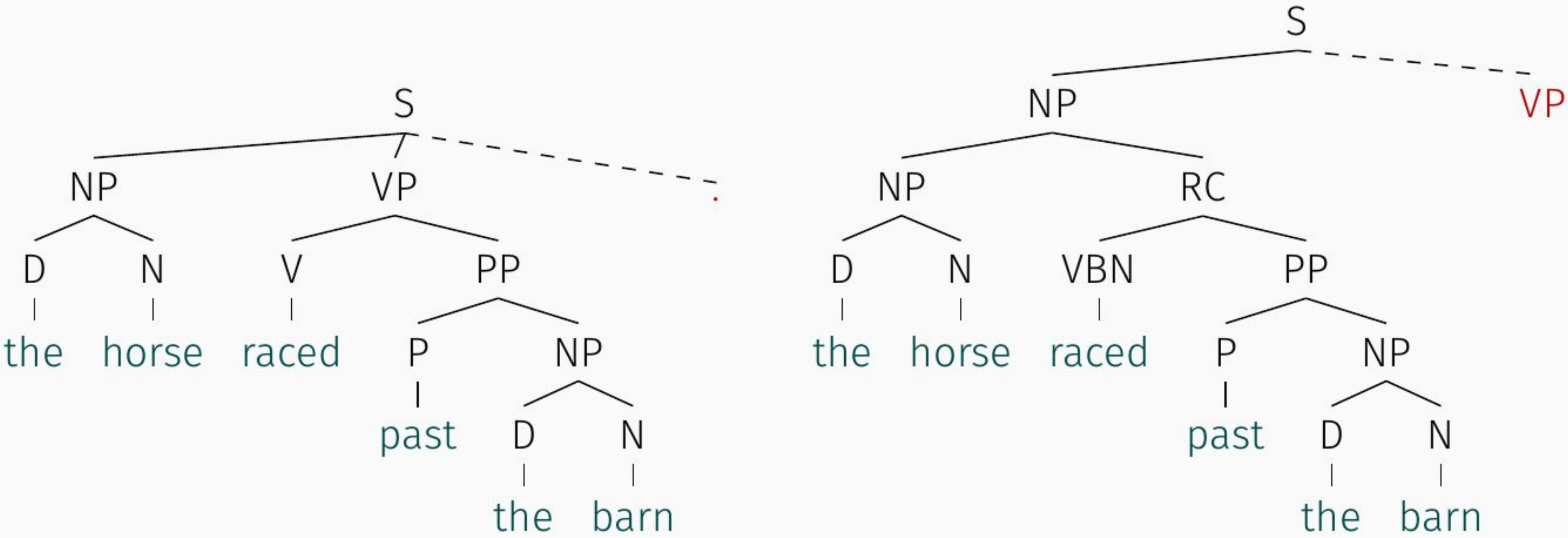
[a]Department of Linguistics, Cornell University
[b]Department of Linguistics and Center for Data Science, New York University

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*


S
NP
VP
.
D
N
V
PP
the
horse
raced
P
NP
past
D
N
the
barn
S
NP
VP
NP
RC
D
N
VBN
PP
the
horse
raced
P
NP
past
D
N
the
barn

Garden paths produce a visceral response

Garden path *responses* exist in the tail of the response distribution

They exist in the tail because

1) the statistics are in the tail (predictability)

OR

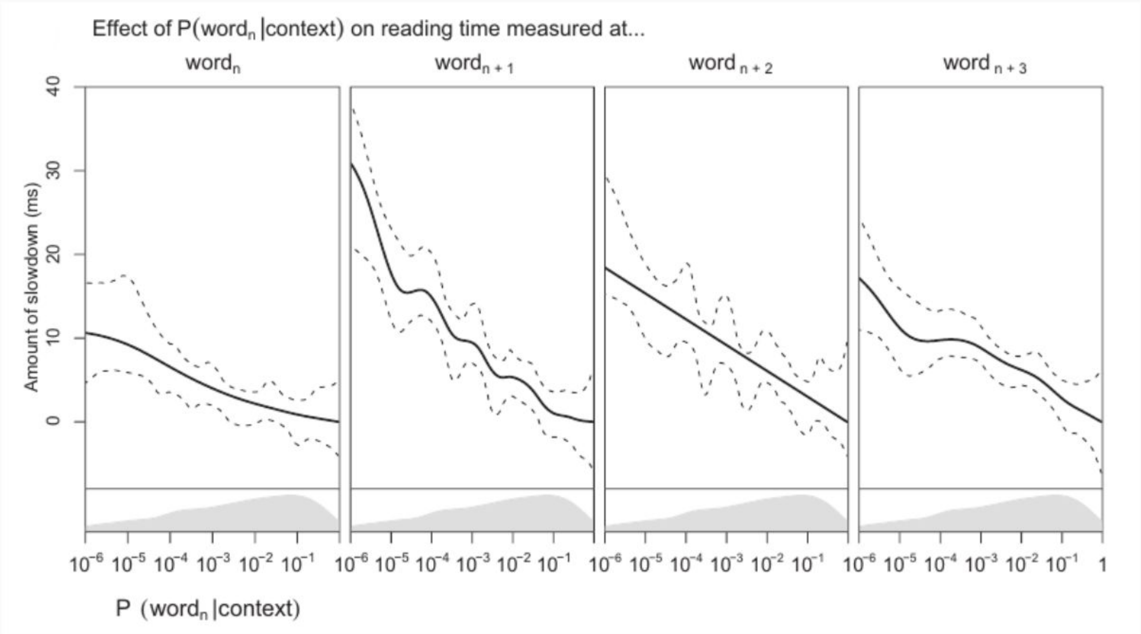
2) the response is unusual (reanalysis)

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*



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

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

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

Smith & Levy, 2013:

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

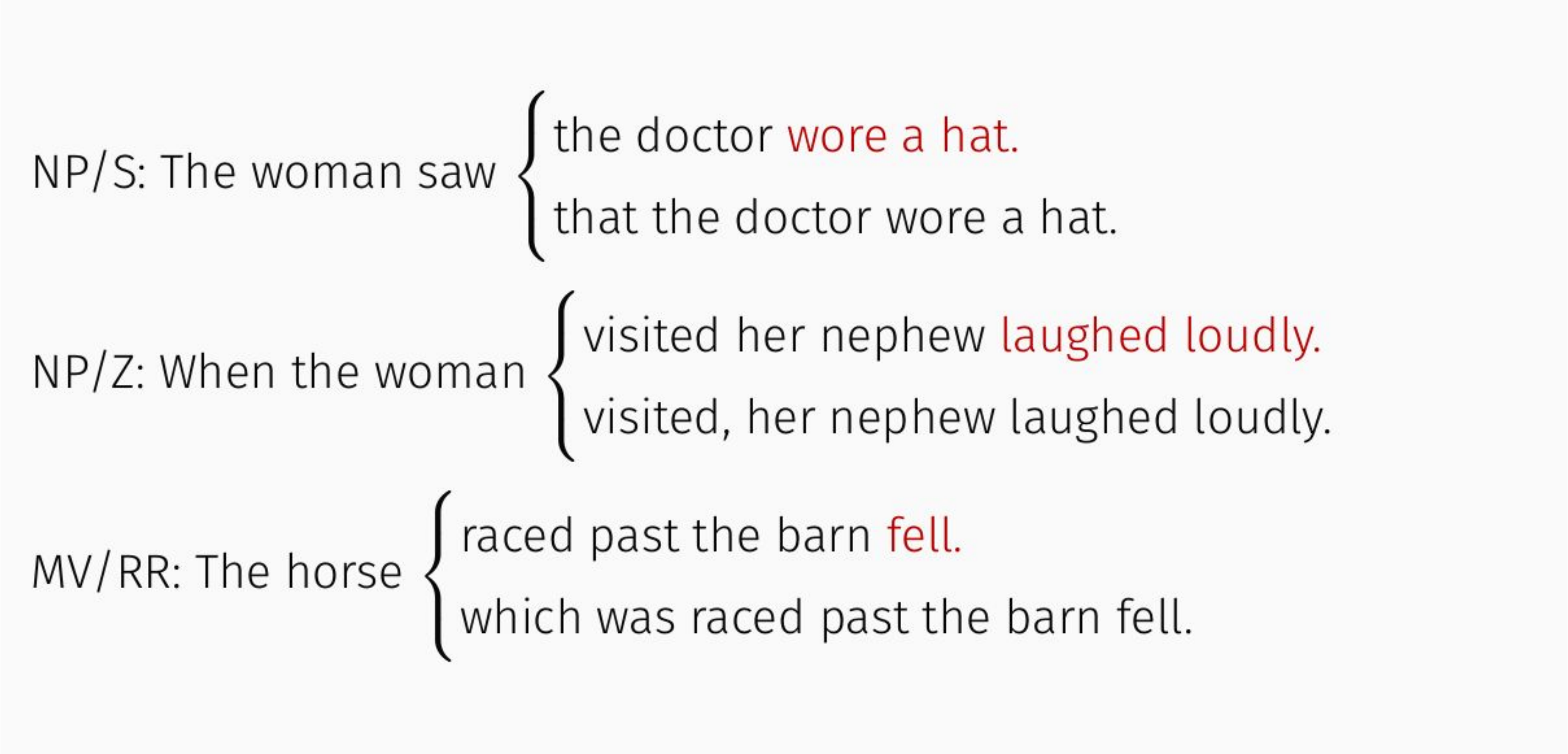
WikiRNN using Prasad & Linzen, 2019:

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

SoapRNN using Prasad & Linzen, 2019:

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

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.

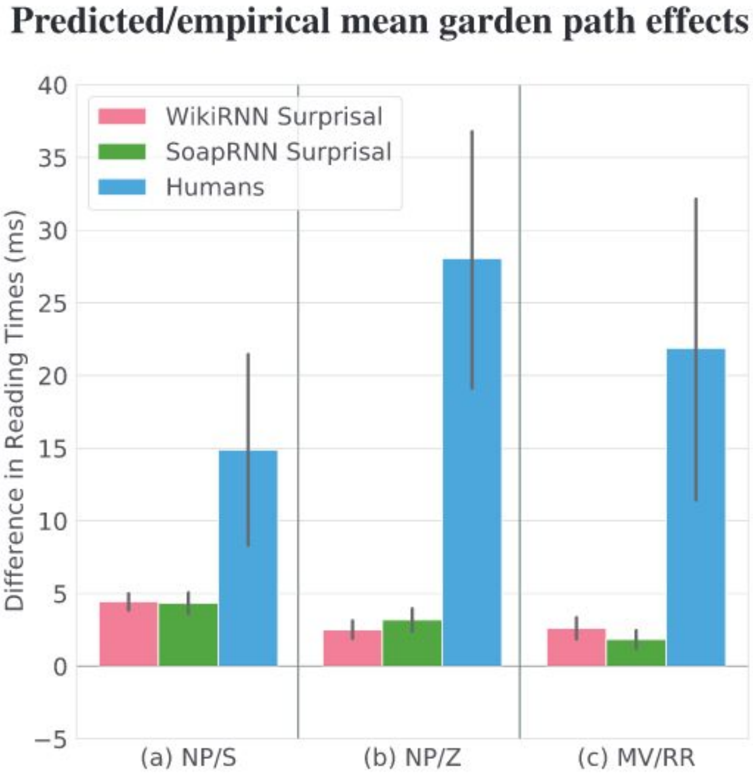
MV/RR: The horse
raced past the barn fell.
which was raced past the barn fell.

The horse raced past the barn fell

The horse which was raced past the barn fell

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

The linear relationship doesn't hold



Paper Conclusions

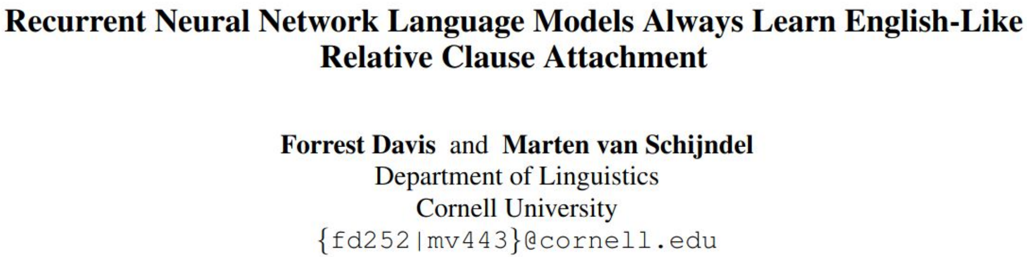
- Conversion rates are fairly similar, but all underestimate human responses
- Suggests human responses influenced by factors beyond predictability

Talk Conclusion

- Algorithmic processing cannot be learned from Language statistics

Computational level requires more than Language stats

Forrest Davis



Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment

Forrest Davis and **Marten van Schijndel**
Department of Linguistics
Cornell University
{fd252|mv443}@cornell.edu

Proceedings of ACL 2020

Does our data match our goal?

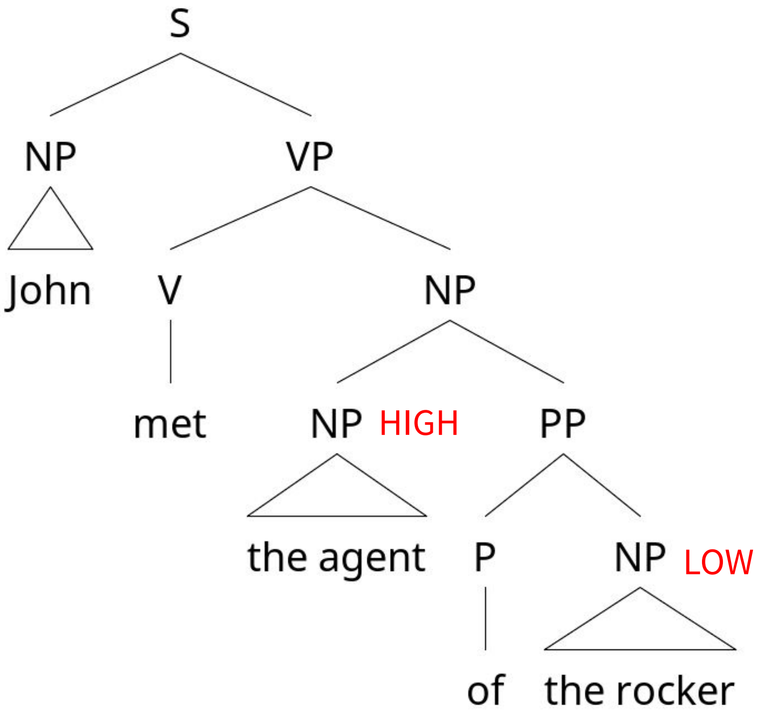
Why can we not predict garden path response sizes?

Because the boggle response is not in the training data

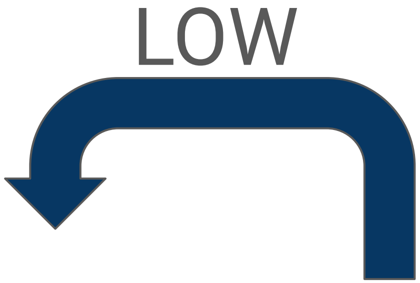
Ambiguous Relative Clause Attachment

John met the agent of the rocker *that is divorced*

Ambiguous Relative Clause Attachment



Ambiguous Relative Clause Attachment

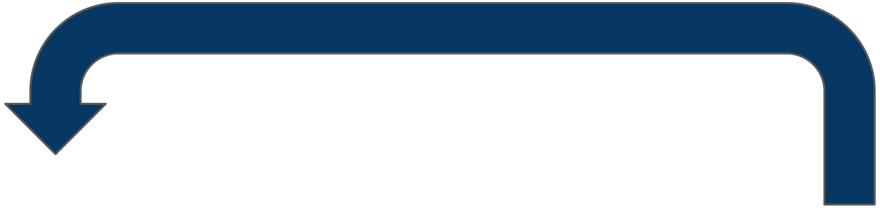


LOW

John met the agent of the rocker *that is divorced*

Ambiguous Relative Clause Attachment

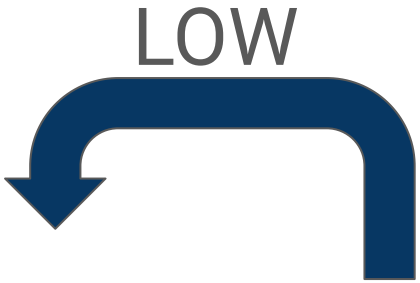
HIGH



John met the agent of the rocker *that is divorced*

Ambiguous Relative Clause Attachment

English speakers have a preference for LOW



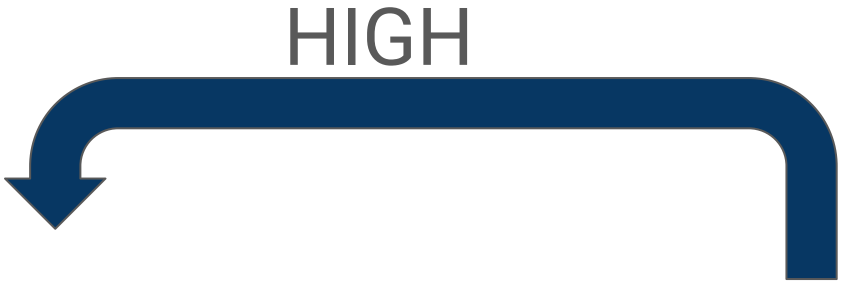
LOW

John met the agent of the rocker *that is divorced*

Carreiras and Clifton, 1993;
Frazier and Clifton, 1996;
Carreiras and Clifton, 1999;
Fernández, 2003

Ambiguous Relative Clause Attachment

Spanish speakers have a preference for HIGH



HIGH

John met the agent of the rocker *that is divorced*

Carreiras and Clifton, 1993;
Frazier and Clifton, 1996;
Carreiras and Clifton, 1999;
Fernández, 2003

Local (LOW)
Non-Local (HIGH)

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

Brysbaert and Mitchell, 1996

Do RNN LMs learn language attachment preferences?

- Used existing stimuli from psycholinguistics (40 sentence frames)
- Balanced for number

1)

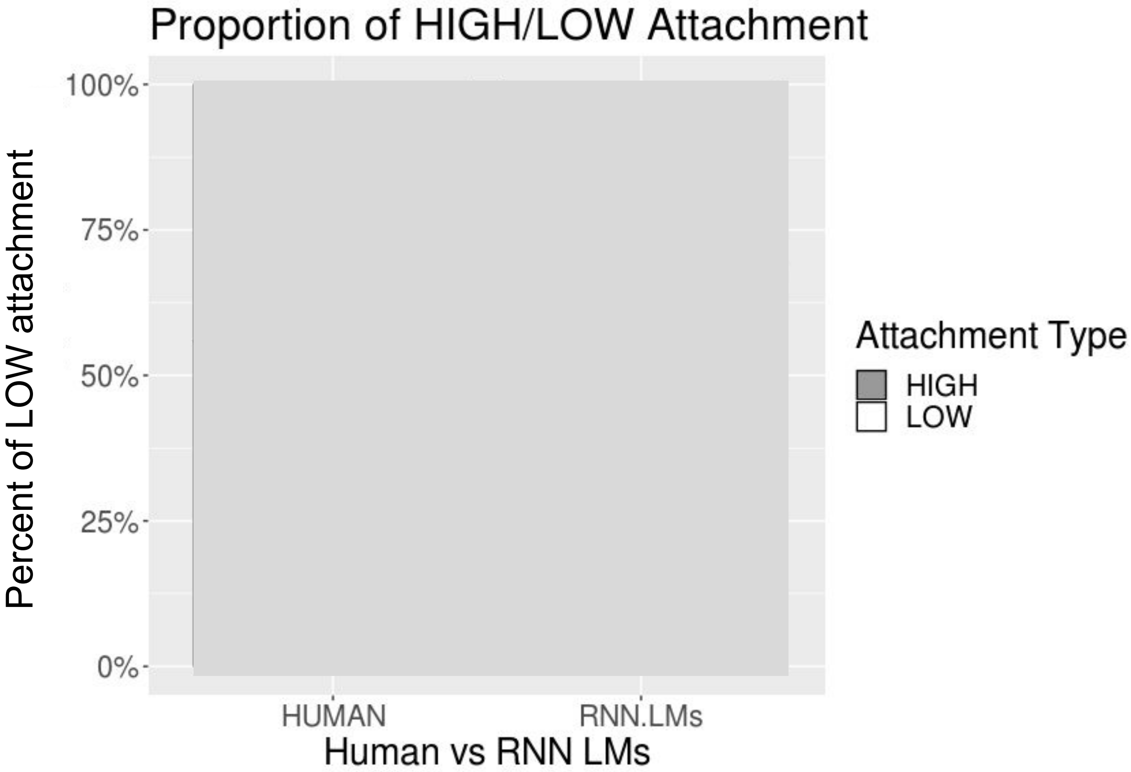
a) Andrew had dinner yesterday with the nephew of the teachers that **was** divorced. HIGH

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

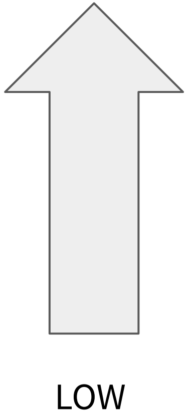
from Fernández (2003)

RNN LMs seem to have a LOW bias

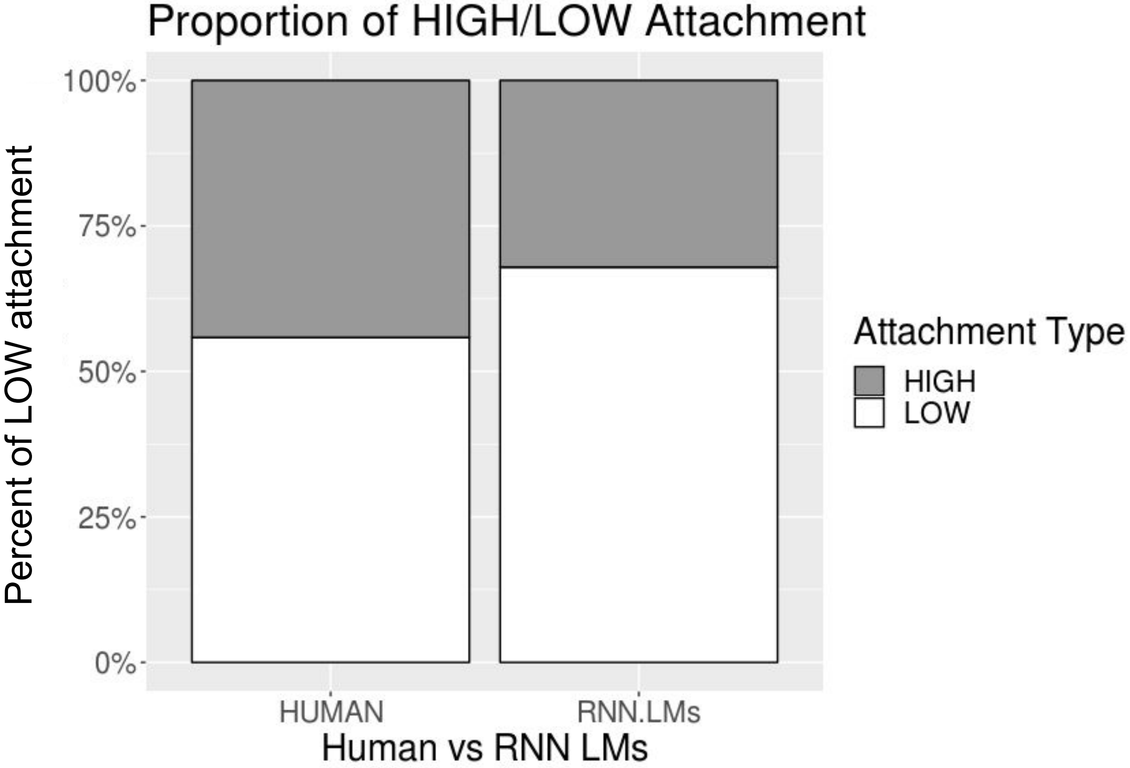
LOW



RNN LMs seem to have a LOW bias



LOW



p-value < 0.00001
Bayes Factor > 100

Do RNN LMs learn Spanish preference?

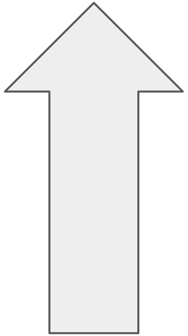
2)

a) André cenó ayer con el sobrino de los maestros que **estaba** divorciado. — HIGH

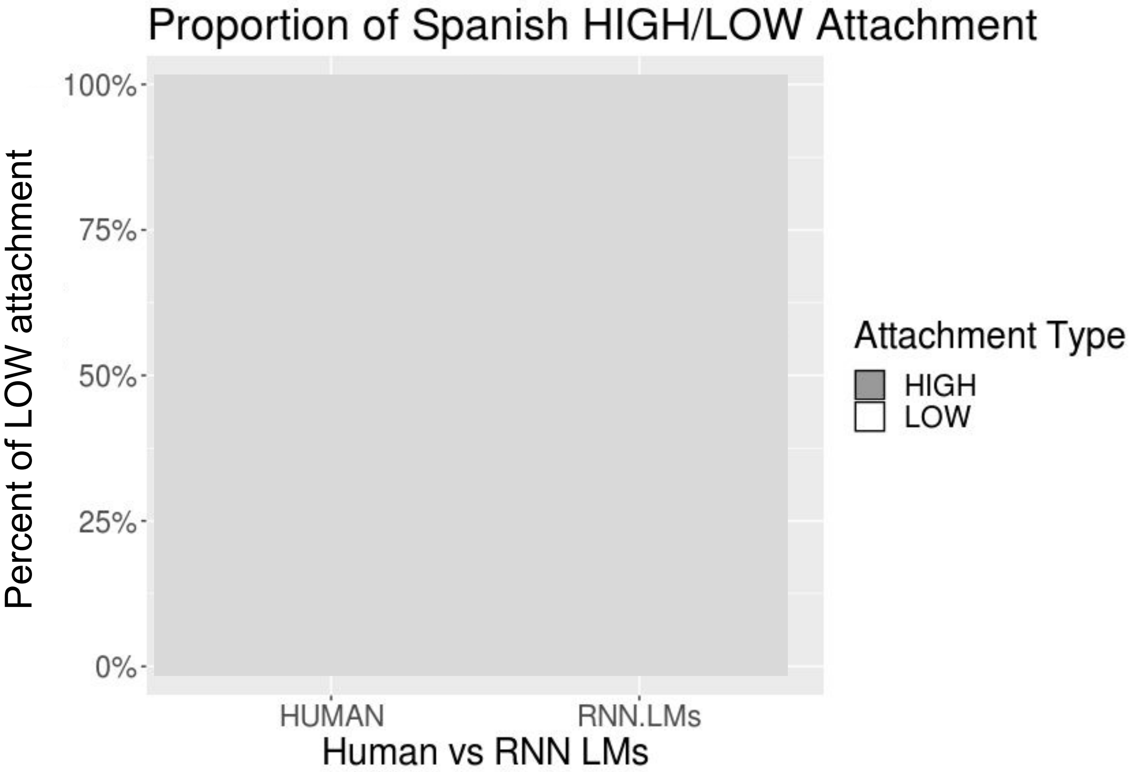
b) André cenó ayer con los sobrinos del maestro que **estaba** divorciado. — LOW

from Fernández (2003)

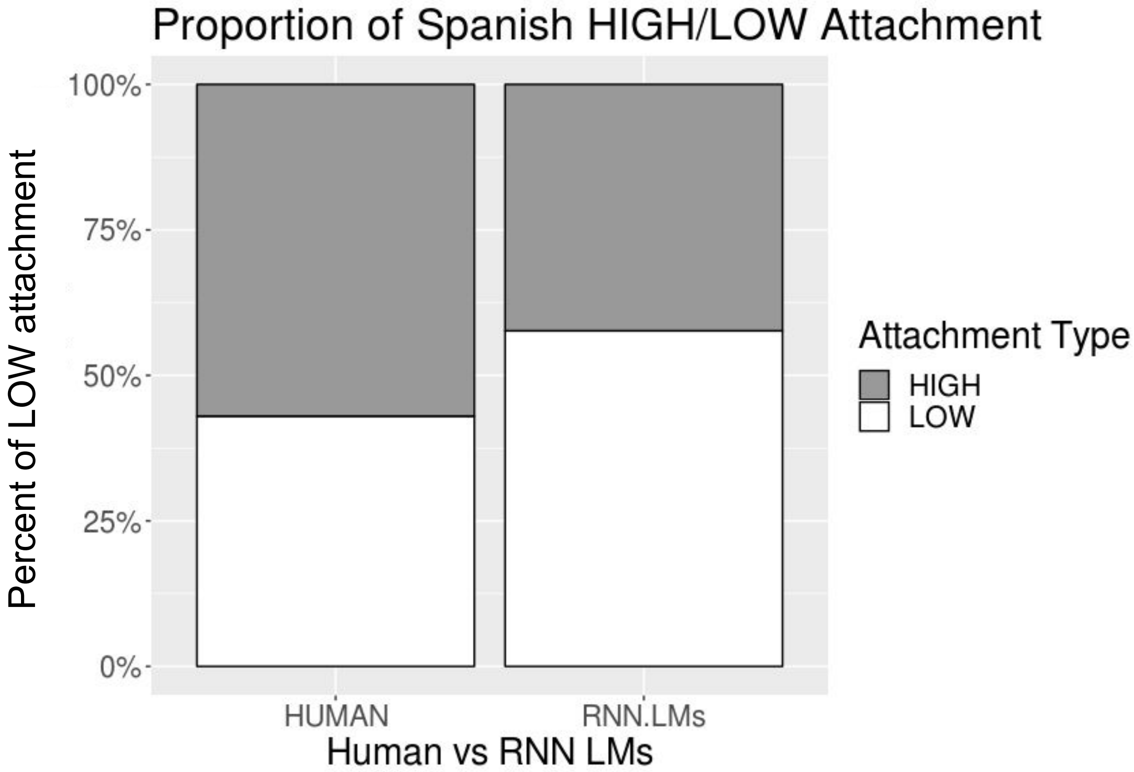
Spanish Results



LOW



Spanish Results



p-value < 0.00001
Bayes Factor > 100

Why can’t the model learn Spanish attachment?

RNN LMs can acquire HIGH or LOW bias when trained on synthetic data

- Synthetic data from PCFG with declarative sentences and sentences with the target RC construction
- 10% of training data had unambiguous RC sentences
    - Incrementing how much of that had HIGH vs LOW
- When at least 50% of RC sentences had HIGH attachment model preferred HIGH attachment

Comprehension signal not in raw text data

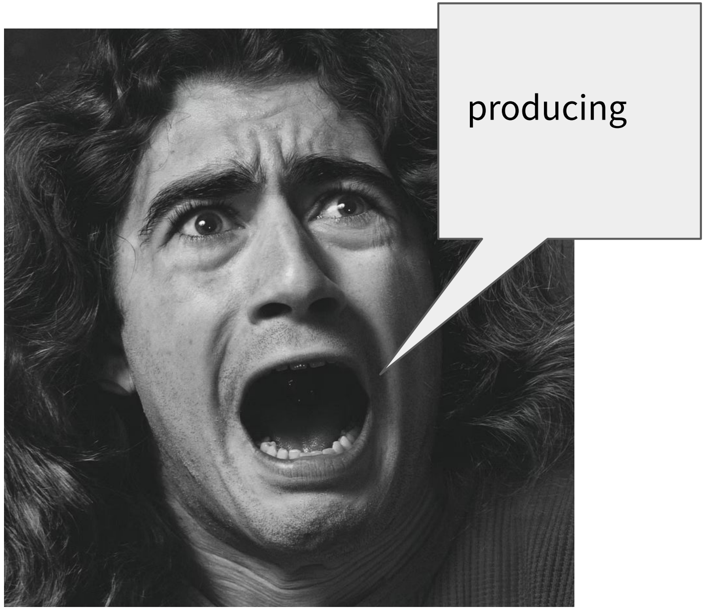
Spanish Wikipedia (training corpus):

LOW 69% more frequent than HIGH

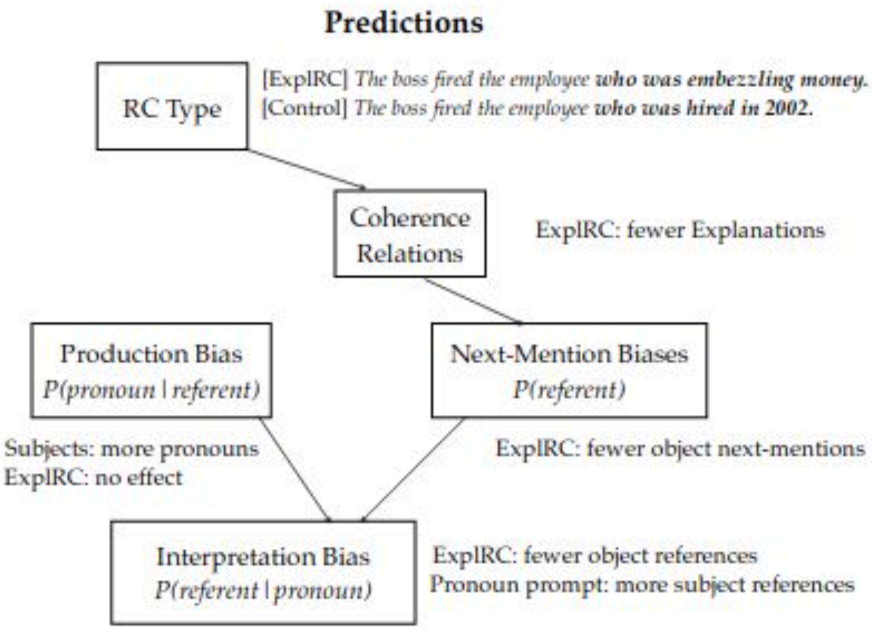
Spanish Newswire data:

LOW 21% more frequent than HIGH

Comprehension and Production



Comprehension is a superset of Production



Predictions

RC Type

[ExplRC] *The boss fired the employee* ***who was embezzling money.***

[Control] *The boss fired the employee* ***who was hired in 2002.***

Coherence Relations

ExplRC: fewer Explanations

Production Bias

P(pronoun | referent)

Subjects: more pronouns

ExplRC: no effect

Next-Mention Biases

P(referent)

ExplRC: fewer object next-mentions

Interpretation Bias

P(referent | pronoun)

ExplRC: fewer object references

Pronoun prompt: more subject references

Rohde et al., 2011

Kehler and Rohde, 2015

Kehler and Rohde, 2019

Conclusions

- Language statistics reflect human production biases
- Most NLP tasks are about comprehension

- What kind of training signal is needed for comprehension?

Thanks!

Tal Linzen

Forrest Davis

C.Psyd

Cornell NLP

Unsplash Images

Slide 2

@amadorloureiroblanco @wocintechchat
@jerry_318 @kaitlynbaker

Slide 40

@krivitskiy @roller1