Harnessing the linguistic signal in predicting within-scale variability in scalar inferences

Judith Degen

May 26, 2021

Scales, degrees and implicature workshop

Scalar implicature

(1) John: Was the exam easy?
Mary: Some of the students failed.
Inference: Some, but not all of the students failed.

(2) John: Who came to the party?
Mary: Ann or Greg.
Inference: Either Ann or Greg came, but not both.

(3) John: How was your date?
Mary: It was OK.
Inference: The date was OK, but not great.

Variability in scalar implicature

Focus on inter-scale variability, attributing variability to properties of the scale

- polarity of scale
- distinctness of scalemates
- semantic similarity of scalemates
- negative strengthening
- propensity to raise QUD about strong alternative

Doran et al 2012; van Tiel et al 2016; Benz et al 2018, Gotzner et al 2018, Sun et al 2018, Westera & Boleda 2020, Ronai & Xiang 2021

Problem

- mixed results in trying to explain SI variability via varying properties of scales (small / noisy effects)
- tested items hand-generated by researchers
- small number of items per scale

Consequence: seeming regularity in inter-scale variability may be due to frequent re-use of the same set of items across experiments, the small number of items per scale, and the possible lack of representativeness of the use of scalar items real listeners encounter in the real world.

What's lacking

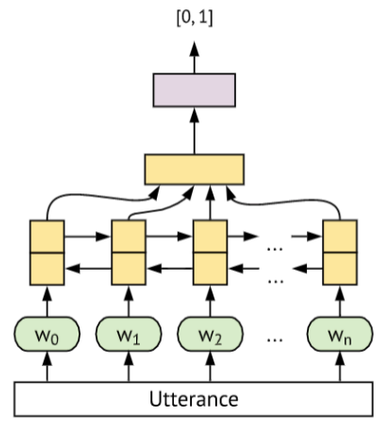
An estimate of intra-scale variability in naturalistic contexts

- an understanding of the naturalistic contexts that speakers produce scalar expressions in
- an estimate of the extent to which listeners make use of the contextual information available to them in naturalistic contexts

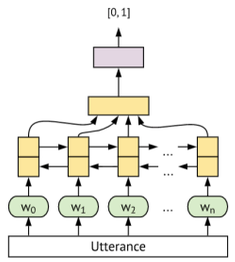
Overview

1. A study combining corpus analysis & web-based experiments on “some”

2. Using a neural network with distributed meaning representations to predict human inference ratings



3. Applying 1. and 2. to “or”



1. Case study: “some”

Degen 2015

Scalar implicatures in the wild

1. I like **some country music**.

   Inference? I like some, but not all, country music

2. It would certainly help them to appreciate **some of the things we have here**.

   Inference? …to appreciate some, but not all…

3. You sound like you have **some small ones** in the background.

   Inference? … some, but not all small ones…

Combining corpora & the web

1. extracted all 1390 utterances containing *some* from the Switchboard corpus of spoken American English

2. collected inference strength ratings for each item on Mechanical Turk (10 judgments per item)

Speaker A: i mean, they just have beautiful, beautiful homes and they have everything. the kids only wear name brand things to school and it's one of these things,

Speaker B: oh me. well that makes it hard for you, doesn't it.

Speaker A: well it does, you know. it really does because i'm a single mom and i have a thirteen year old now and uh, you know, it does.

Speaker B: oh, me.

Speaker A: i mean, we do it to a point but uh, not to where she feels different ,

Speaker B: yeah.

Speaker A:
but some of them are very rich

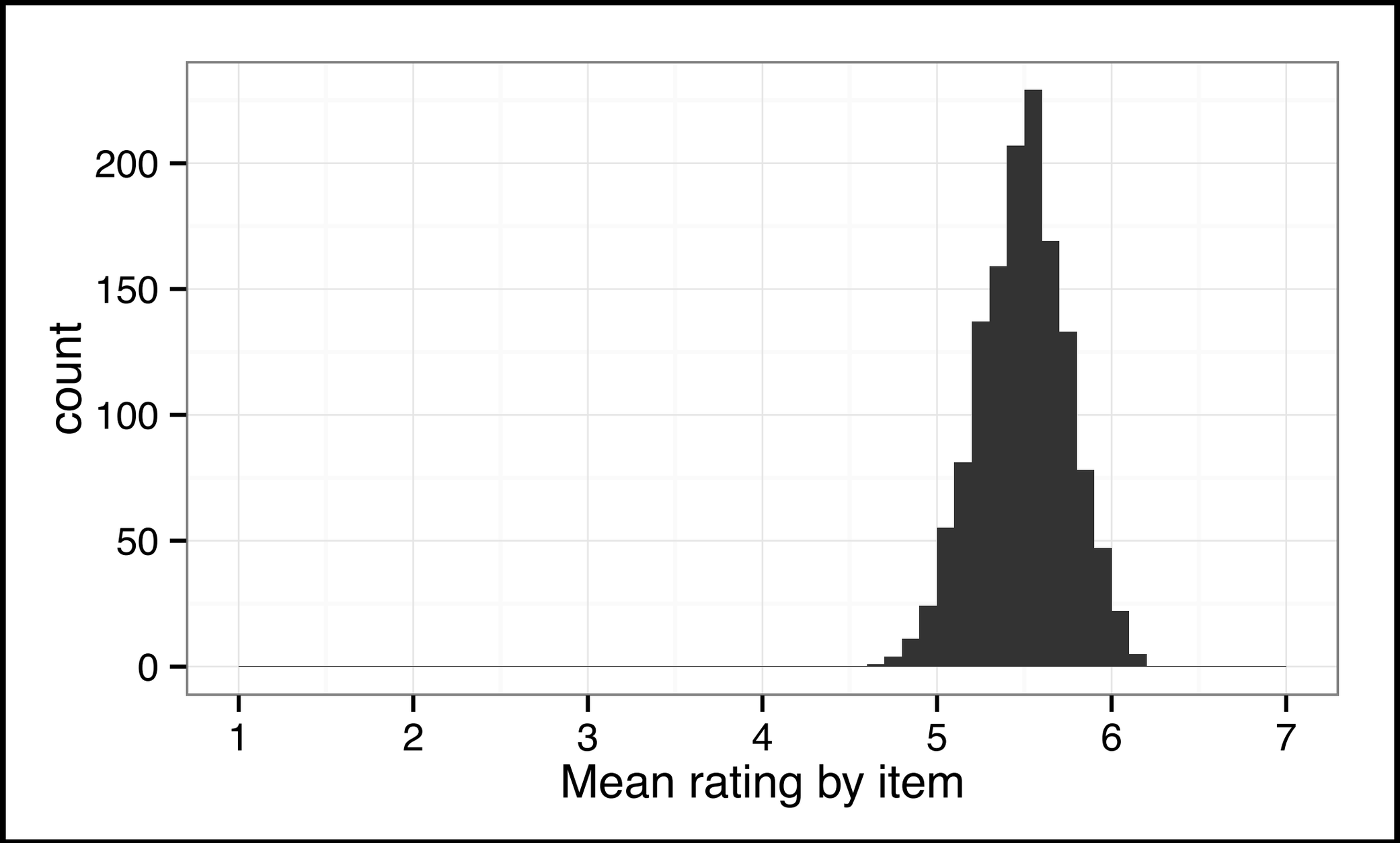
but **some, but not all** of them are very rich

How similar is the statement with 'some, but not all' (green) to the statement with 'some' (red)?

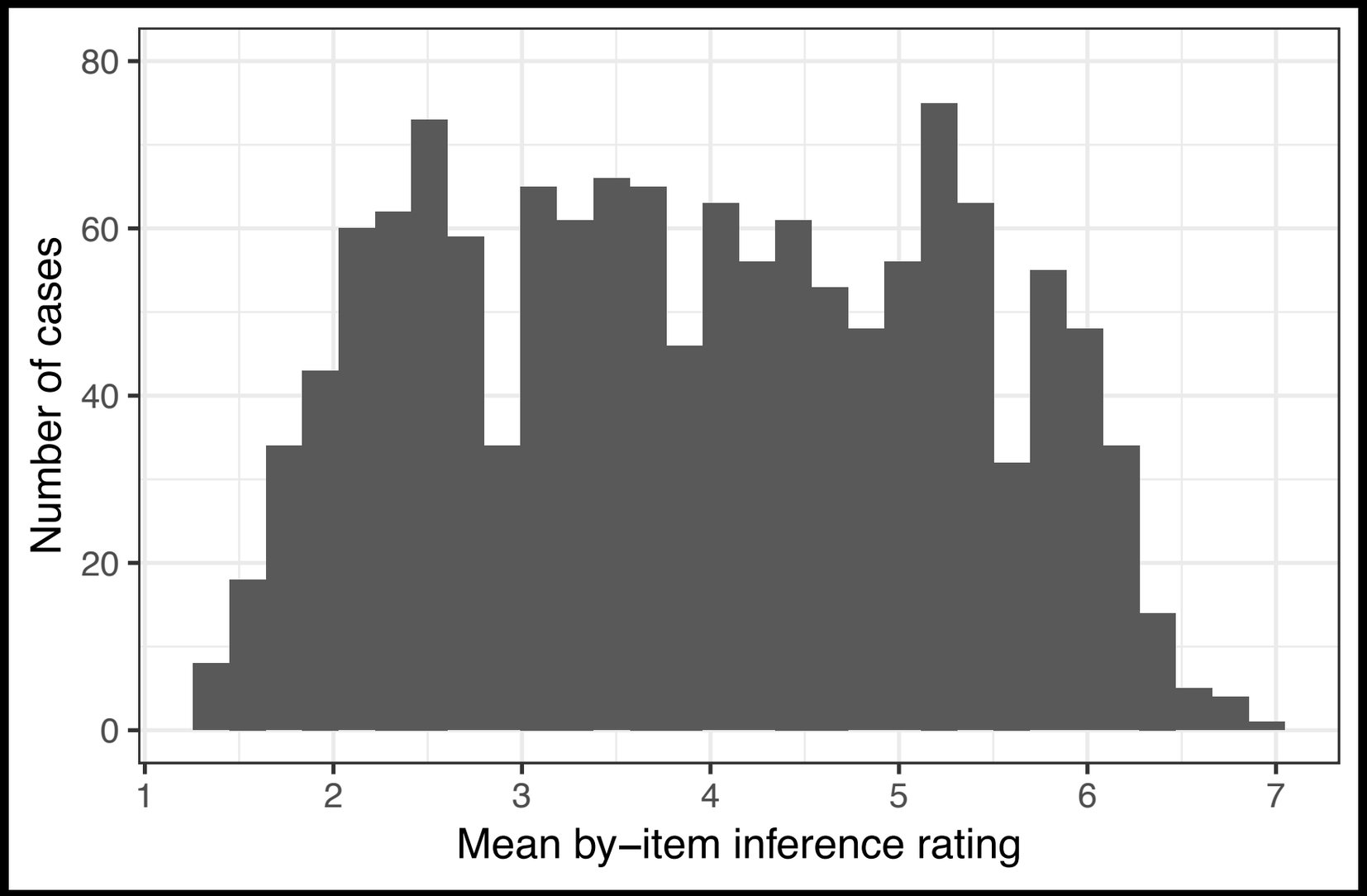
Very different meaning							Same meaning
1	2	3	4	5	6	7	

Continue

Low intra-scale variability prediction



Variability in inference strength



large amount of variability in inference strength

Just noise?

Qualitative investigation

1. I like **some country music**.

6.9

2. It would certainly help them to appreciate **some of the things we have here**.

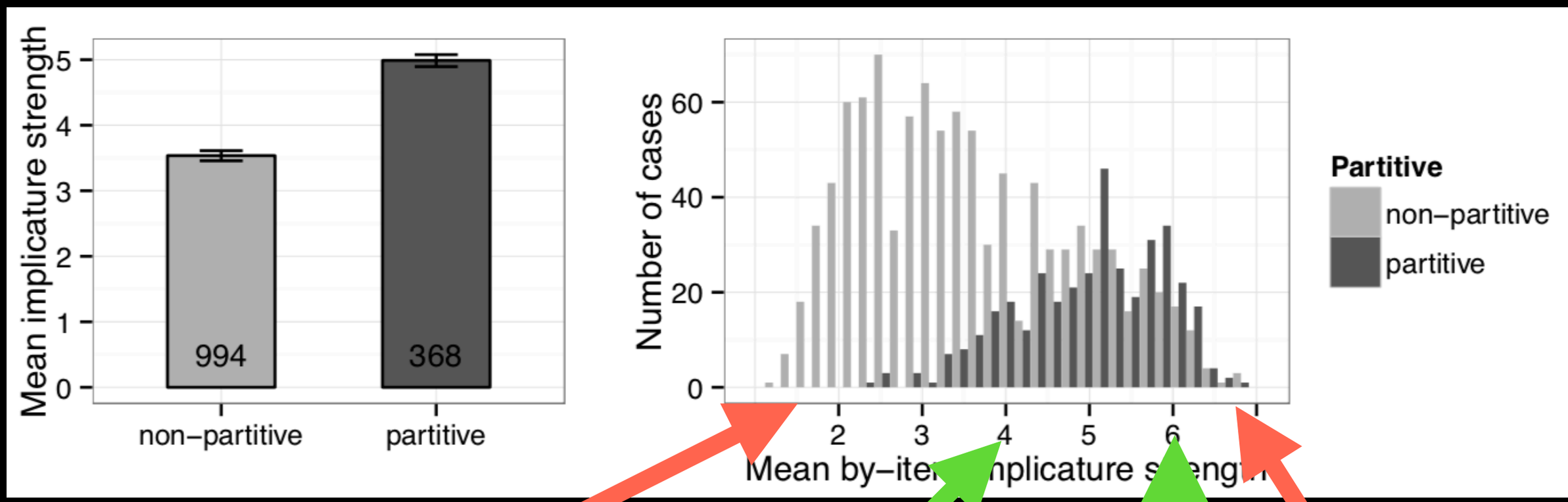
4

3. You sound like you have **some small ones** in the background.

1.5

Stronger inferences…

…with **partitive** *some*-NPs.



*I've seen **some of them** on repeats*

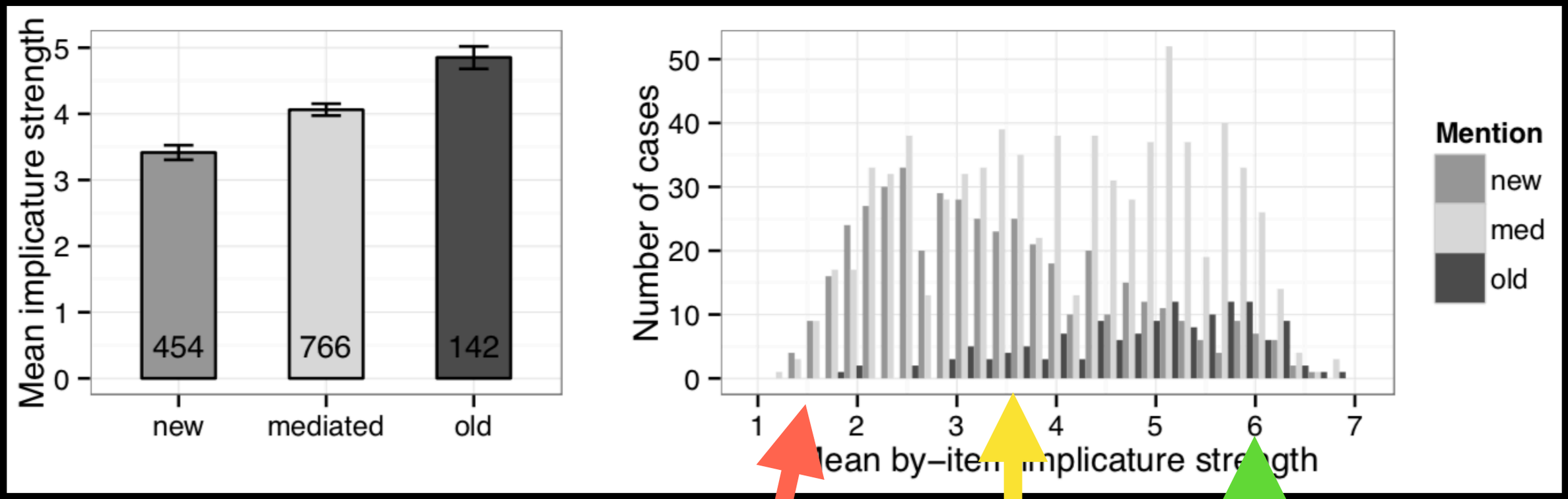
*It would certainly help them to appreciate **some of the things we have here**.*

*You sound like you have **some small ones** in the background.*

*I like **some country music.***

Stronger inferences…

…with **previously mentioned** NP referents.



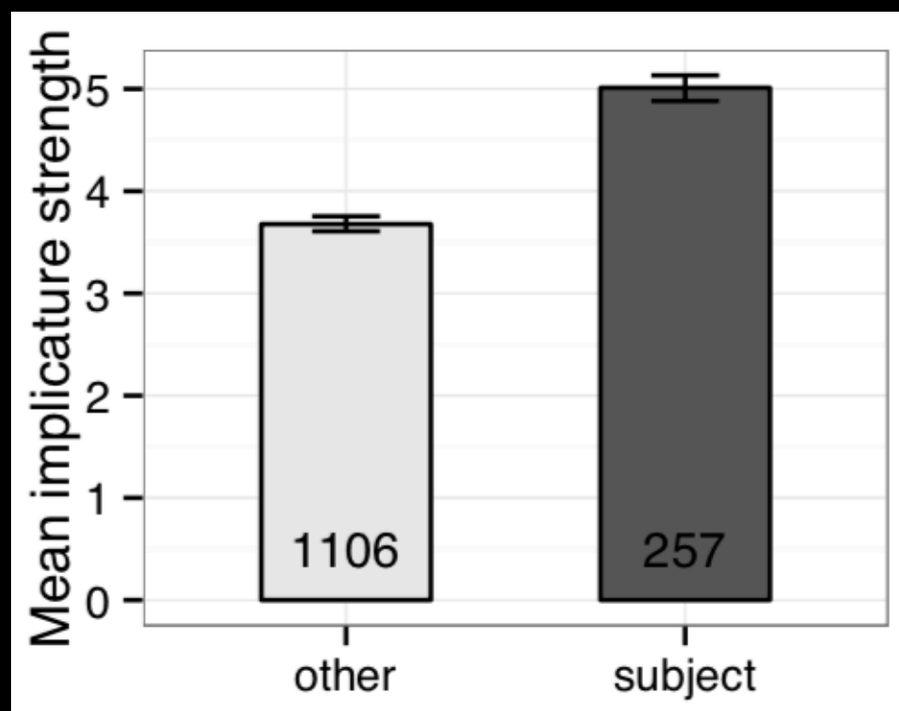
*I've seen **some of them** on repeats*

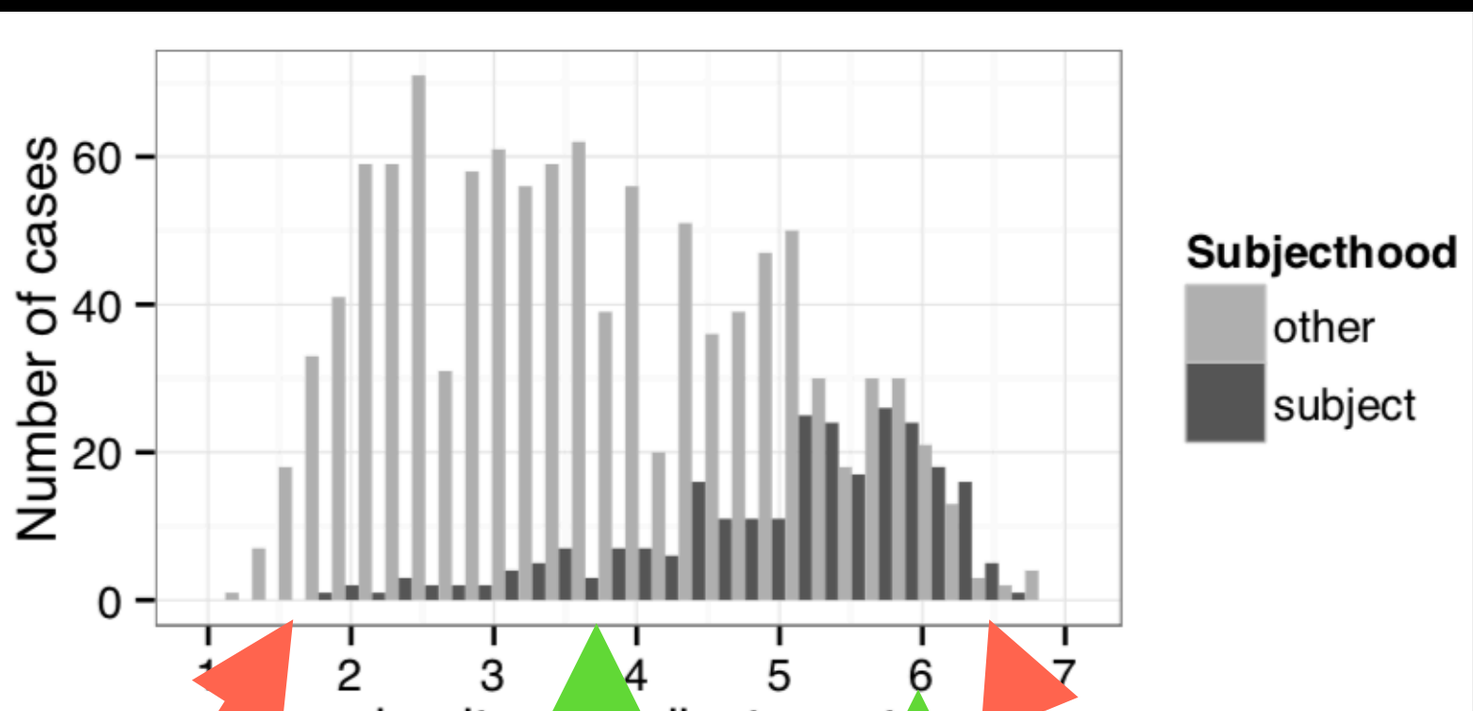
*We've got **some beets**.*

*That would take **some planning**.*

Stronger inferences…

…with *some*-NPs in **subject** position.





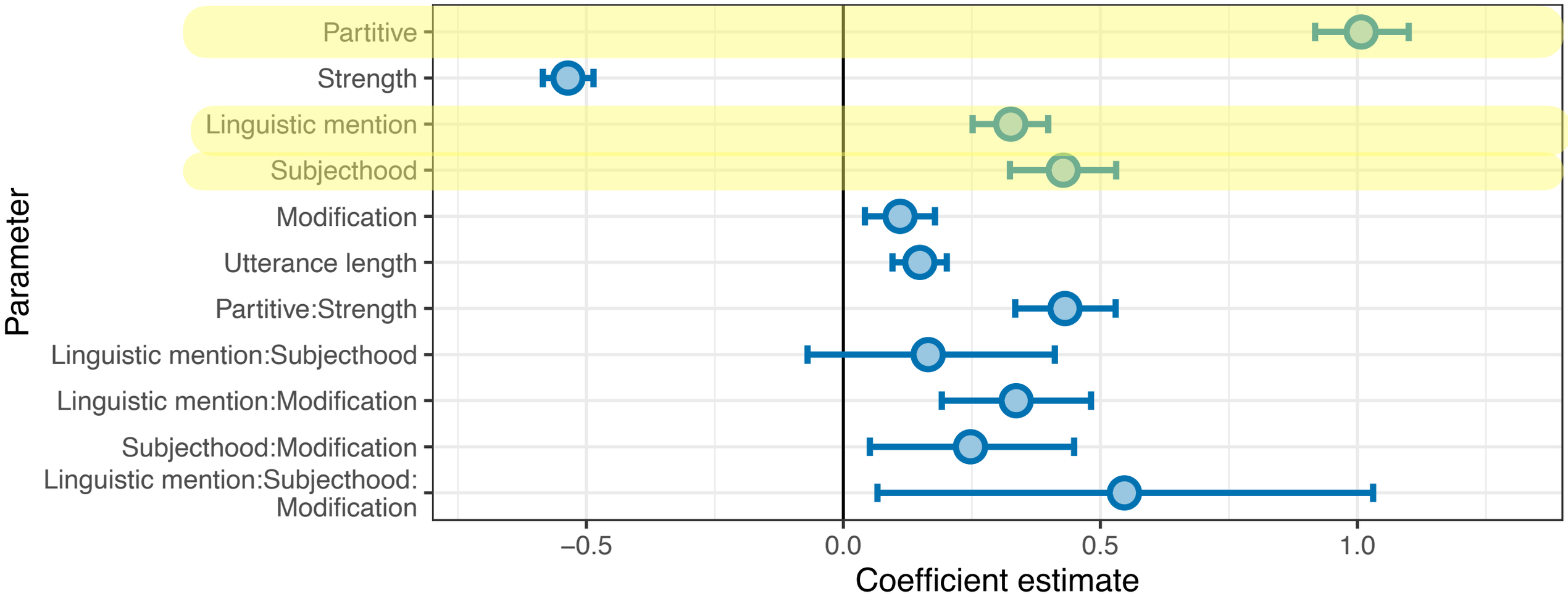
***Some kids** are really having it.*

*Occasionally, **some ice skating** will come on.*

*That would take **some planning**.*

*I like **some country music.***

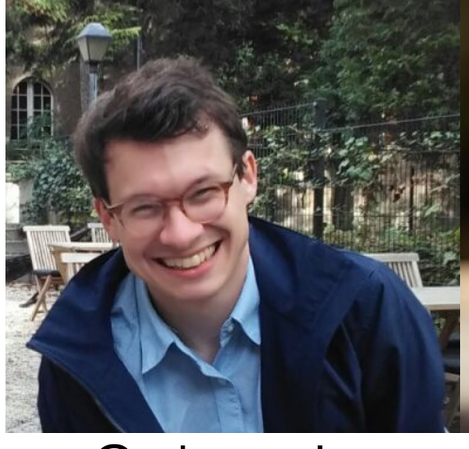
Results overview



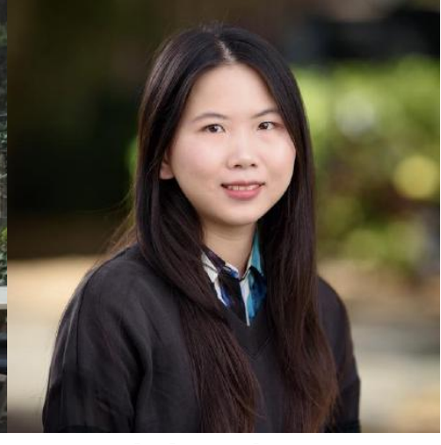
No. Replication by Eiteljoerge et al 2019 in child-directed speech.

Just noise?

No. Variability in ratings is systematically predicted by syntactic, semantic, and pragmatic features of context.



Sebastian Schuster



Yuxing Chen

2. How much information about the interpretation of “some” is contained in the linguistic signal?

Schuster, Chen, & Degen, 2020

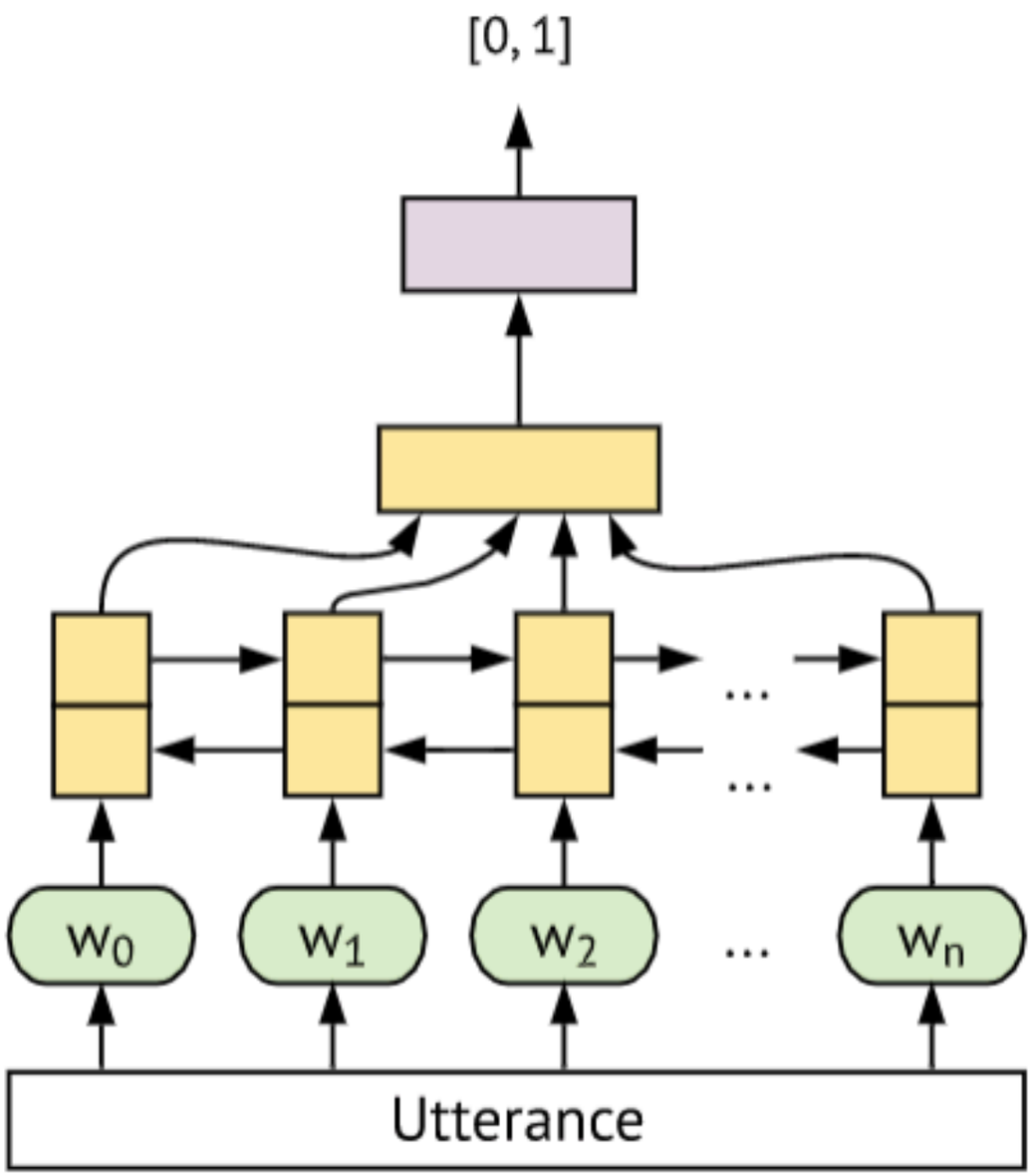
Predicting inference strength from distributed meaning representations

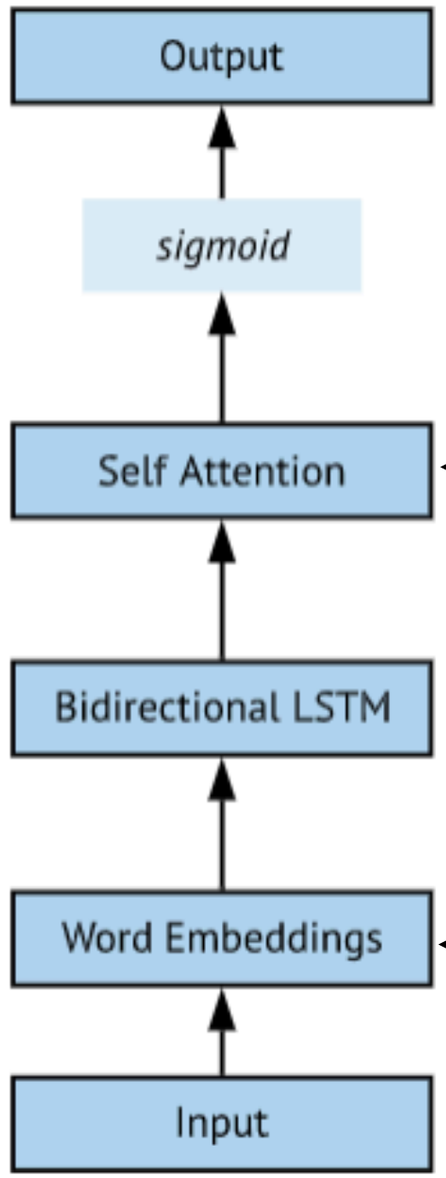
Ultimate goal:
Use distributed vector-based meaning representation methods from NLP to infer which, if any, linguistically encoded features of context listeners use in drawing inferences, to help inform pragmatic theory.

More proximate goal:
Use distributed vector-based meaning representation methods from NLP to test whether any of these methods

- reliably predict inference ratings
- capture the identified context effects

Model architecture





optional

GLoVe
ELMo
BERT

i like some country music

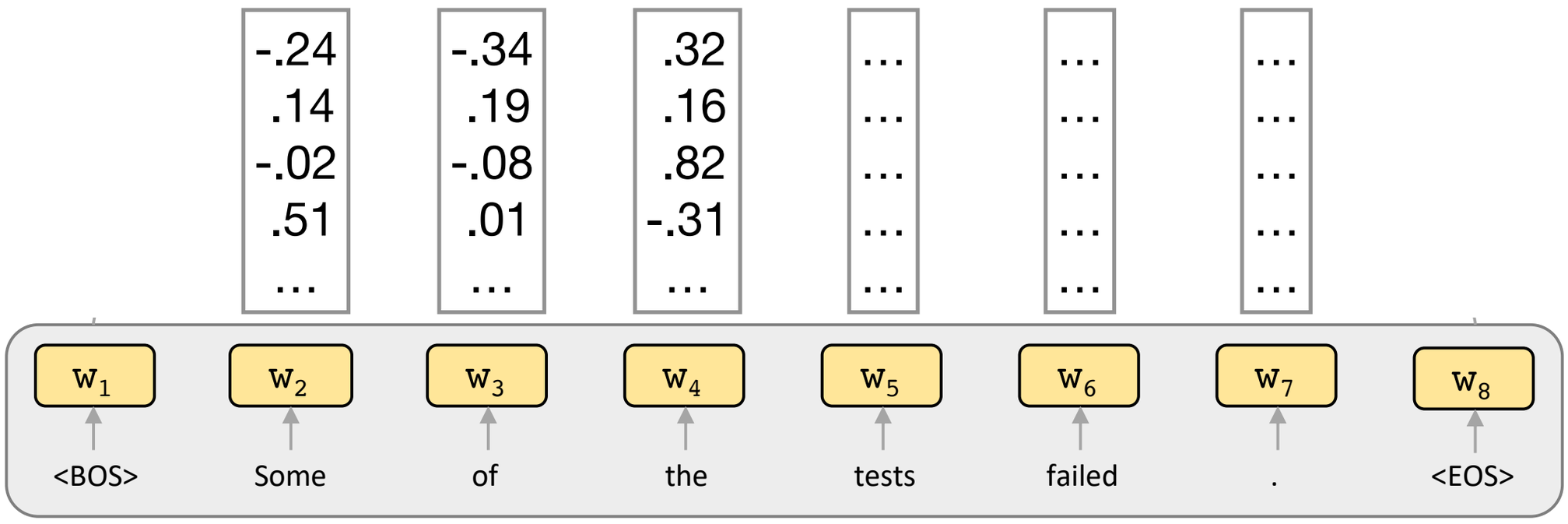
BERT

Bidirectional Encoder Representations from Transformers

Devlin et al., 2019; Wolf et al., 2019

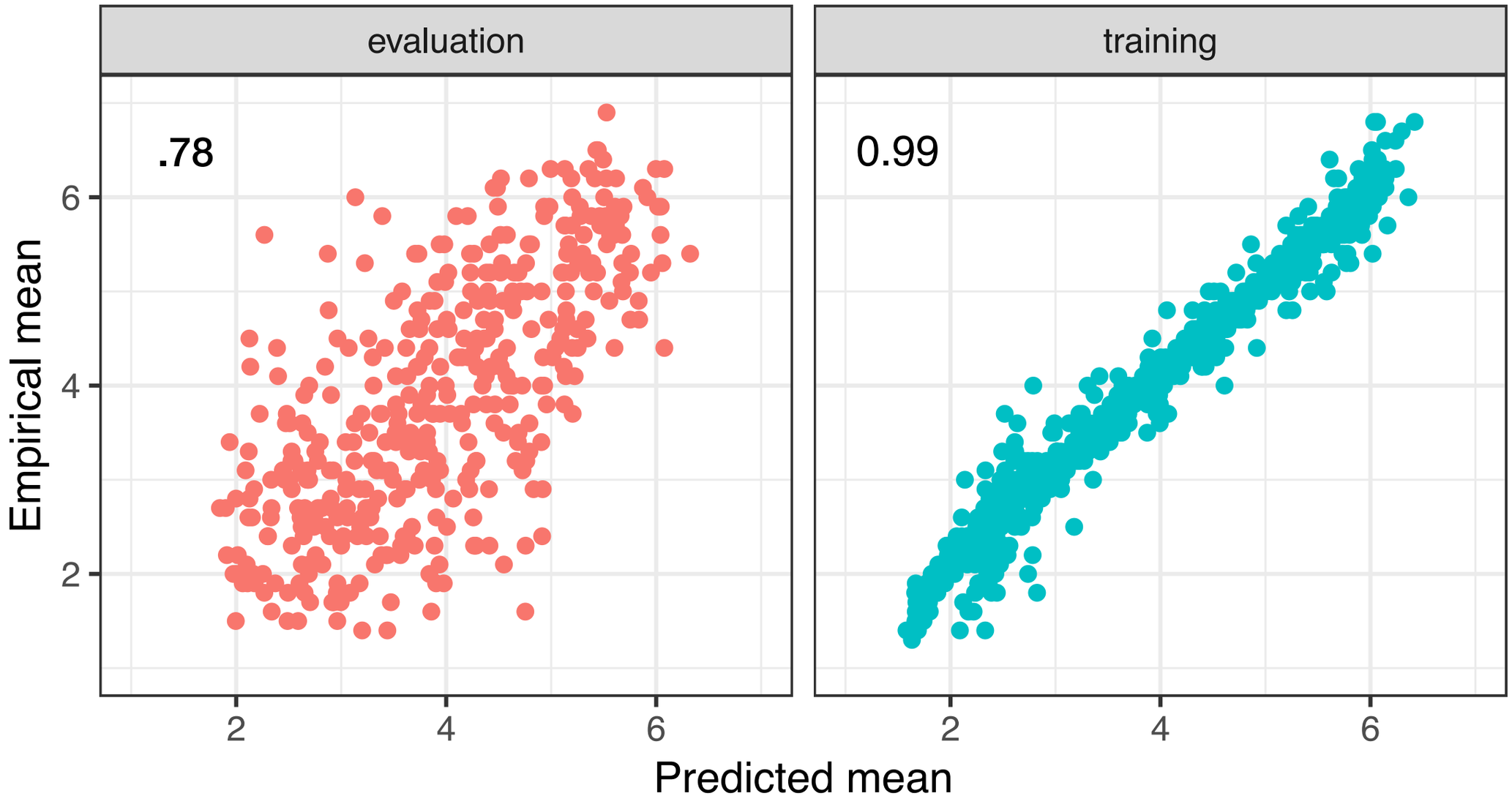
- **contextual** word embeddings (considers entire sentence before assigning a word in it an embedding)
- captures that a word's meanings can vary across sentences

1. **Apple** announced the new iPhone today.
2. **Google** announced a new browser last week.
3. I ate an **apple** for breakfast.
4. I ate an **orange** after dinner.



Model predictions

Best model: BERT — LSTM + attention — no-context

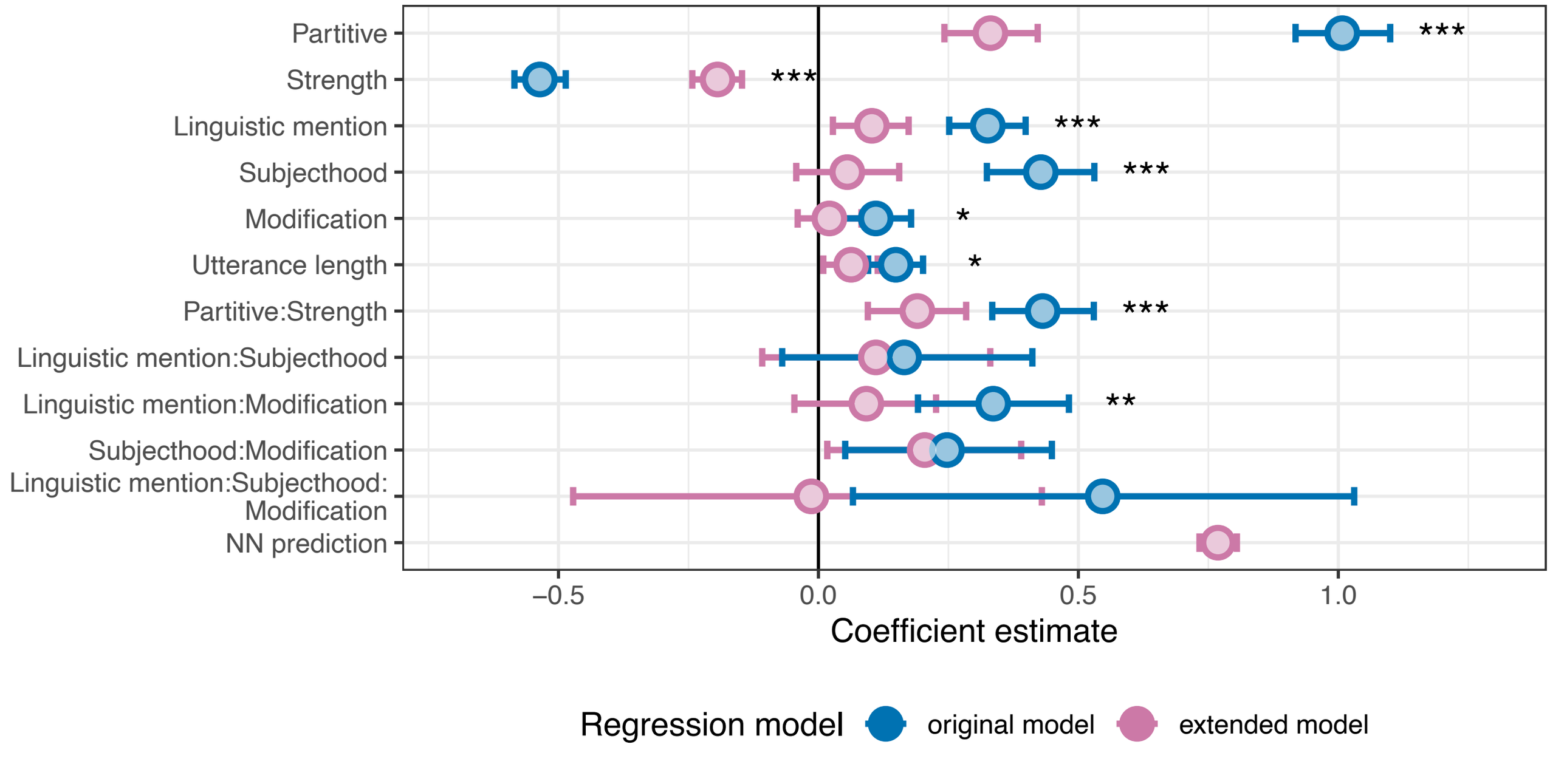


The model learned to predict “not all” inference strength ratings

Quantitative analysis

Is there any evidence that the model captures the same effects that the hand-mined feature model did?

Quantitative comparison with hand-mined model



Quantitative analysis

Is there any evidence that the model captures the same effects that the hand-mined feature model did?

Yes! In fact, most hand-mined feature effects barely survive, and some don’t.

Additional analyses

Attention weight analyses

Lee et al., 2017; Ding et al., 2017; Wiegreffe and Pinter, 2019

Evidence that the model learned to pay attention to a priori relevant utterance tokens (e.g., partitive “of”)

Minimal pair analyses

Linzen et al. 2016; Gulordava et al. 2018; Chowdhury and Zamparelli 2018; Marvin and Linzen 2018; Futrell et al. 2019; Wilcox et al. 2019

Evidence that the model can generalize what it learned to entirely new, artificial sentences

Sebastian Schuster

Elissa Li

Case study: “or”

Li, Schuster, & Degen, 2020

Examples

…but not both?

And I told my husband, I said, you know, it's either ***me or the dog.***

They always like to be able to attract ***the, uh, Einsteins or the Professor Chou****.*

So I began a program a couple, I don't know, probably ***three or four*** *weeks ago.*

But if you have a problem with ***what we did or how we did it****, you can always come back and talk to me about it.*

Methods

1. extracted all 1244 utterances containing *or* from the Switchboard corpus

2. crowd-sourced the position of “but not both”

3. collected inference strength ratings for each item on Mechanical Turk (9 judgments per item)

Speaker #2: . That 's something that , uh , people have seen . Oh , here 's an easy way - - to make some money . But , uh , I do n't know if that 's been challenged in the courts or not . I , I 've heard , fairly recently , uh , some talk about that in this , in , in my state . Uh , the budget problems up here are , are pretty tense . And people are looking for alternate ways of , uh , - en- , enhancing revenue is the , uh , phrase - they use . And they were talking about - selling the D M V lists and there was a lot of , uh , a lot of , uh , consternation about that and the last - I heard they 'd backed down from that idea . But it really makes you wonder what other lists you 're own that have been made , uh , public that you , do n't know about .
Speaker #1: .
Speaker #1: . Well , that 's easy . Whenever you donate money to someone .
Speaker #2: . Uh-huh .
Speaker #1: . They , you become , put on something like a sucker list and you start getting millions of calls or solicitations.

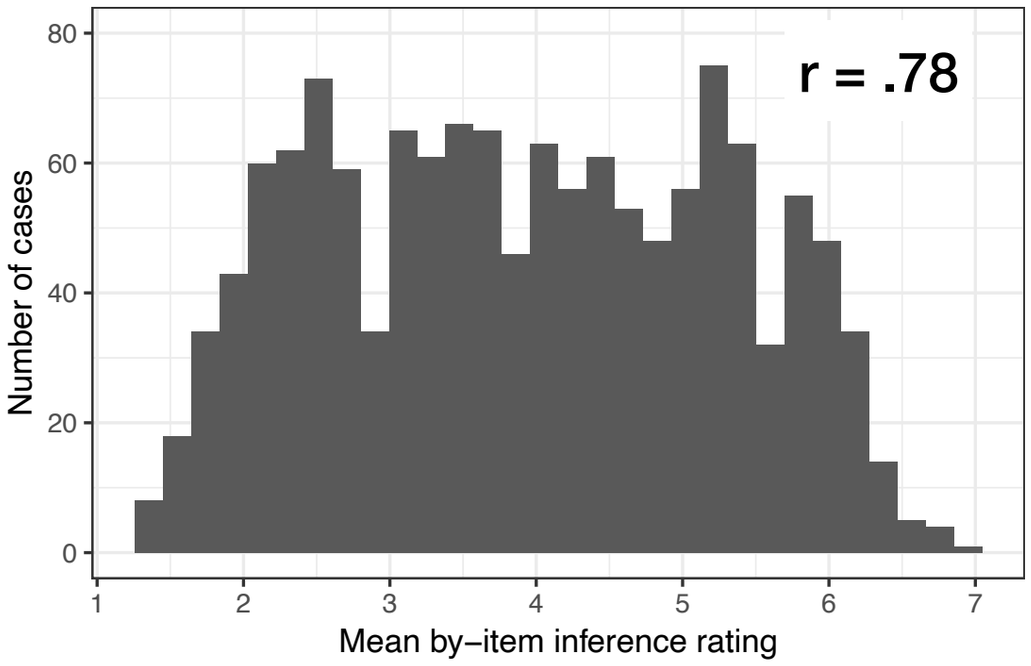
and you start getting millions of calls or solicitations **but not both**.

How similar is the statement in blue (with 'but not both') to the statement in red (without 'but not both')? Please adjust the slider.

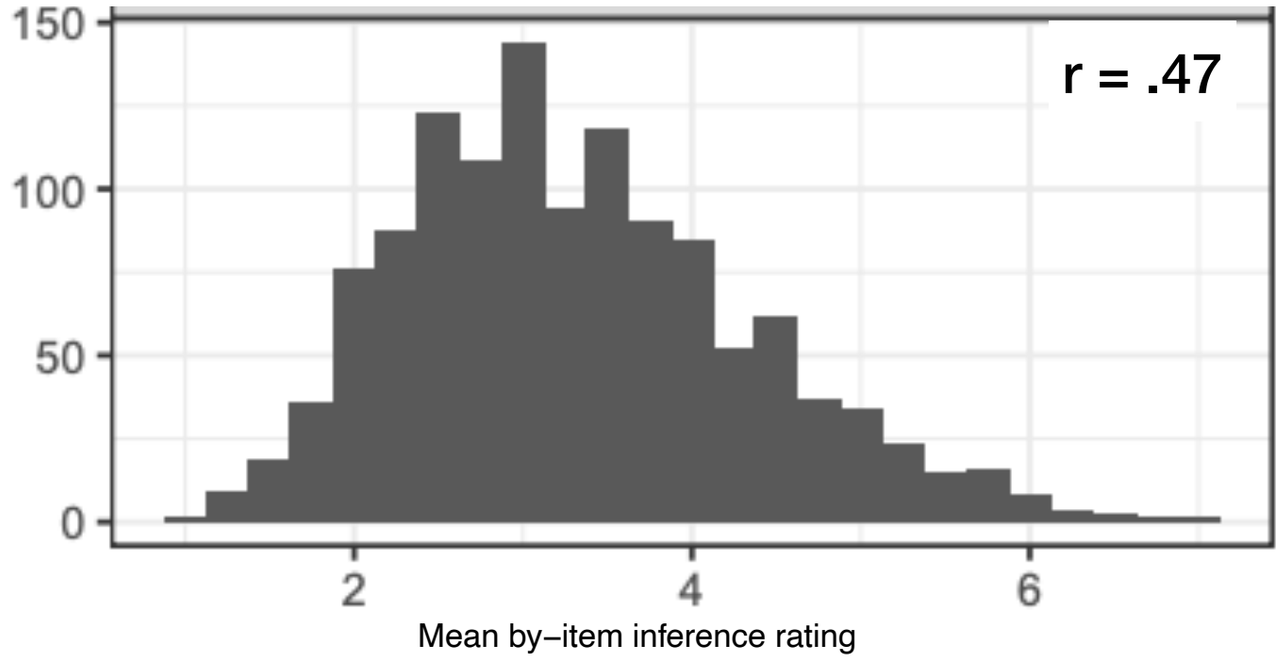
very different meaning same meaning

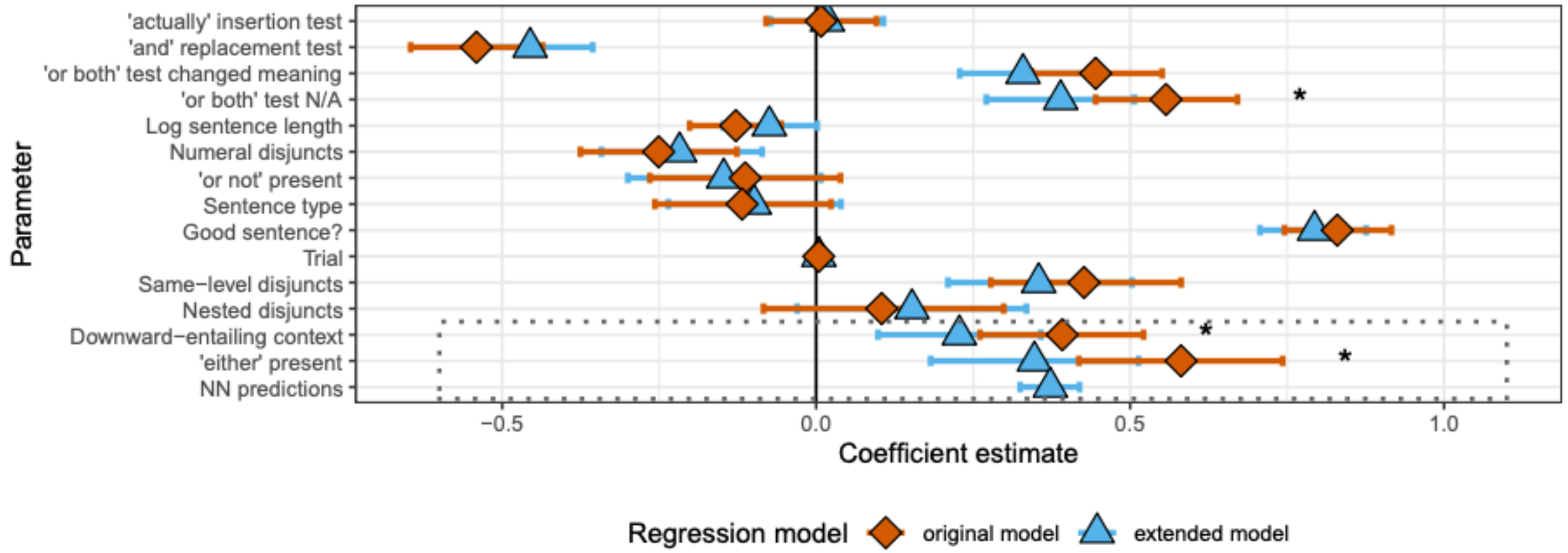
The blue sentence sounds strange: ☐

“some”



“or”





Neural model learned to predict “not both” inference strength ratings, but weaker performance than on “some” dataset

Txurruka and Asher, 2008; Potts & Levy 2015; Ariel and Mauri, 2018; Ariel, 2020

Conclusion

The focus on inter-scale variability may be premature, given the large amount of intra-scale variability in inference strength.

The surprisingly good performance of the neural models suggests that a lot of information about scalar inference is contained in the linguistic signal itself.

Interesting empirical questions:

1. How much pragmatic information is typically extracted from the linguistic signal itself vs from the extra-linguistic utterance context?
2. How big of an explanatory role will "the scale" retain once we better understand intra-scale variability?

Thank you!

Research assistants

Jane Boettcher
Pratyusha Javangula
Lexi Kuppersmith
Leyla Kursat
Cindy Torma
Andrew Watts
Neele Witte

Collaborators

Yuxing Chen
Elissa Li
Sebastian Schuster
Michael Franke

ALPS interActive Language Processing Lab Stanford