

# NLP: Neural Language Priming

Marten van Schijndel  
Department of Linguistics, Cornell University  
January 25 2022

# What are effective language processing representations?



We can explore the space using priming

# What is priming?

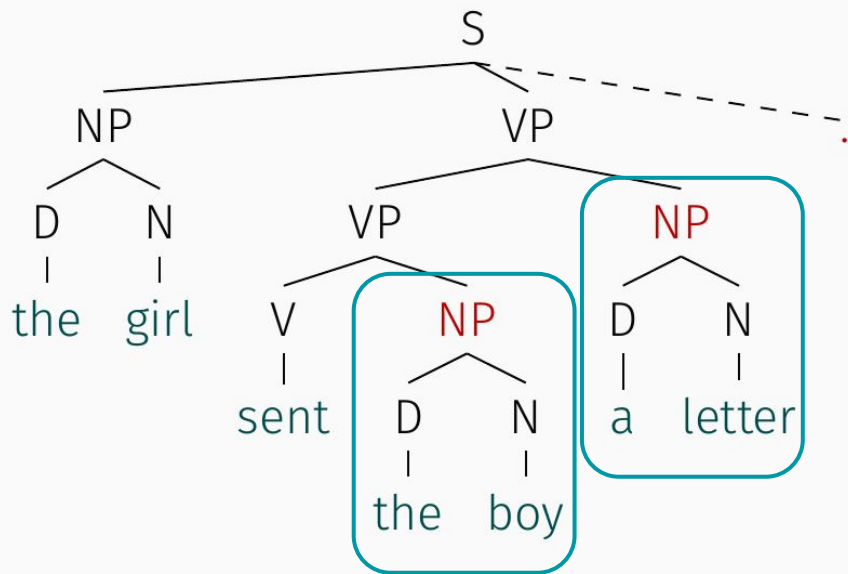
... cassowary ...

Cassowaries ...

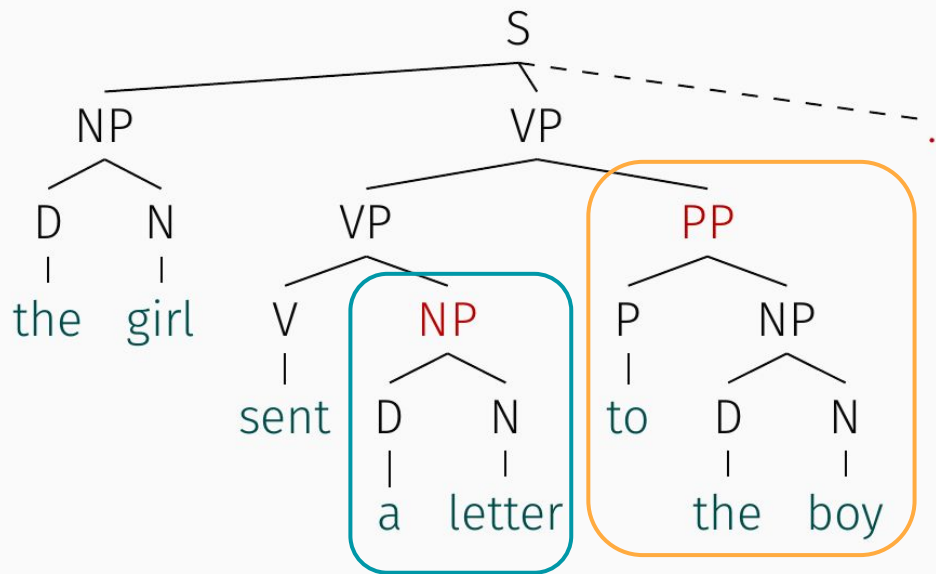
... Just like a cassowary!

... cassowary ...

# What is structural priming?



Double object



Prepositional object

# What is structural priming?

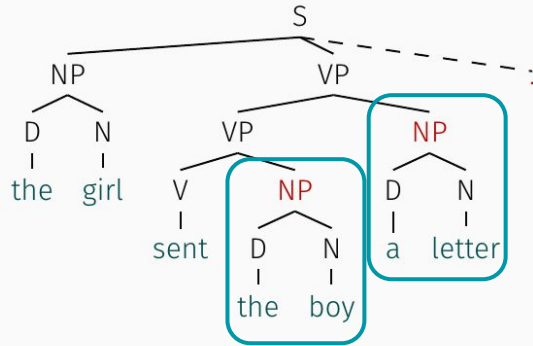
A rockstar sold some cocaine to the undercover agent.

...

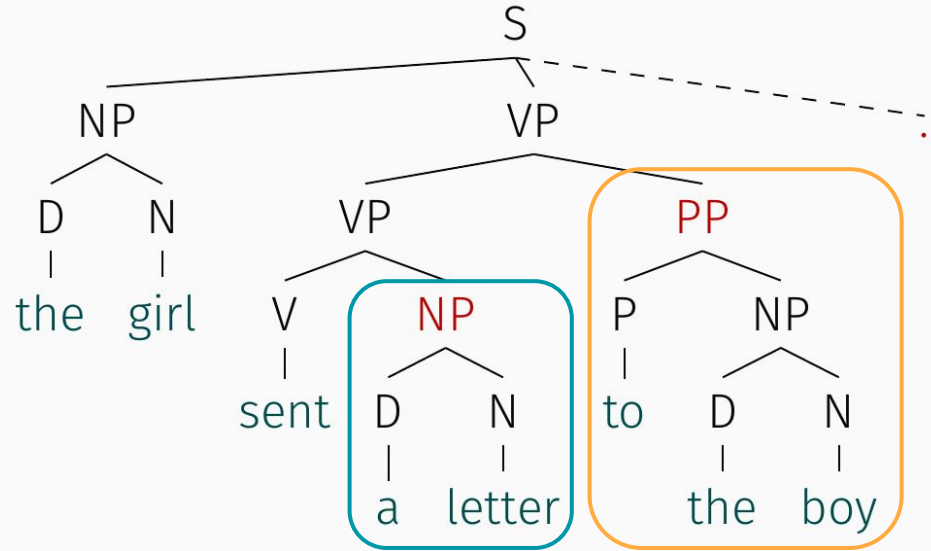
The girl sent a letter to the boy.

(Bock, 1986)

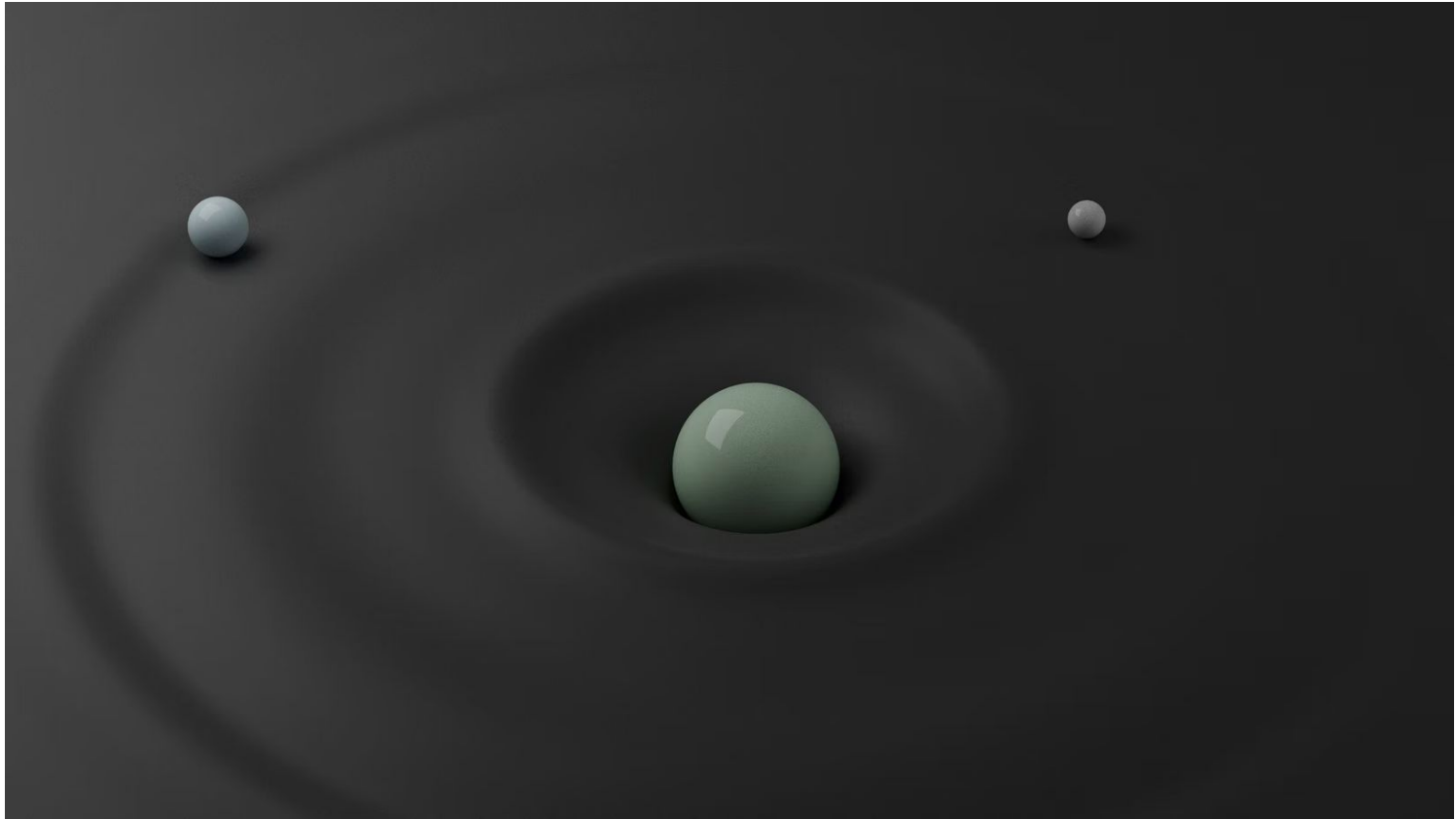
# What is structural priming?



Double object



Prepositional object



# Why expect this to work with NNs?

Neural networks operate over probability spaces

NNs require large amounts of data to train

Small amounts of training data should skew distributions according to sensitivity without wholly retraining them



# Neural language models can be primed



Tal Linzen

## **A Neural Model of Adaptation in Reading**

**Marten van Schijndel**  
Department of Linguistics  
Cornell University  
mv443@cornell.edu

**Tal Linzen**  
Department of Linguistics  
New York University  
linzen@nyu.edu

*Proceedings of EMNLP 2018*

# Adaptation priming algorithm

- 1) Test on a sentence
- 2) Update weights by fine-tuning on that sentence
- 3) Repeat on remaining sentences

The man named  
the child as his  
sole heir.

Wiki  
(1000)

DO  
(100)

The boy threw  
the dog a ball.

Adapt  
Set

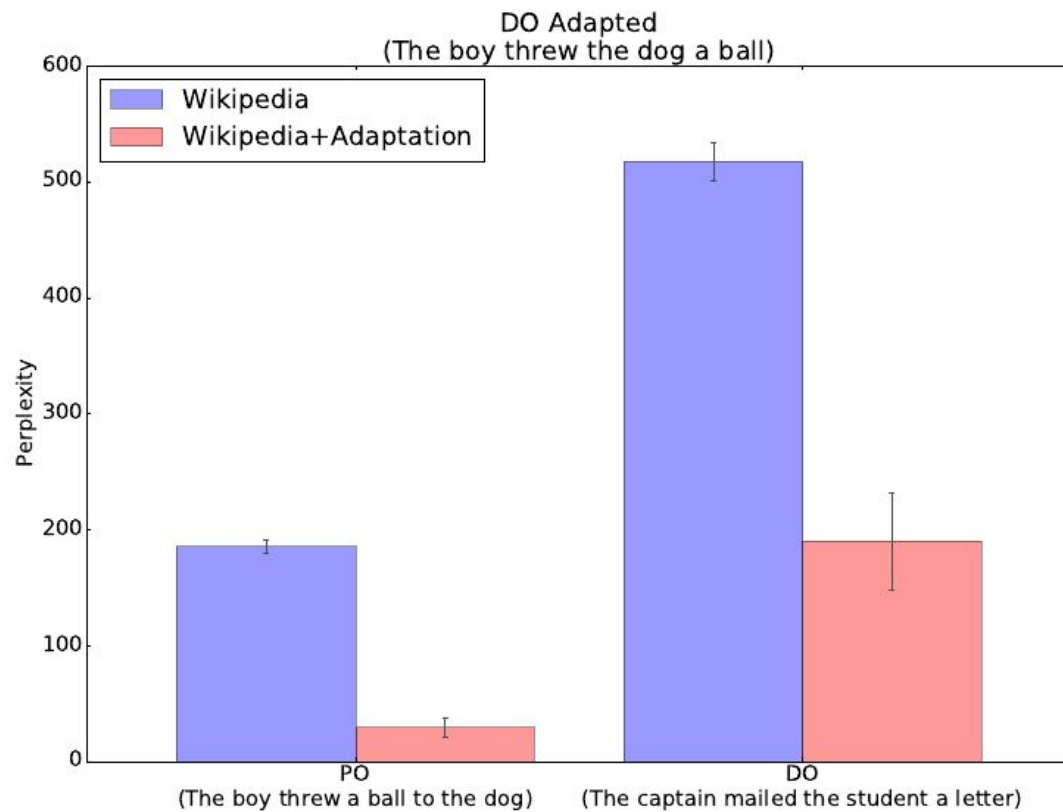
Language  
Model

The boy  
threw a ball to  
the dog.

PO  
(100)

DO  
(100)

The captain mailed  
the student a letter.



# Cumulative priming reveals underlying structure



Grusha Prasad



Tal Linzen

## Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models

**Grusha Prasad**  
Johns Hopkins University  
grusha.prasad@jhu.edu

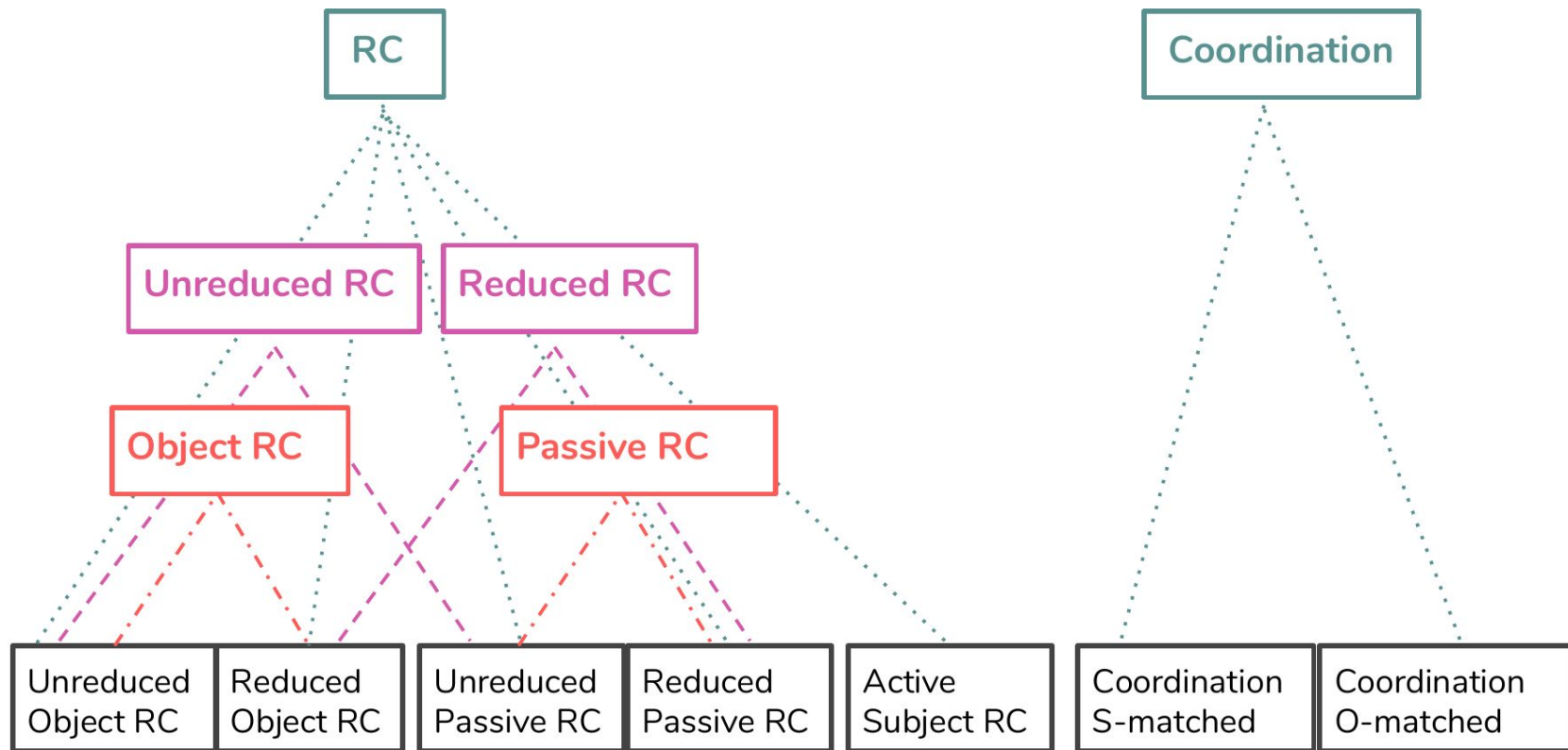
**Marten van Schijndel**  
Cornell University  
mv443@cornell.edu

**Tal Linzen**  
New York University  
linzen@nyu.edu

*Proceedings of CoNLL 2019*

# Cumulative priming reveals underlying structure

Abstract structure	Example
Unreduced Object RC	The conspiracy that the employee welcomed divided the beautiful country.
Reduced Object RC	The conspiracy the employee welcomed divided the beautiful country.
Unreduced Passive RC	The conspiracy that was welcomed by the employee divided the beautiful country.
Reduced Passive RC	The conspiracy welcomed by the employee divided the beautiful country.
Active Subject RC	The employee that welcomed the conspiracy quickly searched the buildings.
PS/ORC-matched Coordination	The conspiracy welcomed the employee and divided the beautiful country.
ASRC-matched Coordination	The employee welcomed the conspiracy and quickly searched the buildings.



# Priming reveals abstraction depth



Deb Bhattacharya

**Filler-gaps that neural networks fail to generalize**

**Debasmita Bhattacharya and Marten van Schijndel**

Department of Linguistics

Cornell University

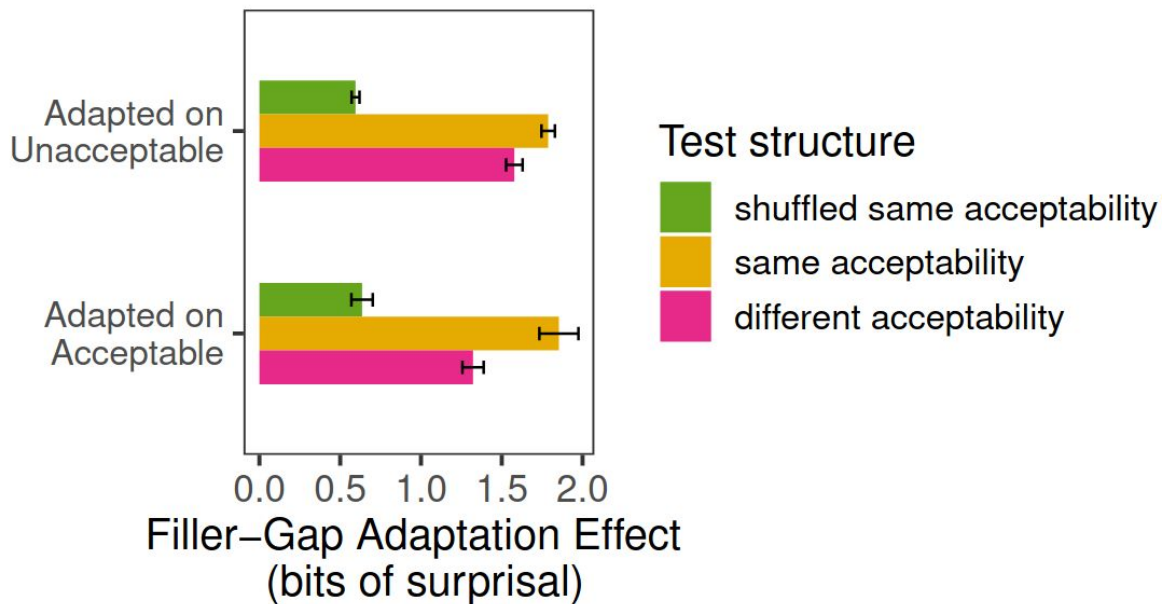
{db758 | mv443}@cornell.edu

*Proceedings of CoNLL 2020*

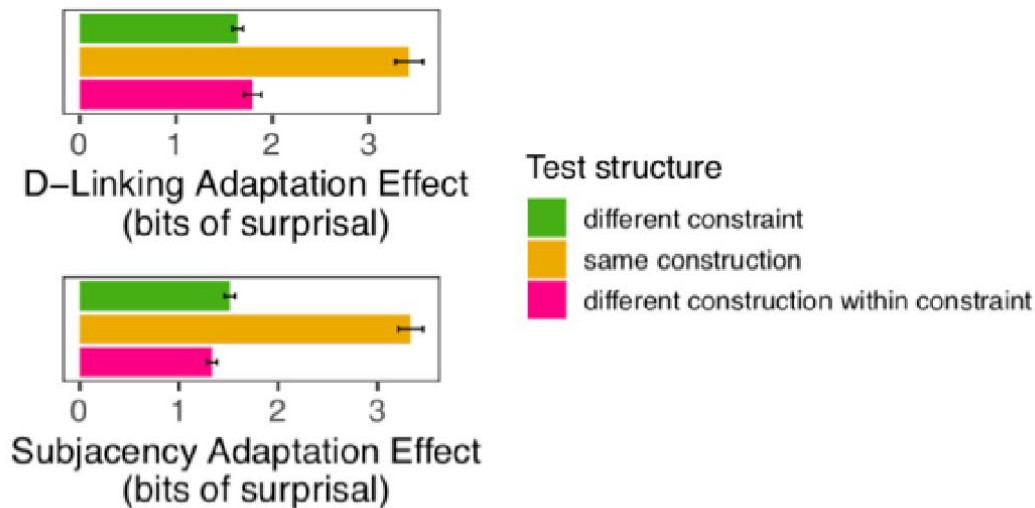


Construction	Examples	L	S	E	D
Adjunct island	*What did you go home because you needed to do __?	-		-	
Wh- island	*Whom did Susan ask why Sam was waiting for __?		-		(+)
Subject island	*Who is that __ went home likely?	-	-	-	
Left branch island	*Whose does Susan like __ account?		-		(+)
Coordinate structure island	*What did Sam eat __ and broccoli?			+	
Complex NP island	*What did you hear the claim that Fred solved __?		-		
Object extraction	Who is it probable that Bill likes __?	+		+	(+)
Non-bridge verb island	*How did she whisper that he had died __?	?	?		

# Networks encode broad FG existence



# Networks do not encode abstract FG constraints



# Priming reveals constraint-like rankings



Forrest Davis

## **Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning**

**Forrest Davis** and **Marten van Schijndel**

Department of Linguistics

Cornell University

`{fd252|mv443}@cornell.edu`

*Proceedings of ACL 2021*

(a) Sally frightened Mary because she was so terrifying.

**Who is terrifying?**

(a) Sally frightened Mary because she was so terrifying.



(a) Sally frightened Mary because she was so terrifying.



(a) Sally frightened Mary because she was so terrifying.



**Who is terrifying?**

**Sally**



(b) Sally feared Mary because she was so terrifying.

**Who is terrifying?**

(b) Sally feared Mary because she was so terrifying.



**Who is terrifying?**

Mary

# Implicit Causality

- Some verbs are **subject-biased**
  - *frightened, amused, astonished, enraged*
- Others are **object-biased**
  - *feared, admired, loved, hated*

# Is it attested cross-linguistically?

Yes

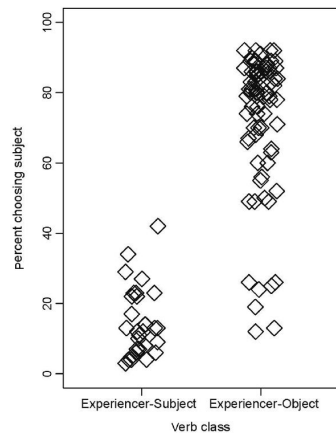


Figure 6. Degree of subject-bias for English emotion verbs reported in Ferstl, Granham, Manouilidou (in press).

**English**

**Dutch**

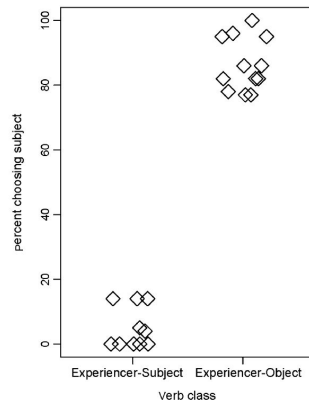


Figure 8. Degree of subject-bias for Dutch emotion verbs reported by Koornneef and van Berkum (2006).

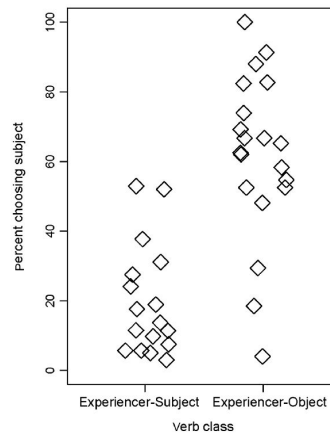


Figure 7. Degree of subject-bias for emotion verbs in Spanish as reported by Goikoetxea et al. (2008).

**Spanish**

**Italian**

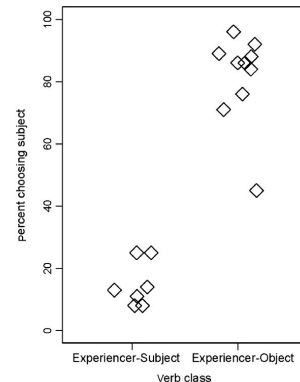


Figure 9. Degree of subject-bias for emotion verbs in Italian as reported by Mannetti and de Grada (1991).

from Hartshorne et al. (2013)

# English

# Evaluating English BERT for IC behavior

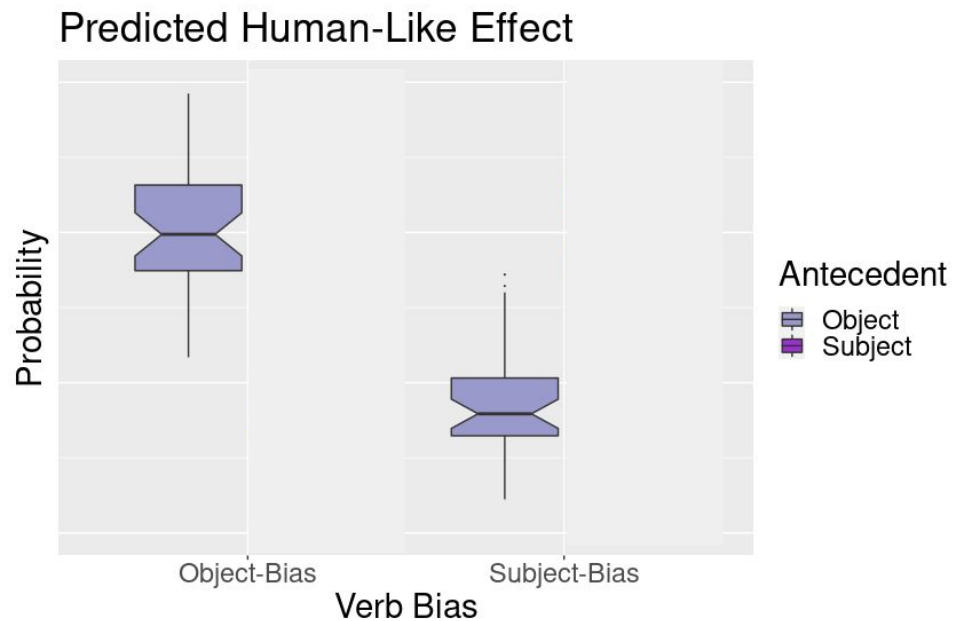
**the man admired the woman because [MASK] was there.**

- ~300 IC verbs from Ferstl et al. (2011)
- 14 pairs of stereotypical masculine and feminine nouns (e.g., king-queen) substituted in for subject and object (balanced for gender and position)
- bert-base-uncased (HuggingFace)
- Expect higher probability for pronouns agreeing with the bias  
(admire = object-biased)
  - $P(\text{she}) > P(\text{he})$

# Evaluating English BERT for IC behavior

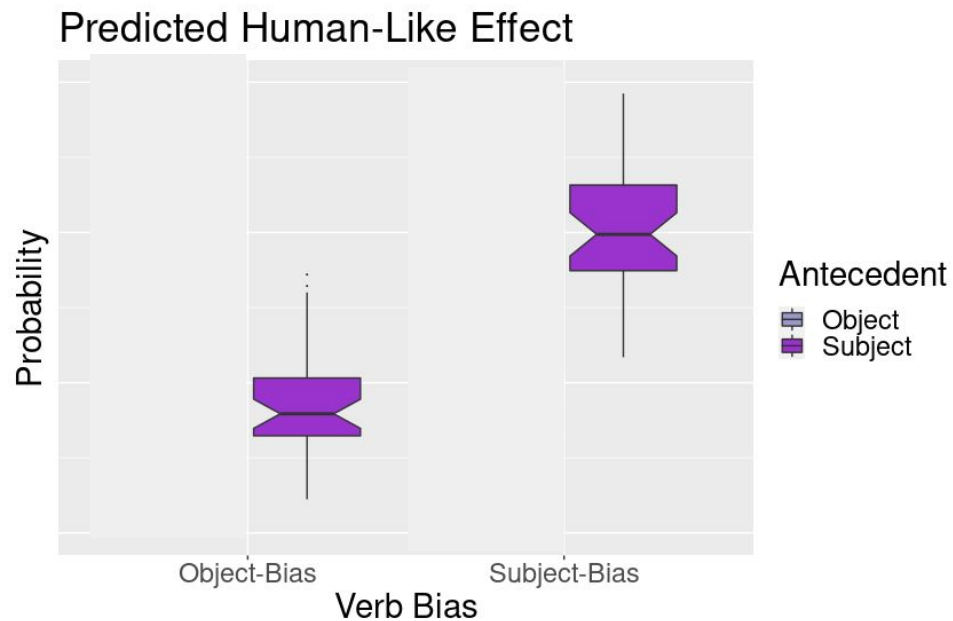


# Evaluating English BERT for IC behavior

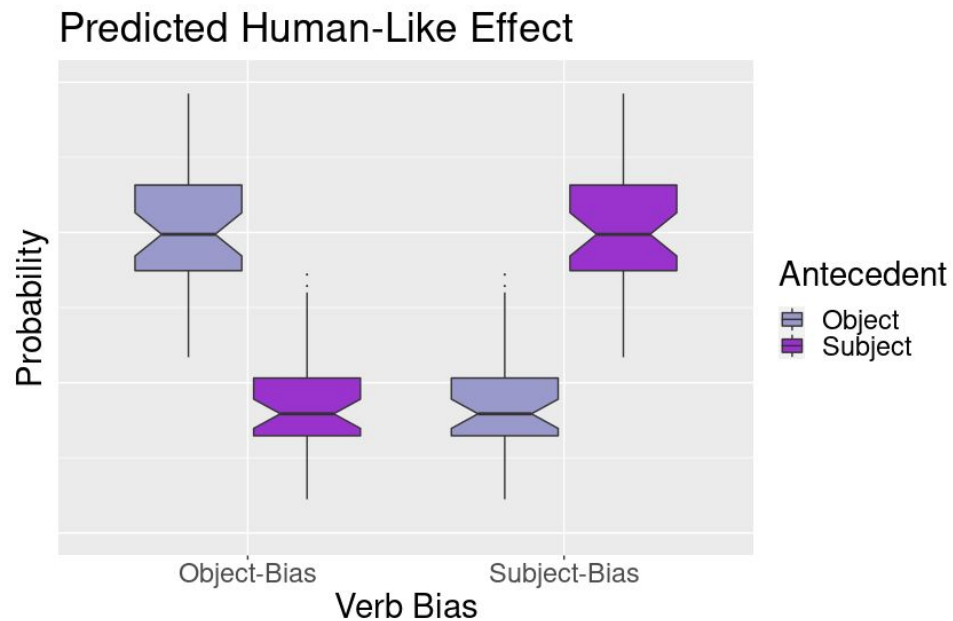




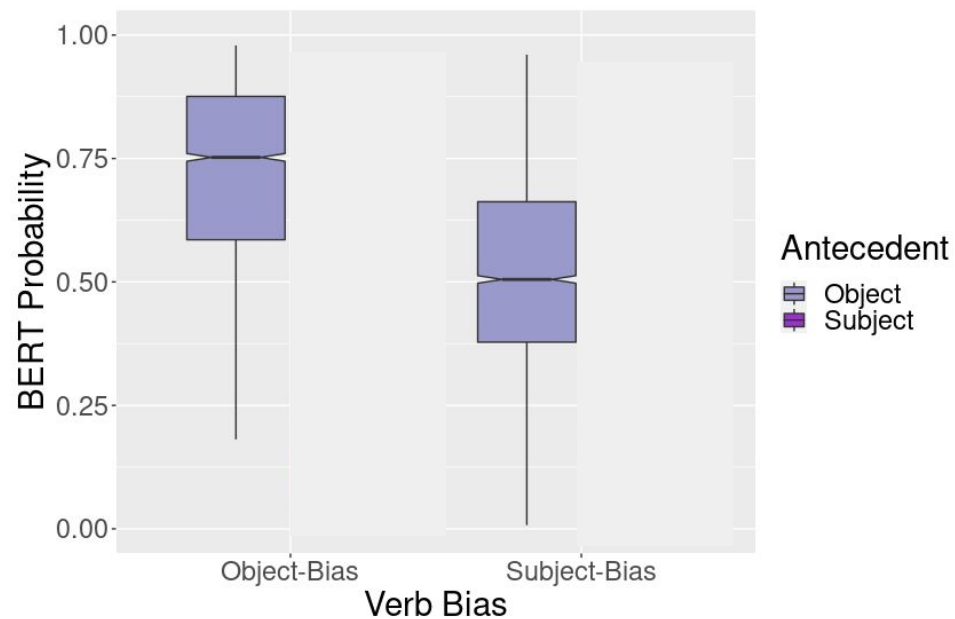
# Evaluating English BERT for IC behavior



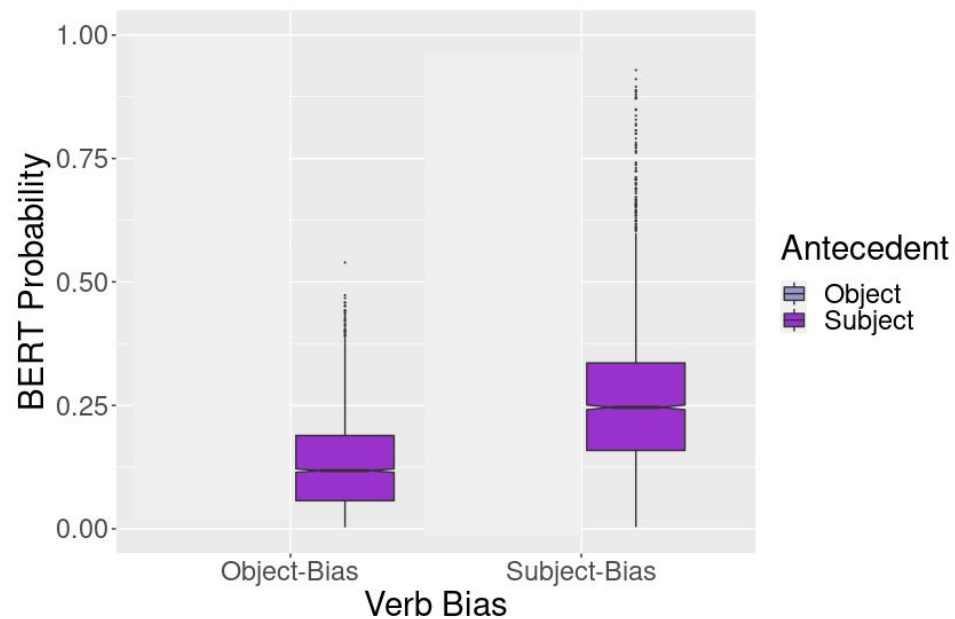
# Evaluating English BERT for IC behavior



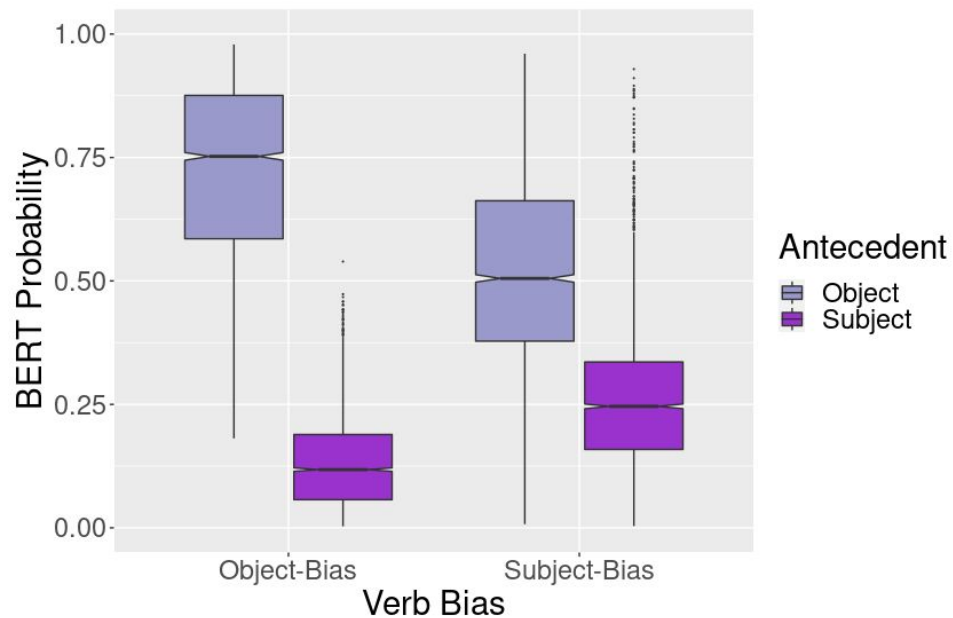
# English BERT shows IC behavior



# English BERT shows IC behavior



# English BERT shows IC behavior



# Spanish & Italian

# Evaluating Spanish BERT for IC behavior

el **hombre** admiró a la **mujer** porque [MASK] estaba allí.

- 100 IC verbs from Goikoetxea et al. (2008)
- 14 pairs of stereotypical m and f nouns subbed for subject and object  
(balanced for gender, position)
- dccuchile/bert-base-spanish-wwm-uncased (aka **BETO**; HuggingFace)
- Expect higher probability for pronouns agreeing with the bias  
(admiró = object-biased)
  - $P(\text{ella}) > P(\text{él})$

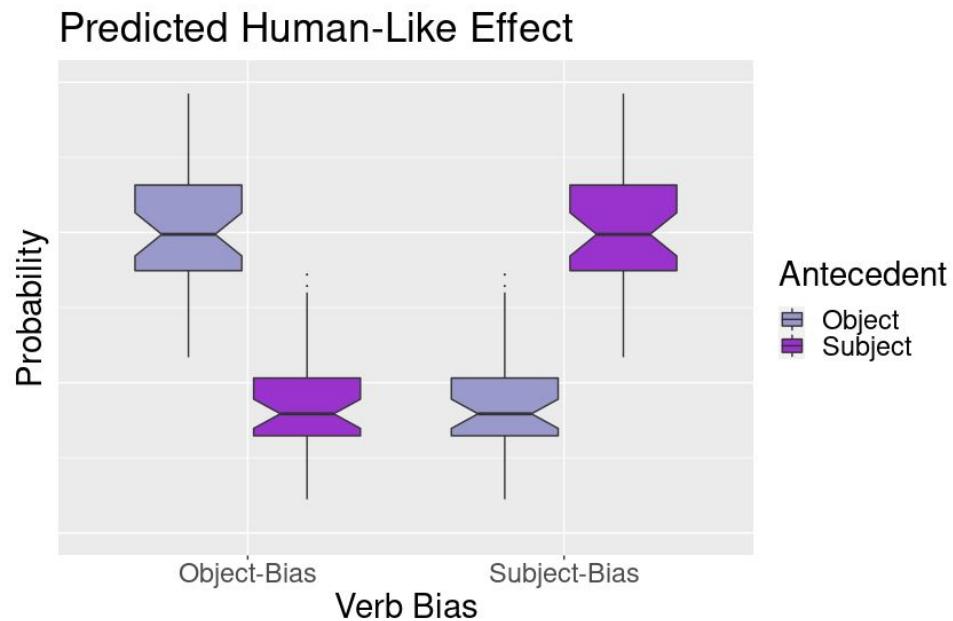
# Evaluating Italian BERT for IC behavior

**l'uomo lusinga la donna a causa del tipo di persona che [MASK] è.**

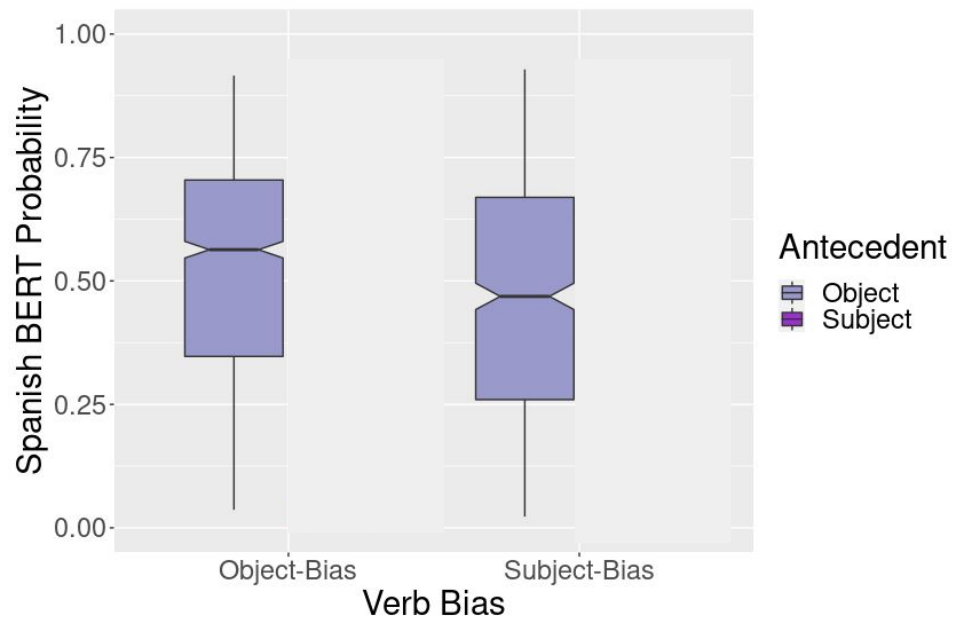
- 40 IC verbs from Mannetti and de Grada (1991)
- 14 pairs of stereotypical m and f nouns subbed for subject and object  
(balanced for gender, position)
- dbmdz/bert-base-italian-uncased (HuggingFace)
- Expect higher probability for pronouns agreeing with the bias  
(lusinga = object-biased)
  - $P(\text{lei}) > P(\text{lui})$



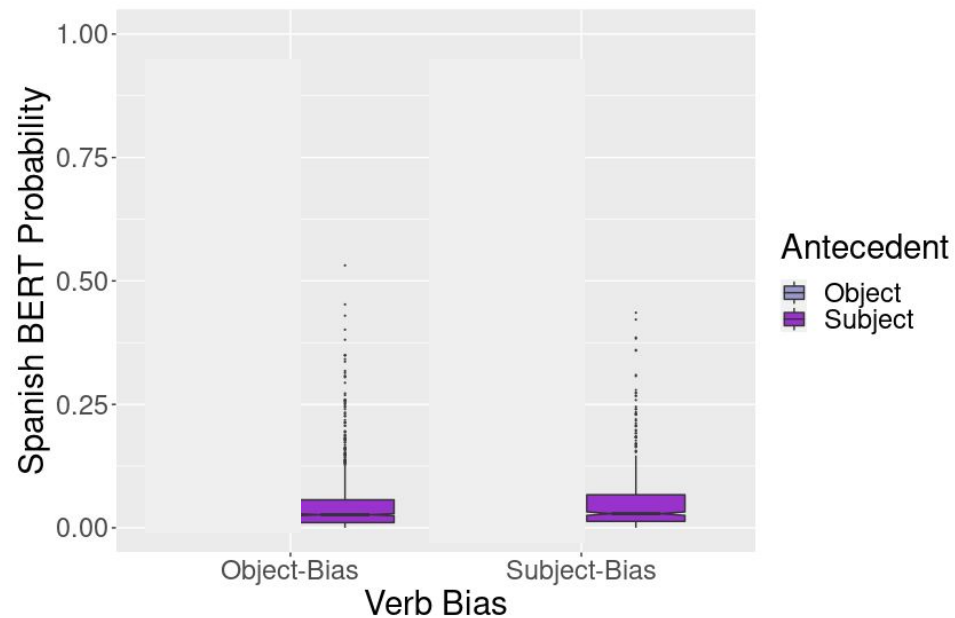
# Recall what we want to see



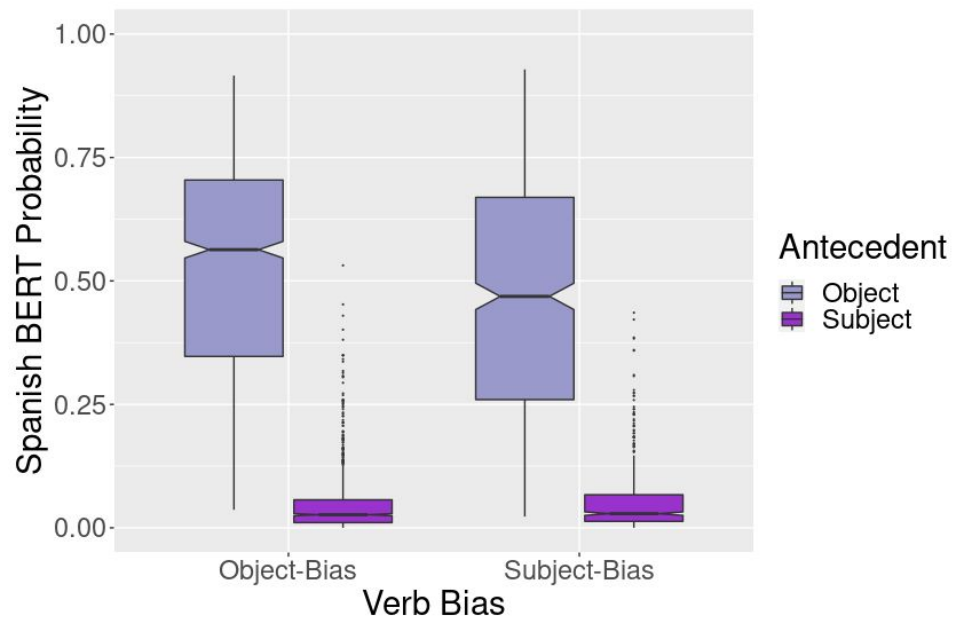
# Spanish BERT partial? IC behavior



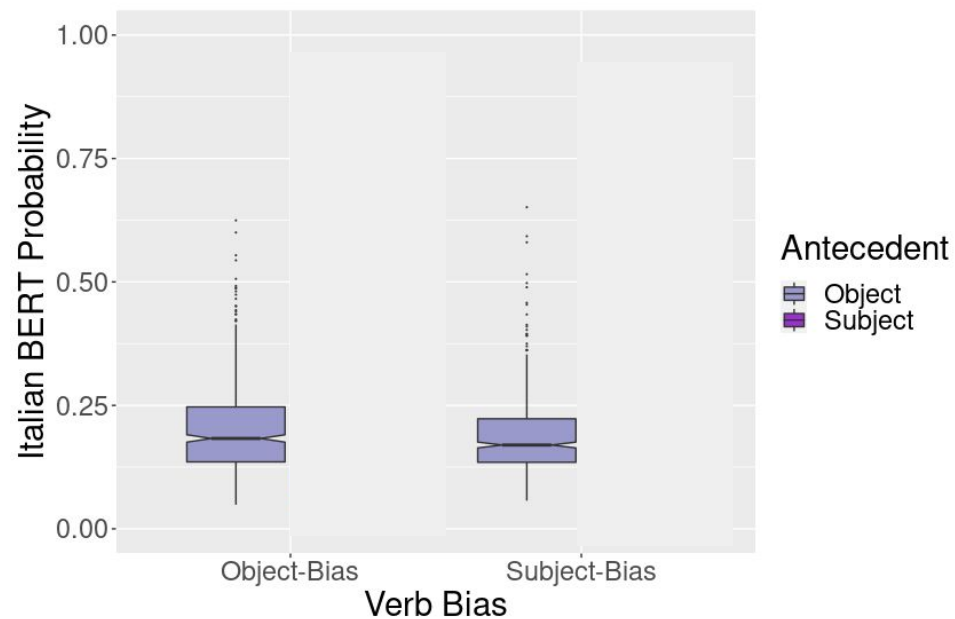
# Spanish BERT partial? IC behavior



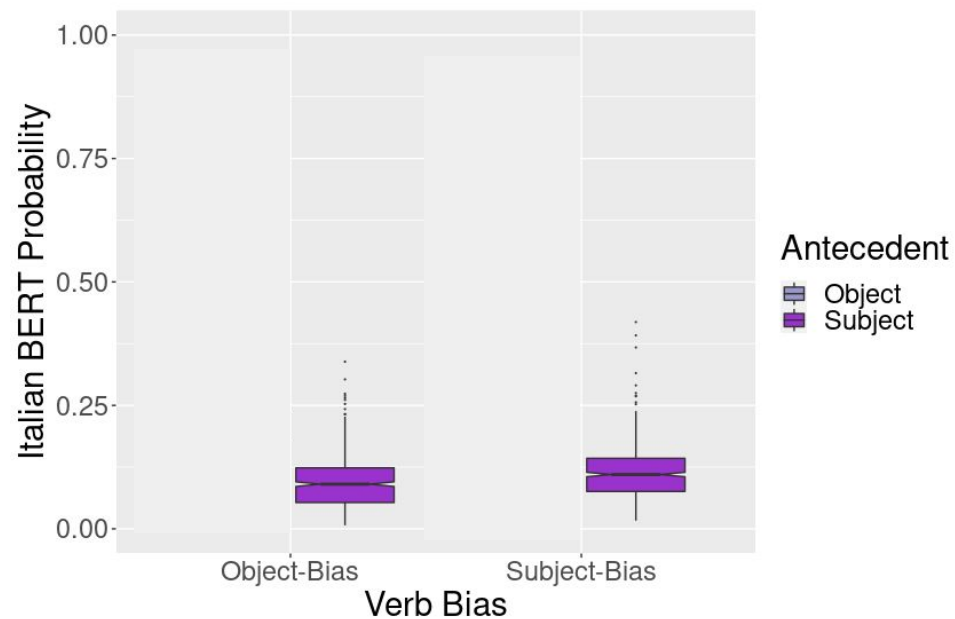
# Spanish BERT partial? IC behavior



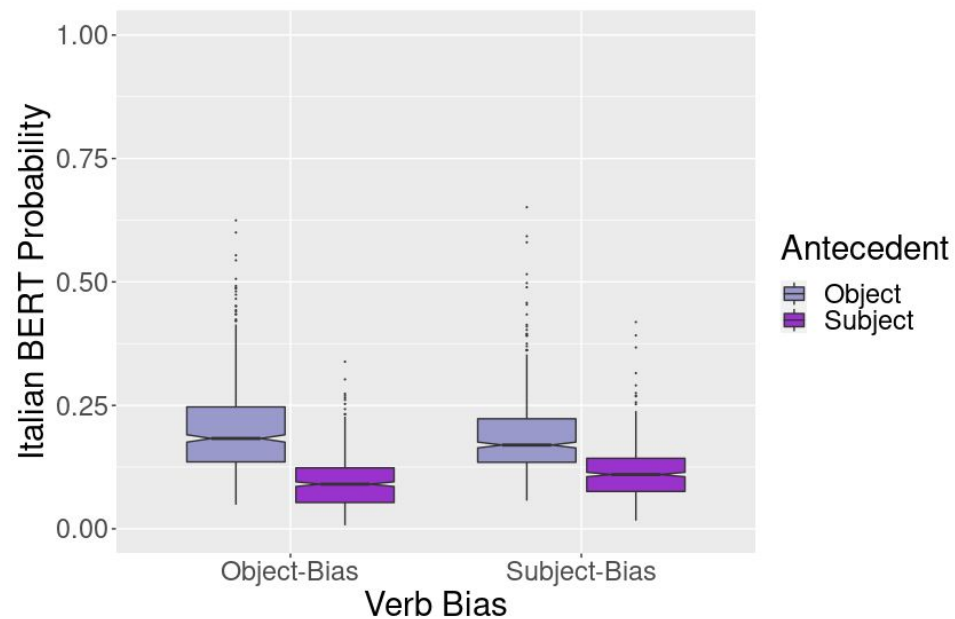
# Italian BERT no IC behavior



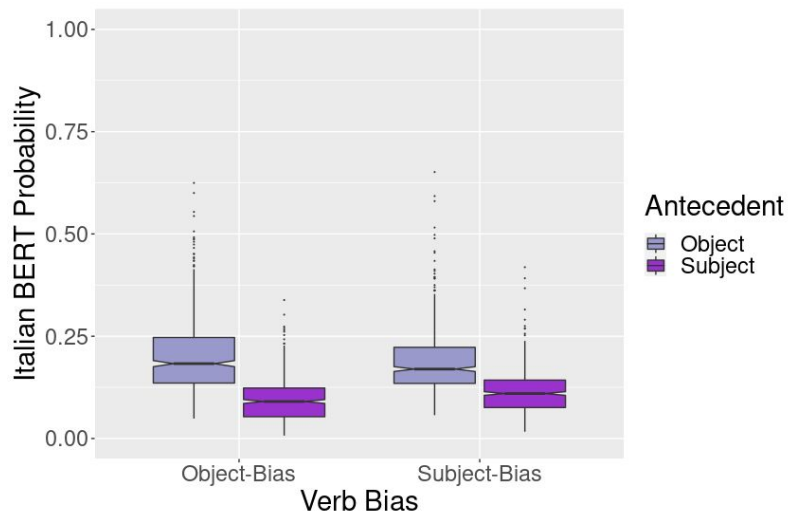
# Italian BERT no IC behavior



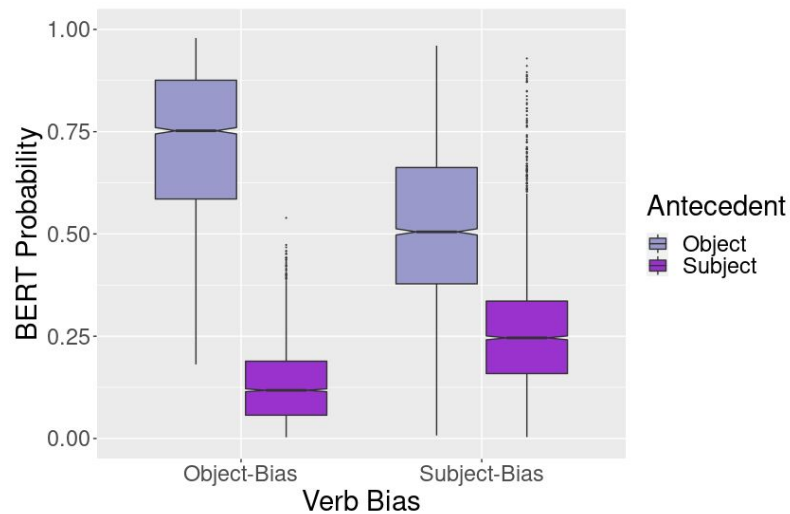
# Italian BERT no IC behavior



# Italian BERT no IC behavior



**Italian**

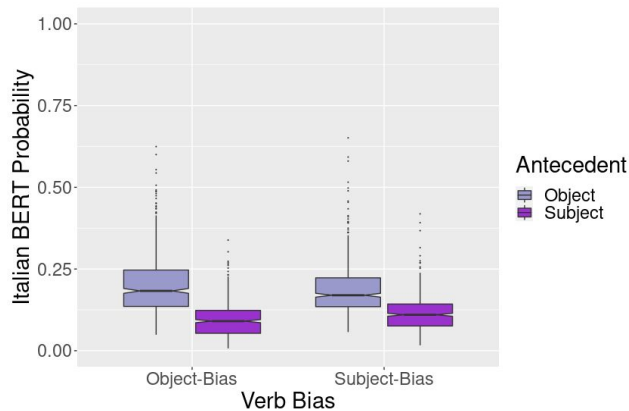


**English**

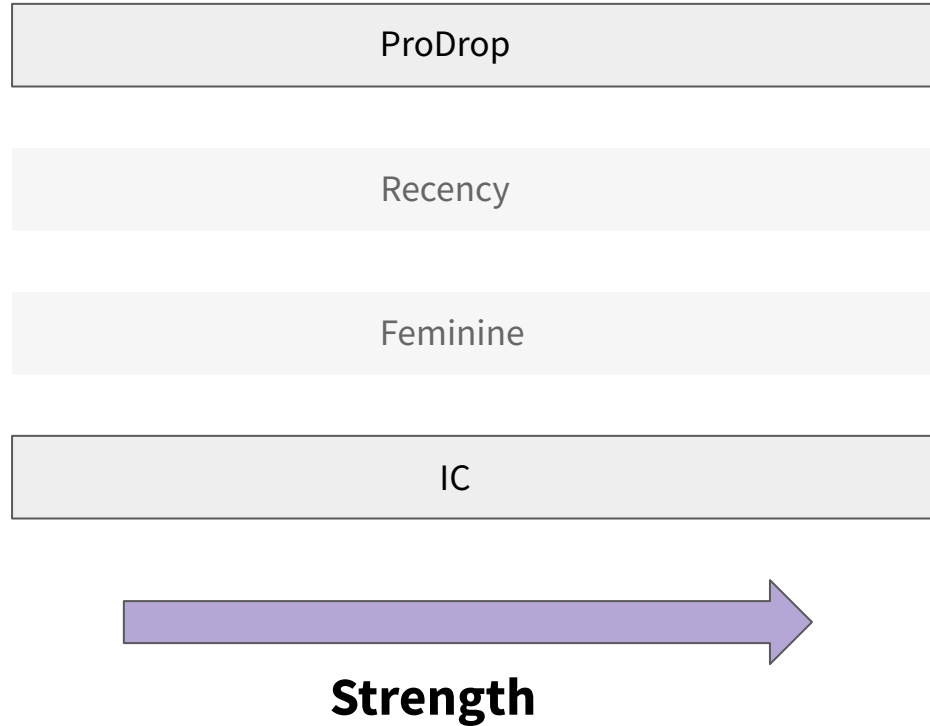


# Why aren't we learning IC in Spanish & Italian

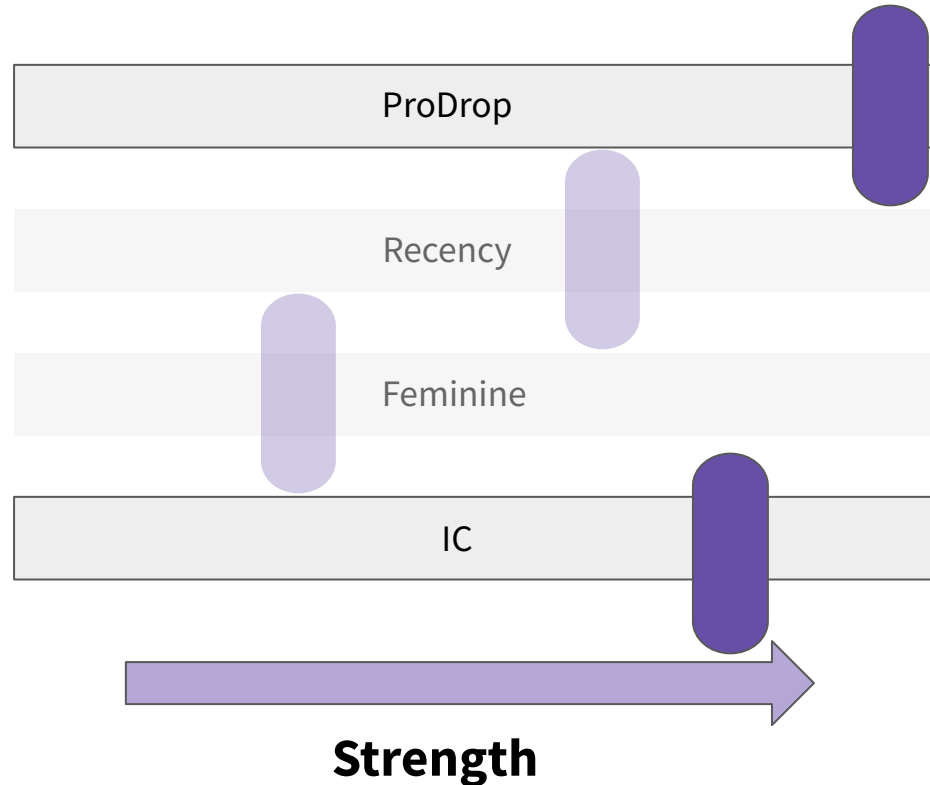
- Implicit Causality is not the only process affecting pronouns
- ProDrop: Overt pronouns in subject position are ungrammatical or dispreferred
  - is happy vs. \*he is happy



# Processes as Constraints



# Processes as Constraints



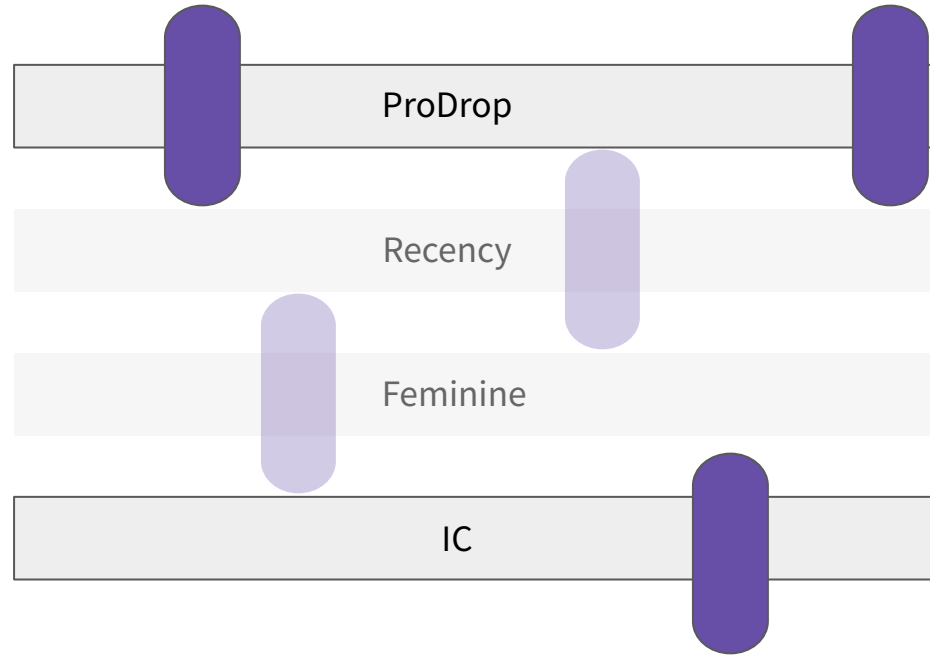
**The man  
admired  
the woman  
because \_\_  
was there**

# ProDrop Competes with IC bias

ProDrop outranks IC

Can we promote IC by demoting ProDrop?

# Processes as Constraints



**The man  
admired  
the woman  
because she  
was there**

# Demoting ProDrop with Fine-tuning

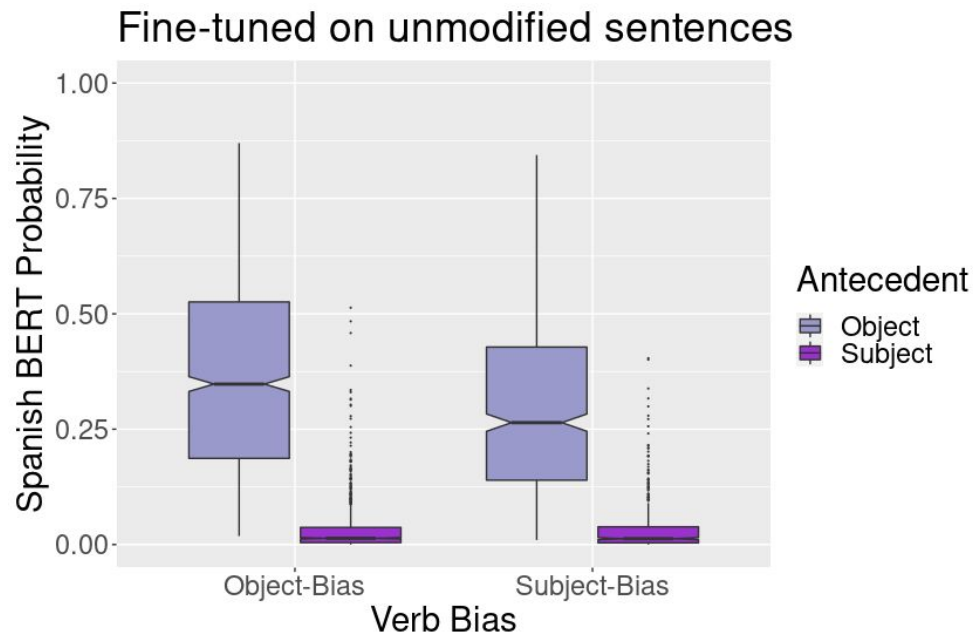
Method: Add pronouns to Spanish/Italian ProDrop sentences

UD datasets (AnCorra Spanish, Italian ISDT and VIT)

- Find finite verbs w/o nsubj relation, filtering out test IC verbs
- Added a pronoun matching the verb's person and number
- Resulted in 4000 Spanish sentences (~3500 he/she)  
4600 Italian sentences (~2000 he/she);  
**0.005%** of original training data
- Fine-tuned Spanish and Italian BERT for 3 epochs (lr=5e-5)

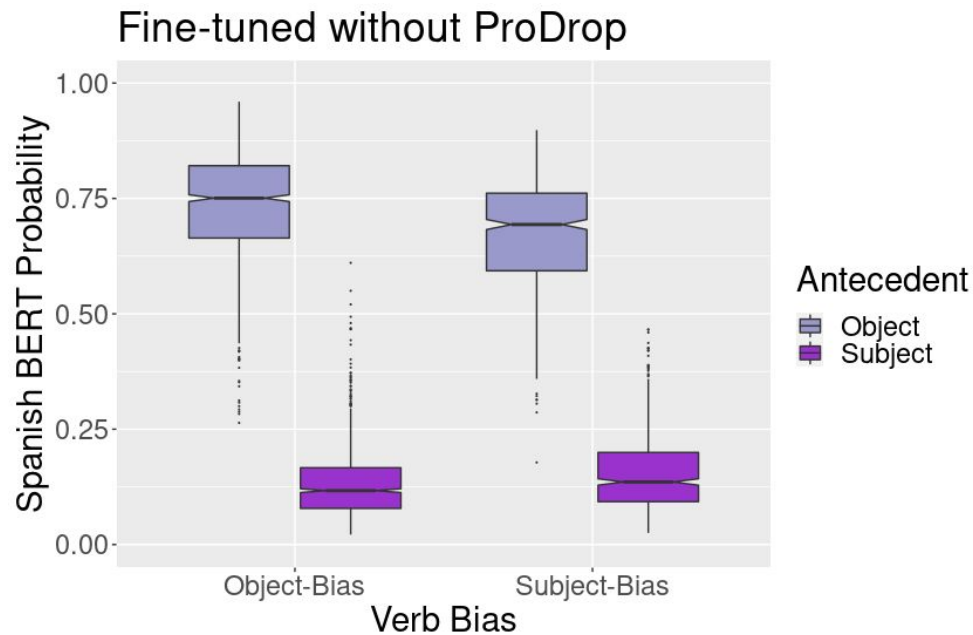
# Spanish Baseline

- Unmodified sentences



# Spanish Fine-tuned

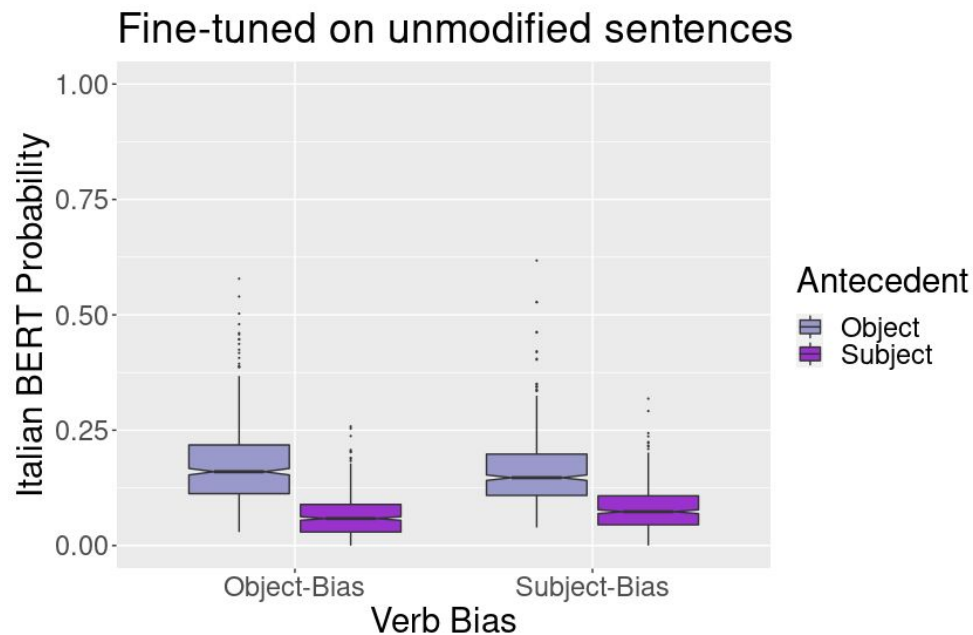
- Sentences without ProDrop





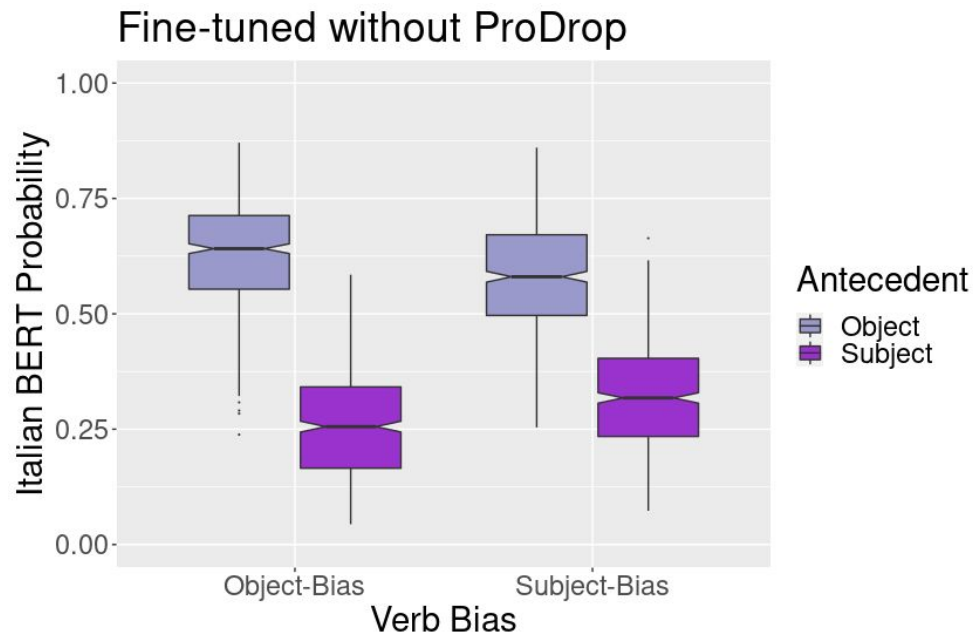
# Italian Baseline

- Unmodified sentences



# Italian Fine-tuned

- Sentences without ProDrop



# Summary

- Standard behavioral probing: IC is only learned by **English** models
- In fact, Spanish and Italian models also learn IC but it conflicts with ProDrop
- **Removing ProDrop** via fine-tuning on a small number of ProDrop violations **revealed IC knowledge**
  - We explicitly train on data for a process we are **not** testing (i.e. we are not training on IC data)

# Conclusion

- Adaptation priming is effective for:

Better text  
predictions

Better human  
predictions

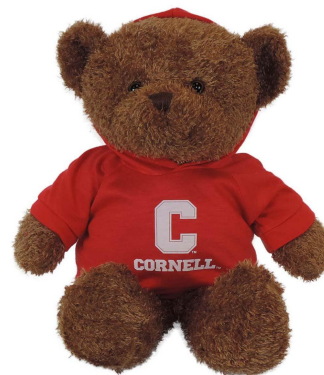
Probing  
representations

- We can probe **feature clusters** and **constraint/feature rankings**
- Field too often probes **single phenomena** (usually in English)  
need to watch for **interactions**  
(signals compete)

# Thanks!



C.Psyd



Cornell NLP