Do RNNs learn human-like abstract worder order preferences?

David Strohmaier

University of Cambridge

2019

Paper

Title:	Do RNNs learn human-like abstract worder order preferences?
Authors:	Richard Futrell & Roger P. Levy
Year:	2019
Journal:	*Proceedings of the Society for Computation in Linguistics*
DOI:	`https://doi.org/10.7275/jb34-9986`

The papers investigates whether RNN language models learn human-like worder order preferences.
All the graphs in the following slides are taken from the original paper.

More on the Task

- Some word order preferences are straightforward (e.g. subject - verb - object).
- The authors are interested in word order preferences that are more abstract. They have to be stated in terms of higher-order linguistic units and abstract features.
- These preferences are (mostly) soft constraints.

Syntactic Phenomena

Alternations and shifts associated with word order preference:

1. heavy NP shift
2. particle shift
3. dative alternation
4. genitive alternation

Shared features:

- short constituents before long
- definite words go earlier
- words referring to animate entities go earlier

Language Models

model	type	corpus size	hidden layers	units
JRNN	LSTM+CNN input	1 billion	2	8196
GRNN	LSTM	90 million	2	650
5-gram	n-gram	1 billion	-	-

Same language models as other papers in the literature.

Metric: Surprisal

A sentence's surprisal equals its contribution to a language model's cross-entropy loss.

$$\begin{aligned} S(x_{i=1}^{n}) &= -\log_2 p(x_{i=1}^{n}) \\ &= -\sum_{i=1}^{n} \log_2 p(x_i | x_{j=1}^{i-j}) \end{aligned} \tag{1}$$

Thus, closely related to the perplexity:

$$2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)} \tag{2}$$

The surprisal sould reflect the dispreference for a sequence according to the language model.

Human Acceptability Ratings

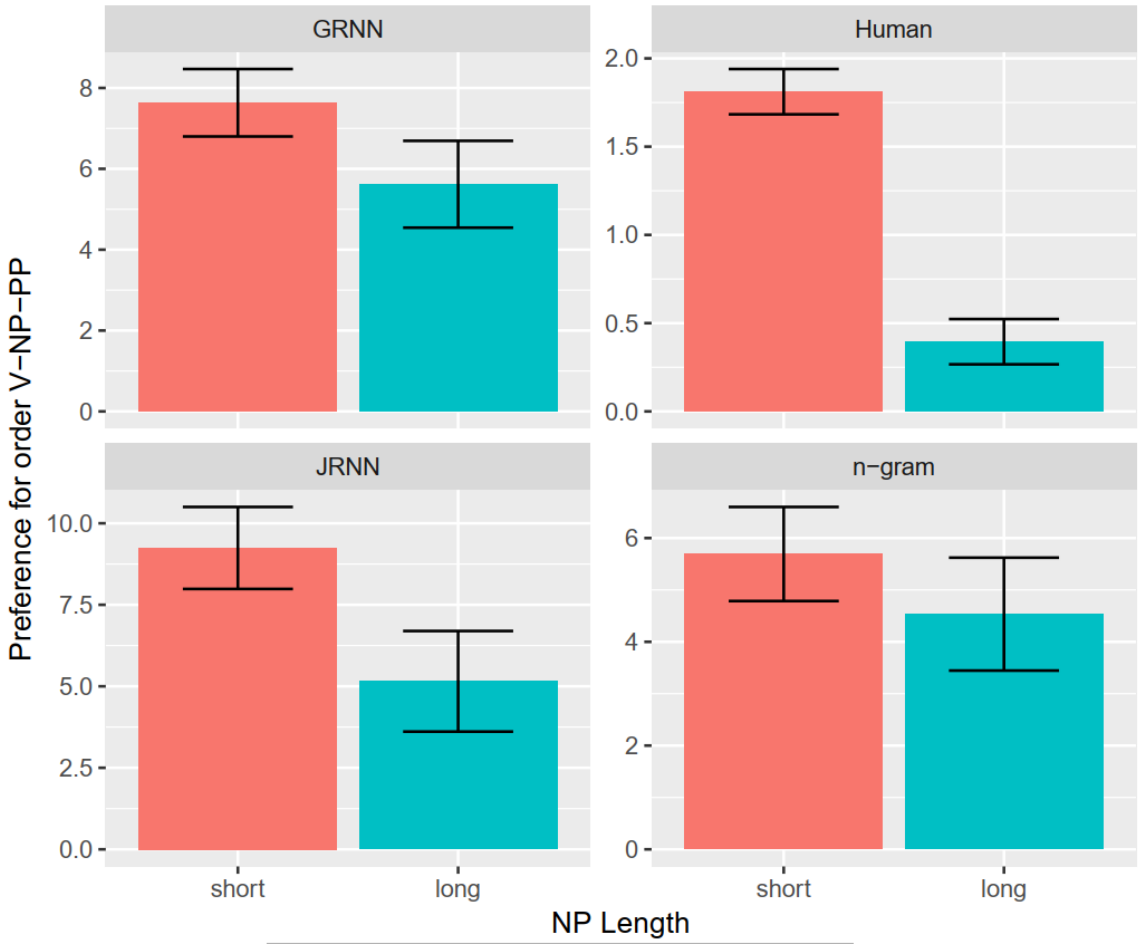
- Using *Amazon Mechanical Turk*
- Number of filtered participants: 156 (using previous data)
- Scale from 1 (least acceptable) to 5 (most acceptable)
- Is the scale ordinal?

Heavy NP Shift

Usual word order: Verb-NP-PP
However if the NP is very long ("heavy"), then Verb-PP-NP becomes acceptable
Example:

1. The publisher announced a book on Thursday.
2. *The publisher announced on Thursday a book.
3. The publisher announced a new book from a famous author who always produced bestsellers on Thursday.
4. The publisher announced on Thursday a new book from a famous author who always produced bestsellers.

Results: Heavy NP Shift



What is preference?

Preference

> *"[p]reference is measured as total sentence surprisal for Verb-NP-PP order minus total sentence surprisal for Verb-PP-NP order"*

No, it is not.

- Surprisal, roughly, measures general dispreference for a sentence.
- Subtracting the average dispreference for the Verb-PP-NP order from the dispreference for Verb-NP-PP, we get how much more generally dispreferred it is.
- The specific preference is the additive inverse of that.

Preference

- I calculated it for the JRNN model according to their description:
- (Dis-)Preference Short: -5.208
- (Dis-)Preference Long : -9.243
- The additive inverse is what we see in the figure.
- Great that they provided data!

For human data: Difference in mean acceptability.

Results: Heavy NP Shift

Their equation:

$$\begin{aligned} I_i = &(S_i(\text{short, Verb-NP-PP}) - S_i(\text{short, Verb-PP-NP})) \\ &- (S_i(\text{long, Verb-NP-PP}) - S_i(\text{long, Verb-PP-NP})) \end{aligned} \tag{3}$$

Again confusion about the order!

JRNN and GRNN show the same effect as the human acceptability judgement. The n-gram model does not show the same word-order bias.

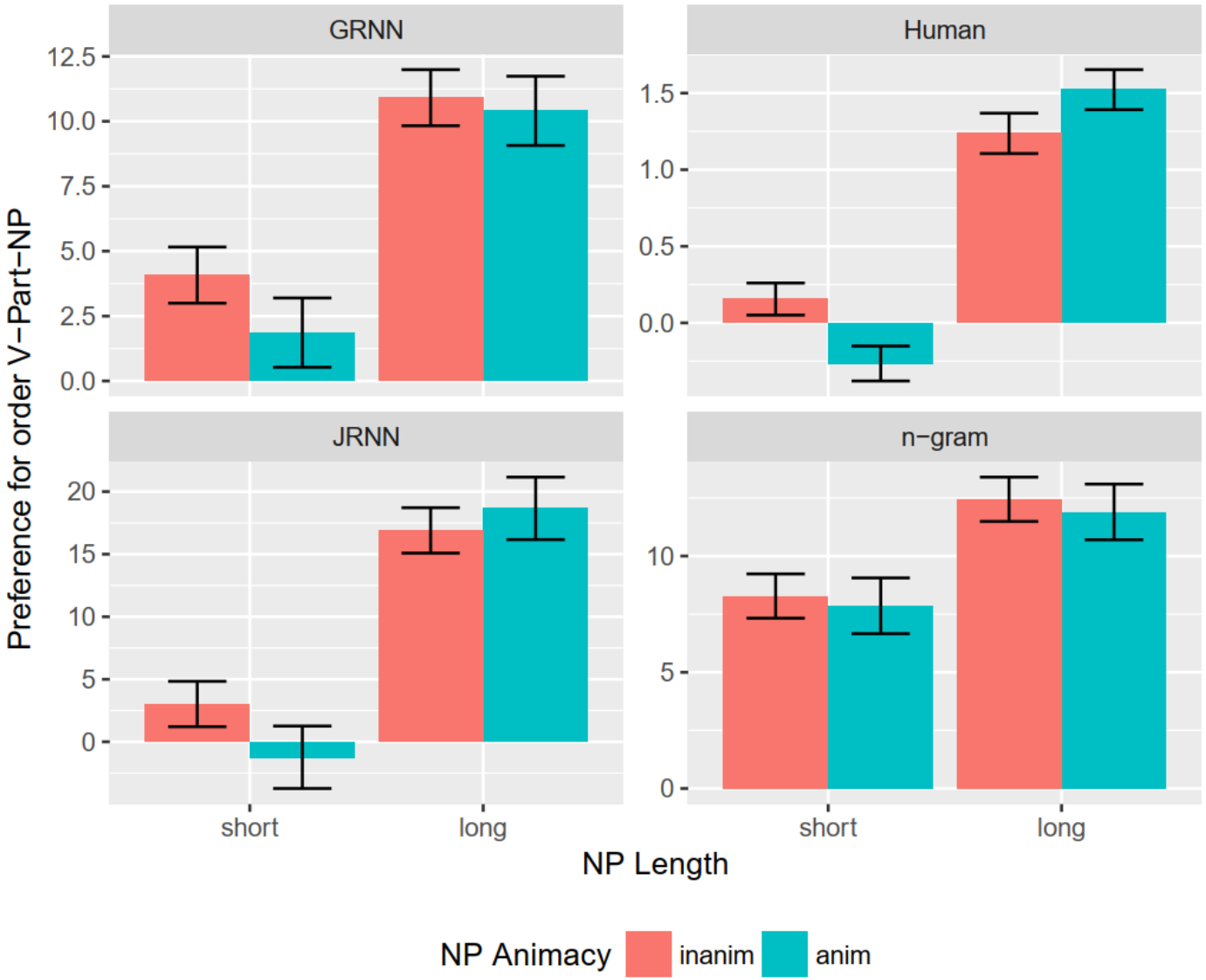
Phrasal Verbs and Particle Shift

The phrasal verbs consist of a verb and a particle. The object NP can appear right after the particle (shifted) or before it (unshifted). Shifted order is generally preferred when the NP is long.
Example:

1. Kim gave up the habit. [shifted]
2. Kim gave the habit up. [unshifted]
3. Kim gave up the habit that was preventing success in the workplace. [shifted]
4. Kim gave the habit that was preventing success in the workplace up. [unshifted]

Whether the NP object picks out an animate or an inanimate object matters as well.

Results: Phrasal Verb Shift



Results: Phrasal Verb Shift

- JRNN, GRNN, and n-gram model show the same direction of the effect as the human acceptability judgement for length (significant).
- The experiments do not find that NP animacy makes a significant difference, not even for human acceptability judgement.

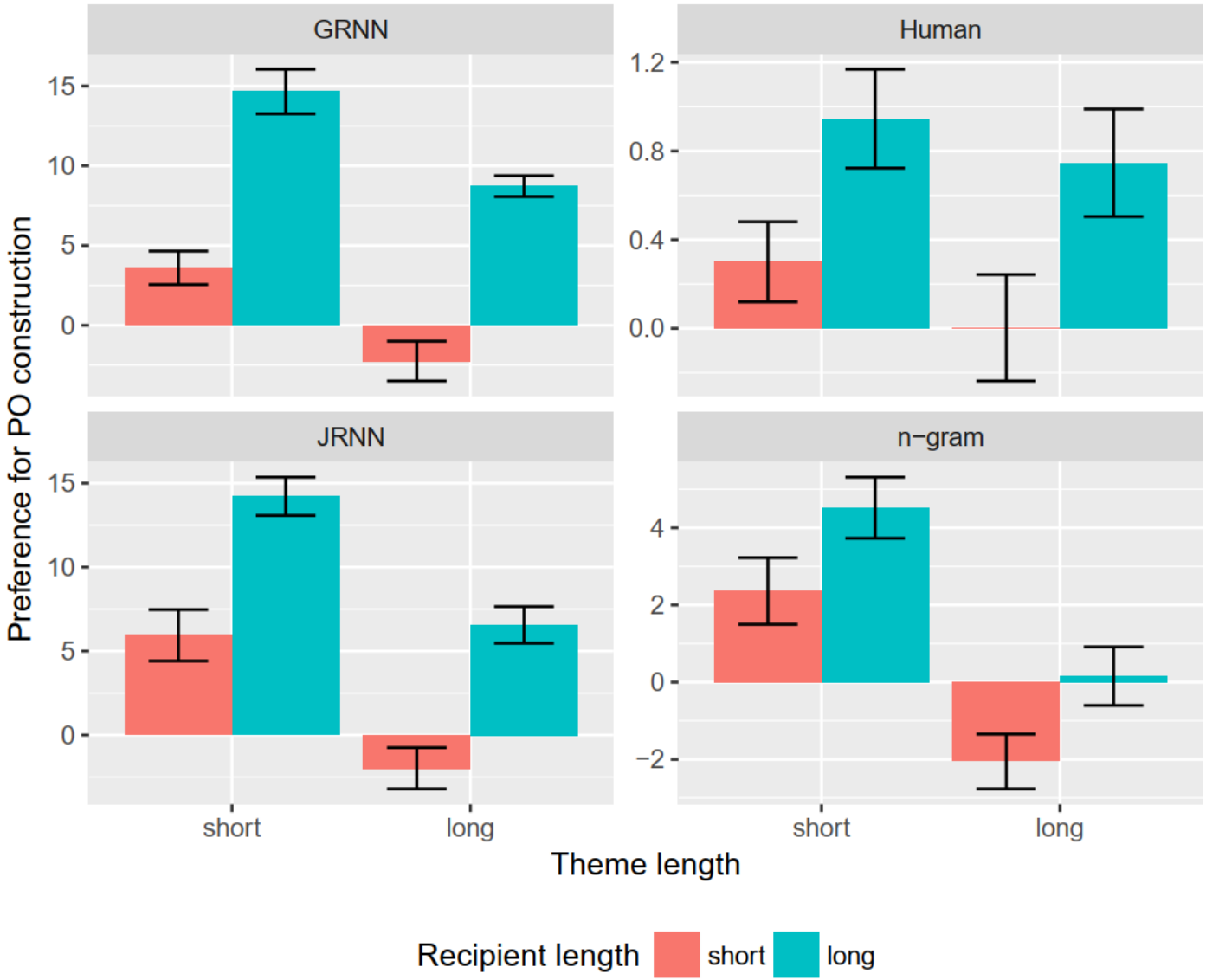
Dative Alternation

Ordering of theme and recipient.

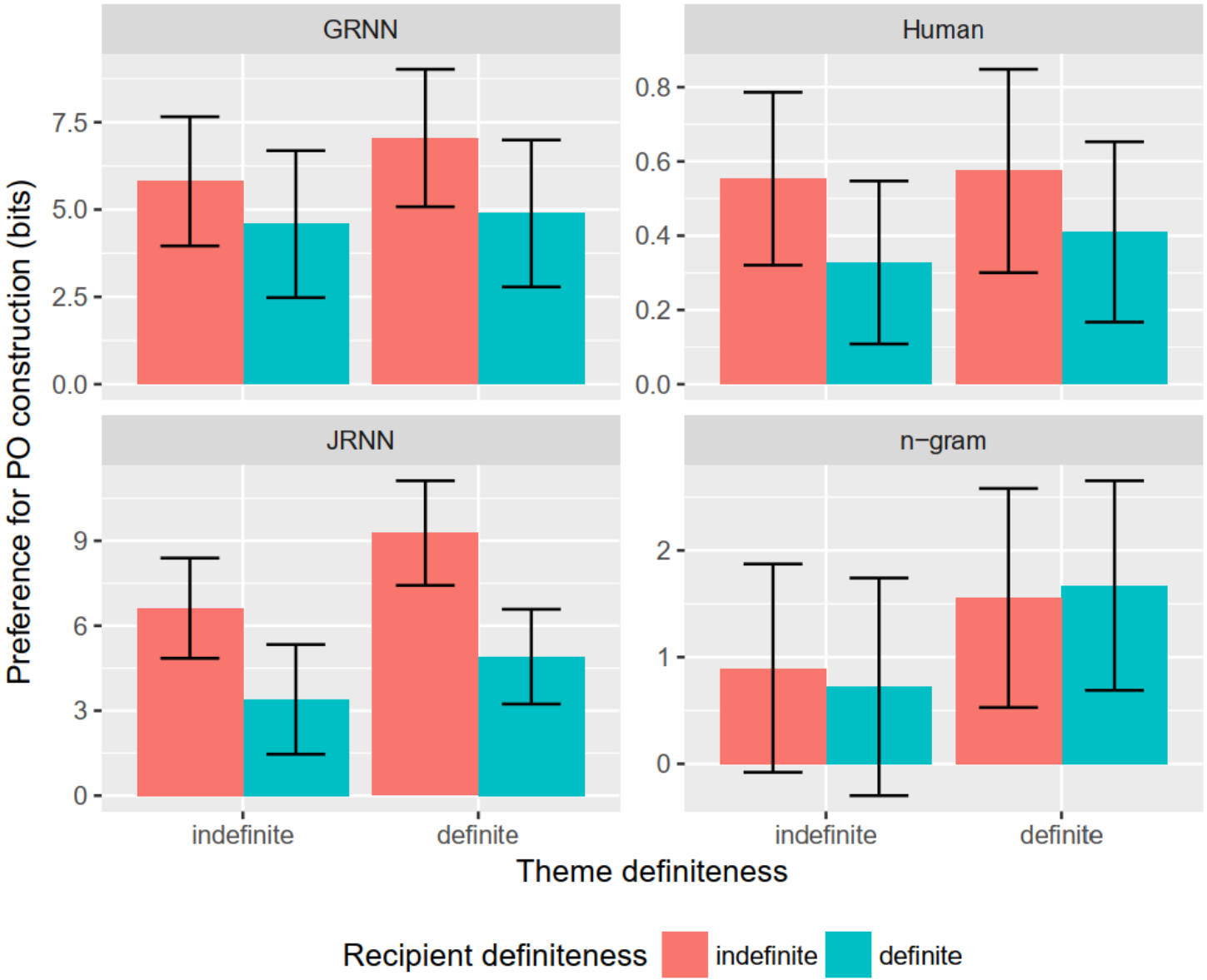
- Double-object (DO): The man gave the woman the book.
- Prepositional-object (PO): The man gave the book to the woman.

Depends on length, definitness, and animacy of recipient (*woman*) and theme(*book*).
Arguably there is a semantic difference.

Results: Dative Alternation



Results: Dative Alternation



Results: Dative Alternation

- Interaction for length effects is statistically significant for all. (If you use $p < 0.05$.)
- Recepient definitness effect present for LSTM models and human data, but not n-gram baseline.
- Theme definiteness effects not significant for human data.

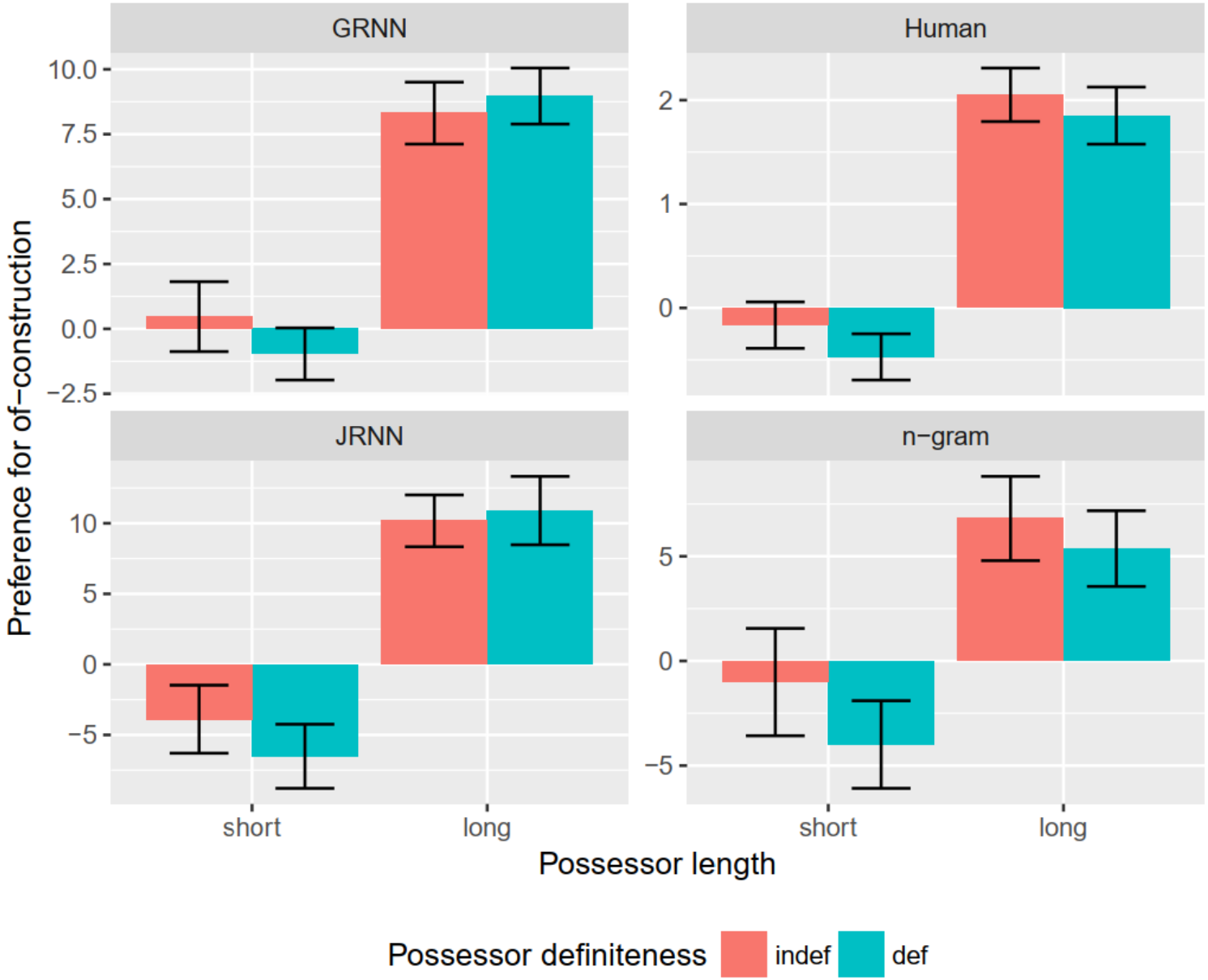
Genitive Alternation

Different word order for possessive constructions:

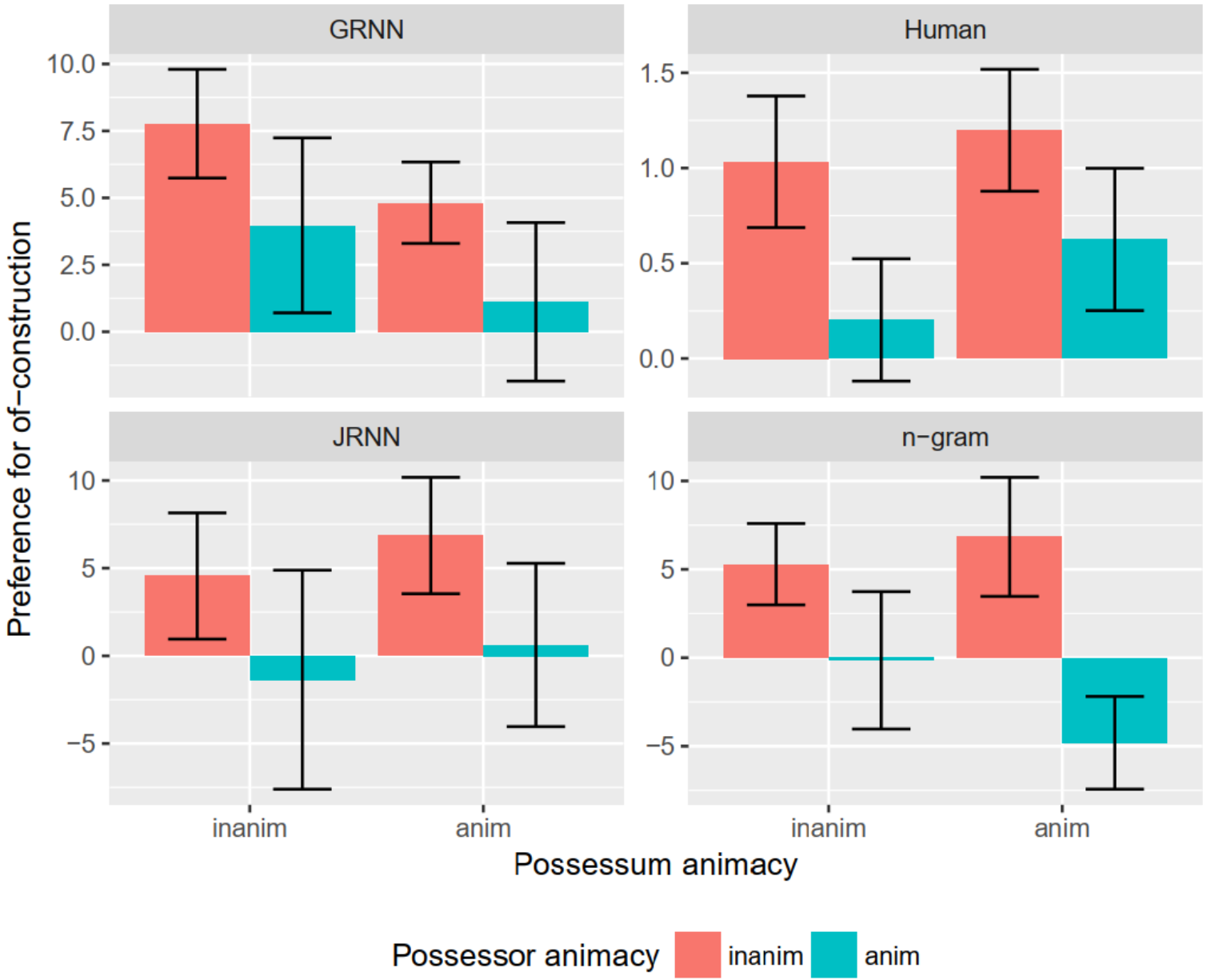
1. The woman's house [*s*-genitive, definite possessor]
2. The house of the woman [*of*-genitive, definite possessor]
3. A woman's house [*s*-genitive, indefinite possessor]
4. The house of a woman [*of*-genitive, indefinite possessor]

Relevant factors: animacy, definiteness, and length of possessor and possessum.

Results: Genitive Alternation



Results: Genitive Alternation



Results: Genitive Alternation

- Possessor length effect is statistically significant for all models and human data.
- Possessor definiteness effects not significant, not even for human data. Contra literature.
- Possessor animacy effect present in JRNN and GRNN, not significant for n-gram model. (Possessum animacy effects not significant.)

Issues: Syntax

- Word order is a syntactic feature, but it can have semantic impact and it can be influenced by semantic features.
- Some of the differences are also not just about word order, e.g. different genitive constructions use different words.
- Hence, showing that RNNs learn the same abstract word order preferences as humans is not exactly the same as showing that RNNs learn syntax. Although it is about learning the use of a syntactic feature!

Issues: Comparison

- We measure surprisal (and interaction) on the side of the RNNs, but acceptability judgements on the side of the human participants (in degrees!).
- At best the paper can say that the effect goes the same direction, but not evalute the strength.
- However, Futrell & Levy write e.g.: "The strongest effects which are most in line with the linguistic literature come from JRNN."
- That matters also because in a number of cases the n-gram model shows the same direction of the effect just weaker.

Conclusion

- They find some evidence for human-like abstract worder preferences using LSTMs.
- They are not always present in LSTM language models and not always absent in n-gram model.
- Arguably overinterpret results.
- How much is it about syntax?

Thank you for your attention!