Discourse structure interacts with reference but not syntax in neural language models – Supplemental Materials

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

1 Overview

We include additional figures and full statistical model outputs. Details of measures, analyses, and statistic models are included in the main paper.

2 Stereotypically gendered nouns used in referential experiments

| male      | female     |
|-----------|------------|
| man       | woman      |
| boy       | girl       |
| father    | mother     |
| uncle     | aunt       |
| husband   | wife       |
| actor     | actress    |
| prince    | princess   |
| waiter    | waitress   |
| lord      | lady       |
| king      | queen      |
| son       | daughter   |
| nephew    | niece      |
| brother   | sister     |
| grandfather | grandmother |

3 Referential Behavioral Results

Statistical models with a categorical IC variable are given for LSTM LMs in Table 1, for TransformerXL in Table 3, and for GPT-2 XL in Table 5. Models with the continuous IC bias measure from Ferstl et al. (2011) are given for LSTM LMs are in Table 2, for TransformerXL in Table 4, and for GPT-2 XL in Table 6.

4 Referential Representational Results

Statistical models with a categorical IC variable are given for LSTM LMs in Table 7, for TransformerXL in Table 9, and for GPT-2 XL in Table 11. Models with the continuous IC bias measure from Ferstl et al. (2011) are given for LSTM LMs are in Table 8, for TransformerXL in Table 10, and for GPT-2 XL in Table 12. The full layer-wise results are given for GPT-2 XL in Figure 2.

5 Syntactic Behavioral Results

The influence of IC on RC verb surprisal is given in Figure 1. Statistical models for the sentence completion experiments from (Rohde et al., 2011) are given for LSTM LMs are given in Table 13, for TransformerXL in Table 14, and for GPT-2 XL in Table 15.

Statistical models for the self-paced reading experiments from (Rohde et al., 2011) are given for LSTM LMs are given in Table 16, for TransformerXL in Table 17, and for GPT-2 XL in Table 18.
Figure 2: Layer-wise representational similarity for GPT-2 XL between pronoun and subject/object; stimuli from Ferstl et al. (2011) (e.g., the man accused the boy because he _). The predicted human-like pattern is given in the rightmost figure. Broken into antecedent (subject vs. object) and IC bias type (subject-bias vs. object-bias). Greater similarity corresponds to greater relationship between pronoun and antecedent.
6 Syntactic Representational Results

Statistical models fitting the similarity between *who* and the possible attachment positions are given for LSTM LMs are given in Table 19, for TransformerXL in Table 20, and for GPT-2 XL in Table 21.

Statistical models fitting the similarity between *was/were* and the possible attachment positions are given for LSTM LMs are given in Table 22, for TransformerXL in Table 23, and for GPT-2 XL in Table 24. Additionally, the full layer-wise results of GPT-2 XL comparing *who* to attachment positions are given in Figure 3 and comparing the RC verb to possible attachment positions are given in Figure 4.

### References

Evelyn C Ferstl, Alan Garnham, and Christina Manouilidou. 2011. *Implicit causality bias in English: A corpus of 300 verbs*. Behavior Research Methods, 43(1):124–135.

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. *Anticipating explanations in relative clause processing*. Cognition, 118(3):339–358.

| LSTM Pronoun Surprisal | (Intercept) | 3.48*** |
|-------------------------|-------------|---------|
|                         | (0.07)      |         |
| hasIC                   | −0.01       |         |
| (0.01)                  |             |         |
| isHigh                  | −0.05***    |         |
| (0.01)                  |             |         |
| gender                  | −0.79***    |         |
| (0.01)                  |             |         |
| hasIC:isHigh            | −0.01       |         |
| (0.01)                  |             |         |
| hasIC:gender            | −0.01       |         |
| (0.01)                  |             |         |
| isHigh:gender           | 0.03**      |         |
| (0.01)                  |             |         |
| hasIC:isHigh:gender     | 0.01        |         |
| (0.01)                  |             |         |
| AIC                     | −4624.15    |         |
| BIC                     | −4548.85    |         |
| Log Likelihood          | 2322.08     |         |
| Num. obs.               | 13776       |         |
| Num. groups: item       | 14          |         |
| Var: item (Intercept)   | 0.08        |         |
| Var: Residual           | 0.04        |         |

Table 1: Linear mixed effects model fitting LSTM surprisal at the pronoun for stimuli from Ferstl et al. (2011). hasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. isHigh corresponds to what position the pronoun refers to (subject or object).
|                      | LSTM Pronoun Surprisal                      | TransformerXL Pronoun Surprisal                       |
|----------------------|---------------------------------------------|-----------------------------------------------------|
| (Intercept)          | 3.4746***                                   | 3.94***                                             |
|                      | (0.0740)                                    | (0.13)                                              |
| bias                 | −0.0001                                     | −0.05**                                             |
|                      | (0.0001)                                    | (0.02)                                              |
| isHigh               | −0.0576***                                   | 0.45***                                             |
|                      | (0.0049)                                    | (0.02)                                              |
| gender               | −0.7911***                                   | −0.99***                                            |
|                      | (0.0049)                                    | (0.02)                                              |
| bias:isHigh          | −0.0001                                     | −0.23***                                            |
|                      | (0.0001)                                    | (0.03)                                              |
| bias:gender          | −0.0001                                     | 0.05                                                |
|                      | (0.0001)                                    | (0.03)                                              |
| isHigh:gender        | −0.46***                                     | −0.46***                                            |
|                      | (0.03)                                      | (0.03)                                              |
| bias:isHigh:gender   | 0.0377***                                    | 0.18***                                             |
|                      | (0.0069)                                    | (0.04)                                              |
| AIC                  | −4593.1510                                  | 22023.43                                            |
| BIC                  | −4517.8442                                  | 22098.74                                            |
| Log Likelihood       | 2306.5755                                   | −11001.72                                           |
| Num. obs.            | 13776                                       | 13776                                               |
| Num. groups: item    | 14                                          | 14                                                  |
| Var: item (Intercept)| 0.0765                                      | 0.24                                                |
| Var: Residual        | 0.0413                                      | 0.29                                                |

Table 2: Linear mixed effects model fitting LSTM surprisal at the pronoun for stimuli from Ferstl et al. (2011). bias corresponds to the IC bias for the verb. isHigh corresponds to what position the pronoun refers to (subject or object).

Table 3: Linear mixed effects model fitting TransformerXL surprisal at the pronoun for stimuli from Ferstl et al. (2011). hasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. isHigh corresponds to what position the pronoun refers to (subject or object).
| TransformerXL Pronoun Surprisal |  |
|----------------------------------|--|
| (Intercept)                      | 3.9097***  |
|                                  | (0.1319)   |
| bias                            | −0.0002    |
|                                  | (0.0002)   |
| isHigh                          | 0.3295***  |
|                                  | (0.0129)   |
| gender                          | −0.9668*** |
|                                  | (0.0129)   |
| bias:isHigh                     | −0.0023*** |
|                                  | (0.0002)   |
| bias:gender                     | 0.0001     |
|                                  | (0.0002)   |
| isHigh:gender                   | −0.3709*** |
|                                  | (0.0182)   |
| bias:isHigh:gender              | 0.0018***  |
|                                  | (0.003)    |

| Log Likelihood                  | −11013.9828 |
| Num. obs.                       | 13776       |
| Num. groups: item               | 14          |
| Var: item (Intercept)           | 0.2425      |
| Var: Residual                   | 0.2861      |

| **p < 0.001; **p < 0.01; *p < 0.05 |

| GPT-2 XL Pronoun Surprisal |  |
|----------------------------|--|
| (Intercept)                | 1.62***  |
|                            | (0.02)   |
| hasIC                      | 0.51***  |
|                            | (0.02)   |
| isHigh                     | 0.24***  |
|                            | (0.02)   |
| gender                     | 0.29***  |
|                            | (0.02)   |
| hasIC:isHigh               | −0.97*** |
|                            | (0.02)   |
| hasIC:gender               | 0.06**   |
|                            | (0.02)   |
| isHigh:gender              | −0.01    |
|                            | (0.02)   |
| hasIC:isHigh:gender        | −0.05    |
|                            | (0.03)   |

| AIC                         | 19612.48  |
| BIC                         | 19687.79  |
| Log Likelihood              | −9796.24  |
| Num. obs.                   | 13776     |
| Num. groups: item           | 14        |
| Var: item (Intercept)       | 0.01      |
| Var: Residual               | 0.24      |

**p < 0.001; **p < 0.01; *p < 0.05

Table 4: Linear mixed effects model fitting TransformerXL surprisal at the pronoun for stimuli from Ferstl et al. (2011). bias corresponds to the IC bias for the verb. isHigh corresponds to what position the pronoun refers to (subject or object).

Table 5: Linear mixed effects model fitting GPT-2 XL surprisal at the pronoun for stimuli from Ferstl et al. (2011). hasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. isHigh corresponds to what position the pronoun refers to (subject or object).
### GPT-2 XL Pronoun Surprisal

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| (Intercept)          | 1.8758***   | (0.0213)  | < 0.001 |
| bias                 | 0.0049***   | (0.0001)  | < 0.001 |
| isHigh               | −0.2594***  | (0.0115)  | < 0.001 |
| gender               | 0.3184***   | (0.0115)  | < 0.001 |
| bias:isHigh          | −0.0093***  | (0.0002)  | < 0.001 |
| bias:gender          | 0.0006**    | (0.0002)  | < 0.01   |
| isHigh:gender        | −0.0402*    | (0.0163)  | < 0.05   |
| bias:isHigh:gender   | −0.0005     | (0.0003)  |         |

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| hasIC:NP             | 0.004       | (0.0005)  | < 0.001 |
| hasIC:layer          | 0.001       | (0.0002)  | < 0.001 |
| NP:layer             | −0.009***   | (0.0002)  | < 0.001 |
| hasIC:gender         | −0.000      | (0.0005)  |         |
| NP:gender            | 0.008       | (0.0005)  |         |
| layer:gender         | 0.037***    | (0.0003)  | < 0.001 |
| hasIC:NP:layer       | −0.005      | (0.0003)  |         |
| hasIC:NP:gender      | −0.000      | (0.0007)  |         |
| NP:layer:gender      | −0.012***   | (0.0003)  | < 0.001 |
| hasIC:NP:layer:gender| −0.000      | (0.0005)  |         |

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| hasIC:layer:gender   | 0.000       | (0.0003)  |         |
| NP:layer:gender      | −0.012***   | (0.0003)  | < 0.001 |
| hasIC:NP:layer:gender| −0.000      | (0.0005)  |         |

**AIC** | 18897.7612  
**BIC** | 18973.0681  
**Log Likelihood** | −9438.8806  
**Num. obs.** | 13776  
**Num. groups: item** | 14  
**Var: item (Intercept)** | 0.0054  
**Var: Residual** | 0.2284  

**p < 0.001; **p < 0.01; *p < 0.05

Table 6: Linear mixed effects model fitting GPT-2 XL surprisal at the pronoun for stimuli from Ferstl et al. (2011). bias corresponds to the IC bias of the verb. isHigh corresponds to what position the pronoun refers to (subject or object).

### LSTM Pronoun Similarity

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| (Intercept)          | 0.013       | (0.011)   |         |
| hasIC                | −0.001      | (0.004)   |         |
| NP                   | −0.063***   | (0.004)   | < 0.001 |
| layer                | 0.167***    | (0.002)   | < 0.001 |
| gender               | −0.076***   | (0.015)   | < 0.001 |
| hasIC:NP             | 0.004       | (0.005)   |         |
| hasIC:layer          | 0.001       | (0.002)   |         |
| NP:layer             | −0.009***   | (0.002)   | < 0.001 |
| hasIC:gender         | −0.000      | (0.005)   |         |
| NP:gender            | 0.008       | (0.005)   |         |
| layer:gender         | 0.037***    | (0.003)   | < 0.001 |
| hasIC:NP:layer       | −0.005      | (0.003)   |         |
| hasIC:NP:gender      | −0.000      | (0.007)   |         |
| NP:layer:gender      | −0.012***   | (0.003)   | < 0.001 |
| hasIC:NP:layer:gender| −0.000      | (0.005)   |         |

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| hasIC:layer:gender   | 0.000       | (0.003)   |         |
| NP:layer:gender      | −0.012***   | (0.003)   | < 0.001 |
| hasIC:NP:layer:gender| −0.000      | (0.005)   |         |

|                      | Coefficient | Std. Err. | p-value |
|----------------------|-------------|-----------|---------|
| hasIC:layer:gender   | 0.000       | (0.003)   |         |
| NP:layer:gender      | −0.012***   | (0.003)   | < 0.001 |
| hasIC:NP:layer:gender| −0.000      | (0.005)   |         |

**AIC** | −89045.922  
**BIC** | −88897.893  
**Log Likelihood** | 44540.961  
**Num. obs.** | 27552  
**Num. groups: item** | 14  
**Var: item (Intercept)** | 0.001  
**Var: Residual** | 0.002  

**p < 0.001; **p < 0.01; *p < 0.05

Table 7: Linear mixed effects model fitting LSTM similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). hasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. NP corresponds to what position the pronoun is compared to (subject or object). layer corresponds to the hidden layer in the model.
### LSTM Pronoun Similarity

| Term            | Estimate | Std. Error | z value | p-value |
|-----------------|----------|------------|---------|---------|
| (Intercept)     | 0.0732***| (0.0112)   |         | <0.001 |
| bias            | -0.0000  | (0.0001)   |         |         |
| NP              | -0.0609***| (0.0026)   |         | <0.001 |
| layer           | 0.1796***| (0.0026)   |         | <0.001 |
| gender          | -0.0848***| (0.0159)   |         | <0.001 |
| bias:NP         | 0.0000   | (0.0000)   |         |         |
| bias:layer      | 0.0000   | (0.0000)   |         |         |
| NP:layer        | -0.0118***| (0.0016)  |         | <0.001 |
| bias:gender     | -0.0000  | (0.0001)   |         |         |
| NP:gender       | 0.0082*  | (0.0036)   |         | <0.01  |
| layer:gender    | 0.0490***| (0.0036)   |         | <0.001 |
| bias:NP:layer   | -0.0000  | (0.0000)   |         |         |
| bias:NP:gender  | -0.0000  | (0.0001)   |         |         |
| bias:layer:gender | 0.0000  | (0.0001)   |         |         |
| NP:layer:gender | -0.0123***| (0.0023)  |         | <0.001 |
| bias:NP:layer:gender | 0.0000  | (0.0000)   |         |         |

AIC: -88963.6651
BIC: -88815.6302
Log Likelihood: 44499.8326
Num. obs.: 27552
Num. groups: item: 14
Var: item (Intercept): 0.0008
Var: Residual: 0.0023

### TransformerXL Pronoun Similarity

| Term            | Estimate | Std. Error | z value | p-value |
|-----------------|----------|------------|---------|---------|
| (Intercept)     | 0.1615***| (0.0129)   |         | <0.001 |
| hasIC           | -0.0008  | (0.0014)   |         |         |
| NP              | -0.0283***| (0.0014)  |         | <0.001 |
| layer           | 0.0189***| (0.0001)   |         | <0.001 |
| gender          | -0.0093***| (0.0182)  |         | <0.001 |
| hasIC:NP        | 0.0074***| (0.0020)   |         | <0.001 |
| hasIC:layer     | 0.0000   | (0.0001)   |         |         |
| NP:layer        | 0.0074***| (0.0001)   |         | <0.001 |
| hasIC:gender    | -0.0007  | (0.0020)   |         |         |
| NP:gender       | 0.0005   | (0.0020)   |         |         |
| layer:gender    | 0.0015***| (0.0001)   |         | <0.001 |
| hasIC:NP:layer  | -0.0017***| (0.0002)  |         | <0.001 |
| hasIC:NP:gender | 0.0013   | (0.0028)   |         |         |
| hasIC:layer:gender | 0.0002  | (0.0002)   |         |         |
| NP:layer:gender | -0.0001  | (0.0002)   |         |         |
| hasIC:NP:layer:gender | -0.0003| (0.0003)   |         |         |

AIC: -533928.0402
BIC: -533740.4612
Log Likelihood: 266982.0201
Num. obs.: 247968
Num. groups: item: 14
Var: item (Intercept): 0.0012
Var: Residual: 0.0068

* ***p < 0.001; **p < 0.01; *p < 0.05

Table 8: Linear mixed effects model fitting LSTM similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). bias corresponds to the IC bias of the verb. NP corresponds to what position the pronoun is compared to (subject or object). layer corresponds to the hidden layer in the model.

Table 9: Linear mixed effects model fitting TransformerXL similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). hasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. NP corresponds to what position the pronoun is compared to (subject or object). layer corresponds to the hidden layer in the model.
| TransformerXL Pronoun Similarity | GPT-2 XL Pronoun Similarity |
|----------------------------------|-----------------------------|
| (Intercept)                      | (Intercept)                 |
| 0.1855***                       | 0.6427***                   |
| (0.0129)                        | (0.0096)                    |
| bias                            | hasIC                       |
| −0.0001***                     | −0.0001                    |
| (0.0000)                       | (0.0007)                    |
| NP                              | NP                          |
| −0.0249***                     | 0.0420***                   |
| (0.0010)                       | (0.0007)                    |
| layer                           | layer                       |
| 0.0124***                      | 0.0001***                   |
| (0.0001)                       | (0.0000)                    |
| gender                          | gender                      |
| −0.0919***                     | −0.0227                     |
| (0.0183)                       | (0.0135)                    |
| bias:NP                         | hasIC:NP                    |
| 0.0001***                      | −0.0011                     |
| (0.0000)                       | (0.0009)                    |
| bias:layer                      | hasIC:layer                 |
| 0.0000***                      | 0.0002***                   |
| (0.0000)                       | (0.0000)                    |
| NP:layer                        | NP:layer                    |
| 0.0065***                      | −0.0014***                  |
| (0.0001)                       | (0.0000)                    |
| bias:gender                     | hasIC:gender                |
| −0.0000                       | 0.0003                      |
| (0.0000)                       | (0.0009)                    |
| NP:gender                       | NP:gender                   |
| 0.0012                       | 0.0031**                    |
| (0.0014)                       | (0.0009)                    |
| layer:gender                    | layer:gender                |
| 0.0018***                      | 0.0002***                   |
| (0.0002)                       | (0.0000)                    |
| bias:NP:layer                   | hasIC:NP:layer              |
| −0.0000***                     | −0.0001*                    |
| (0.0000)                       | (0.0000)                    |
| bias:NP:gender                  | hasIC:NP:gender             |
| 0.0000                       | 0.0008                      |
| (0.0000)                       | (0.0013)                    |
| bias:layer:gender               | hasIC:layer:gender          |
| 0.0000                       | −0.0000                     |
| (0.0000)                       | (0.0000)                    |
| NP:layer:gender                 | NP:layer:gender             |
| −0.0002                       | 0.0001*                     |
| (0.0001)                       | (0.0000)                    |
| bias:NP:layer:gender            | hasIC:NP:layer:gender       |
| −0.0000                       | −0.0000                     |
| (0.0000)                       | (0.0000)                    |

AIC | −534228.7812 |
BIC | −534041.2022 |
Log Likelihood | 267132.3906 |
Num. obs. | 247968 |
Num. groups: item | 14 |
Var: item (Intercept) | 0.0012 |
Var: Residual | 0.0068 |

Table 10: Linear mixed effects model fitting TransformerXL similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). Bias corresponds to the IC bias of the verb. NP corresponds to what position the pronoun is compared to (subject or object). Layer corresponds to the hidden layer in the model.

| GPT-2 XL Pronoun Similarity |
|-------------------------------|
| (Intercept) | 0.6427*** |
| (0.0096) |
| hasIC | −0.0001 |
| (0.0007) |
| NP | 0.0420*** |
| (0.0007) |
| layer | 0.0001*** |
| (0.0000) |
| gender | −0.0227 |
| (0.0135) |
| hasIC:NP | −0.0011 |
| (0.0009) |
| hasIC:layer | 0.0002*** |
| (0.0000) |
| NP:layer | −0.0014*** |
| (0.0000) |
| hasIC:gender | 0.0003 |
| (0.0009) |
| NP:gender | 0.0031** |
| (0.0009) |
| layer:gender | 0.0002*** |
| (0.0000) |
| hasIC:NP:layer | −0.0001* |
| (0.0000) |
| hasIC:NP:gender | 0.0008 |
| (0.0013) |
| hasIC:layer:gender | −0.0000 |
| (0.0000) |
| NP:layer:gender | 0.0001* |
| (0.0000) |
| hasIC:NP:layer:gender | −0.0000 |
| (0.0000) |

AIC | −1714643.8420 |
BIC | −1714438.6080 |
Log Likelihood | 857339.9210 |
Num. obs. | 661248 |
Num. groups: item | 14 |
Var: item (Intercept) | 0.0006 |
Var: Residual | 0.0044 |

Table 11: Linear mixed effects model fitting GPT-2 XL similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). HasIC corresponds to a categorical bias where 0 means object-biased and 1 subject-biased. NP corresponds to what position the pronoun is compared to (subject or object). Layer corresponds to the hidden layer in the model.
Table 12: Linear mixed effects model fitting GPT-2 XL similarity between the pronoun and referents for stimuli from Ferstl et al. (2011). bias corresponds to the IC bias of the verb. NP corresponds to what position the pronoun is compared to (subject or object). layer corresponds to the hidden layer in the model.

|                | (Intercept) | bias     | NP       | layer   | gender  | bias:NP | bias:layer | NP:layer | bias:gender | NP:gender | layer:gender | bias:NP:layer | bias:NP:gender | bias:layer:gender | NP:layer:gender | bias:NP:layer:gender |
|----------------|-------------|----------|----------|---------|---------|---------|------------|----------|-------------|-----------|---------------|----------------|-----------------|------------------|-----------------|---------------------|
|                | 0.60127249*** | (0.000959026) | 0.04138159*** | 0.00162516*** | -0.02603201 | -0.00000145 | -0.000001291 | 0.000162516*** | 0.00145307*** | 0.000000373*** | 0.0000000456 | -0.000000132*** | 0.0000000297 | -0.000000003 | -0.0000000065 | 0.0000007309** | -0.000000014 | 0.000000041 |

Table 13: Linear mixed effects model fitting LSTM scores for sentence completion for stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the singular noun is (higher vs. lower nominal).

|                | (Intercept) | hasIC | isHIGH | hasIC:isHIGH | AIC | BIC | Log Likelihood | Num. obs. | Num. groups: item | Var: item (Intercept) | Var: Residual |
|----------------|-------------|-------|--------|--------------|-----|-----|----------------|-----------|------------------|----------------------|--------------|
|                | 0.77***     | -0.02 | -0.51*** | -0.01 | -10.64 | 5.68 | 11.32 | 112     | 14         | 0.01     | 0.04       |

Table 14: Linear mixed effects model fitting TransformerXL scores for sentence completion for stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the singular noun is (higher vs. lower nominal).

|                | (Intercept) | hasIC | isHIGH | hasIC:isHIGH | AIC | BIC | Log Likelihood | Num. obs. | Num. groups: item | Var: item (Intercept) | Var: Residual |
|----------------|-------------|-------|--------|--------------|-----|-----|----------------|-----------|------------------|----------------------|--------------|
|                | 0.87***     | 0.05  | -0.23*** | -0.06 | 42.50  | 58.81 | 15.25 | 112     | 14         | 0.02     | 0.06       |
### GPT-2 XL Sentence Completion Scores

|                          | Estimate | Std. Error |      |      |
|--------------------------|----------|------------|------|------|
| (Intercept)              | 0.82***  | (0.06)     |      |      |
| hasIC                    | 0.03     | (0.07)     |      |      |
| isHIGH                   | -0.35*** | (0.07)     |      |      |
| hasIC:isHIGH             | -0.04    | (0.09)     |      |      |
| AIC                      | 39.97    |            |      |      |
| BIC                      | 56.28    |            |      |      |
| Log Likelihood           | -13.98   |            |      |      |
| Num. obs.                | 112      |            |      |      |
| Num. groups: item        | 14       |            |      |      |
| Var: item (Intercept)    | 0.01     |            |      |      |
| Var: Residual            | 0.06     |            |      |      |

**p < 0.001; **p < 0.01; *p < 0.05

Table 15: Linear mixed effects model fitting GPT-2 XL scores for sentence completion for stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the singular noun is (higher vs. lower nominal).

### LSTM RC Surprisal

|                          | Estimate | Std. Error |      |      |
|--------------------------|----------|------------|------|------|
| (Intercept)              | 3.59***  | (0.08)     |      |      |
| hasIC                    | -0.12    | (0.12)     |      |      |
| isHIGH                   | 1.98***  | (0.12)     |      |      |
| num                      | -0.22    | (0.12)     |      |      |
| hasIC:isHIGH             | -0.19    | (0.16)     |      |      |
| hasIC:num                | 0.34*    | (0.16)     |      |      |
| isHIGH:num               | 0.32     | (0.16)     |      |      |
| hasIC:isHIGH:num         | 0.28     | (0.23)     |      |      |
| AIC                      | 234.29   |            |      |      |
| BIC                      | 266.86   |            |      |      |
| Log Likelihood           | -107.14  |            |      |      |
| Num. obs.                | 192      |            |      |      |
| Num. groups: item        | 12       |            |      |      |
| Var: item (Intercept)    | 0.00     |            |      |      |
| Var: Residual            | 0.16     |            |      |      |

**p < 0.001; **p < 0.01; *p < 0.05

Table 16: Linear mixed effects model fitting LSTM surprisal at RC verb (was/were) for self-paced reading stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the verb agrees with (higher vs. lower nominal). num corresponds to the number of the RC verb.
| TransformerXL RC Surprisal | GPT-2 XL RC Surprisal |
|---------------------------|-----------------------|
| (Intercept)               | 4.67***               |
|                           | (0.34)                |
| hasIC                     | 0.02                  |
|                           | (0.47)                |
| isHIGH                    | 1.39**                |
|                           | (0.47)                |
| num                       | −1.44**               |
|                           | (0.47)                |
| hasIC:isHIGH              | 0.31                  |
|                           | (0.67)                |
| hasIC:num                 | 0.22                  |
|                           | (0.67)                |
| isHIGH:num                | −0.15                 |
|                           | (0.67)                |
| hasIC:isHIGH:num          | −0.38                 |
|                           | (0.95)                |
| AIC                       | 750.55                |
| BIC                       | 783.13                |
| Log Likelihood            | −365.28               |
| Num. obs.                 | 192                   |
| Num. groups: item         | 12                    |
| Var: item (Intercept)     | 0.00                  |
| Var: Residual             | 2.70                  |
|                           |                       |
| (Intercept)               | 3.05***               |
|                           | (0.28)                |
| hasIC                     | 0.64                  |
|                           | (0.35)                |
| isHIGH                    | 2.57***               |
|                           | (0.35)                |
| num                       | −0.25                 |
|                           | (0.35)                |
| hasIC:isHIGH              | 0.16                  |
|                           | (0.50)                |
| hasIC:num                 | −0.43                 |
|                           | (0.50)                |
| isHIGH:num                | −1.36**               |
|                           | (0.50)                |
| hasIC:isHIGH:num          | −0.41                 |
|                           | (0.70)                |
| AIC                       | 652.88                |
| BIC                       | 685.46                |
| Log Likelihood            | −316.44               |
| Num. obs.                 | 192                   |
| Num. groups: item         | 12                    |
| Var: item (Intercept)     | 0.20                  |
| Var: Residual             | 1.48                  |

***p < 0.001; **p < 0.01; *p < 0.05

Table 17: Linear mixed effects model fitting TransformerXL surprisal at RC verb (was/were) for self-paced reading stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the verb agrees with (higher vs. lower nominal). num corresponds to the number of the RC verb.

Table 18: Linear mixed effects model fitting GPT-2 XL surprisal at RC verb (was/were) for self-paced reading stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. isHIGH corresponds to what position the verb agrees with (higher vs. lower nominal). num corresponds to the number of the RC verb.
Table 19: Linear mixed effects model fitting LSTM similarity between who and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). layer corresponds to the hidden layer in the model.

|                     | LSTM who Similarity | TransformerXL who Similarity |
|---------------------|---------------------|-----------------------------|
| (Intercept)         | −0.07***            | −0.011                      |
|                     | (0.01)              | (0.009)                     |
| hasIC               | −0.06**             | −0.019                      |
|                     | (0.02)              | (0.011)                     |
| NP                  | −0.03               | −0.031**                    |
|                     | (0.02)              | (0.011)                     |
| layer               | 0.15***             | 0.031***                    |
|                     | (0.01)              | (0.001)                     |
| hasIC:NP            | 0.05*               | 0.018                       |
|                     | (0.03)              | (0.016)                     |
| hasIC:layer         | 0.04**              | 0.004***                    |
|                     | (0.01)              | (0.001)                     |
| NP:layer            | 0.06***             | 0.010***                    |
|                     | (0.01)              | (0.001)                     |
| hasIC:NP:layer      | −0.03               | −0.003*                     |
|                     | (0.02)              | (0.001)                     |
| AIC                 | −1248.55            | −5070.279                   |
| BIC                 | −1209.04            | −5008.801                   |
| Log Likelihood      | 634.27              | 2545.140                    |
| Num. obs.           | 384                 | 3456                        |
| Num. groups: item   | 12                  | Num. groups: item           |
| Var: item (Intercept)| 0.00               | 12                          |
| Var: Residual       | 0.00                | Var: item (Intercept)       |

***p < 0.001; **p < 0.01; *p < 0.05

Table 20: Linear mixed effects model fitting TransformerXL similarity between who and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). layer corresponds to the hidden layer in the model.
Table 21: Linear mixed effects model fitting GPT-2 XL similarity between *who* and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). layer corresponds to the hidden layer in the model.

| GPT-2 XL *who* Similarity          |         |
|-----------------------------------|---------|
| (Intercept)                       | 0.6472*** (0.0051) |
| hasIC                             | -0.0037 (0.0043) |
| NP                                | 0.0172*** (0.0043) |
| layer                             | -0.0004*** (0.0001) |
| hasIC:NP                          | 0.0020 (0.0061) |
| hasIC:layer                       | 0.0005** (0.0002) |
| NP:layer                          | 0.0011*** (0.0002) |
| hasIC:NP:layer                    | -0.0005* (0.0002) |
| AIC                               | -22141.1534 |
| BIC                               | -22069.8664 |
| Log Likelihood                    | 11080.5767 |
| Num. obs.                         | 9216 |
| Num. groups: item                 | 12 |
| Var: item (Intercept)             | 0.0002 |
| Var: Residual                     | 0.0052 |

***p < 0.001; **p < 0.01; *p < 0.05

Table 22: Linear mixed effects model fitting LSTM similarity between *was/were* and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). isHIGH corresponds to attachment point that the RC verb agrees with. layer corresponds to the hidden layer in the model.

| LSTM *was/were* Similarity       |         |
|----------------------------------|---------|
| (Intercept)                      | -0.18*** (0.03) |
| hasIC                            | -0.01 (0.04) |
| NP                               | 0.00 (0.02) |
| layer                            | 0.19*** (0.02) |
| isHIGH                           | 0.05 (0.04) |
| hasIC:NP                         | 0.01 (0.03) |
| hasIC:layer                      | -0.00 (0.03) |
| NP:layer                         | 0.00 (0.01) |
| hasIC:NP:layer                   | 0.02 (0.06) |
| hasIC:isHIGH                     | -0.04 (0.03) |
| layer:isHIGH                     | -0.02 (0.03) |
| hasIC:NP:layer:isHIGH            | 0.00 (0.02) |
| hasIC:NP:isHIGH                  | -0.01 (0.04) |
| hasIC:layer:isHIGH               | -0.01 (0.04) |
| NP:layer:isHIGH                  | 0.02 (0.02) |
| hasIC:NP:layer:isHIGH            | 0.00 (0.02) |
| AIC                              | -2522.85 |
| BIC                              | -2439.26 |
| Log Likelihood                   | 1279.42 |
| Num. obs.                        | 768 |
| Num. groups: item                | 12 |
| Var: item (Intercept)            | 0.00 |
| Var: Residual                    | 0.00 |

***p < 0.001; **p < 0.01; *p < 0.05
TransformerXL was/were Similarity

| Term          | Coefficient | Standard Error | t-value | p-value |
|---------------|-------------|----------------|---------|---------|
| (Intercept)   | -0.0710***  | 0.0199         | 0.0001  | <0.001 |
| hasIC         | -0.0560*    | 0.0279         | 0.0001  | <0.05  |
| NP            | 0.0081      | 0.0125         | 0.0001  |         |
| layer         | 0.0255***   | 0.0018         | 0.0001  | <0.001 |
| isHIGH        | 0.0974***   | 0.0279         | 0.0001  | <0.001 |
| hasIC:NP      | 0.0274      | 0.0176         | 0.0001  | <0.05  |
| hasIC:layer   | 0.0063*     | 0.0026         | 0.0001  | <0.05  |
| NP:layer      | 0.0077***   | 0.0012         | 0.0001  | <0.001 |
| hasIC:isHIGH  | 0.0426      | 0.0395         | 0.0001  | <0.05  |
| NP:isHIGH     | -0.0800***  | 0.0176         | 0.0001  | <0.001 |
| layer:isHIGH  | 0.0017      | 0.0026         | 0.0001  |         |
| hasIC:NP:layer| -0.0029     | 0.0016         | 0.0001  |         |
| hasIC:NP:isHIGH| -0.0209    | 0.0250         | 0.0001  |         |
| hasIC:layer:isHIGH| -0.0003   | 0.0036         | 0.0001  |         |
| NP:layer:isHIGH| 0.0005     | 0.0016         | 0.0001  |         |
| hasIC:NP:layer:isHIGH| 0.0001  | 0.0023         | 0.0001  |         |

AIC: -9002.3174
BIC: -8879.1792
Log Likelihood: 4519.1587
Num. obs: 6912
Num. groups: item: 12
Var: item (Intercept): 0.0001
Var: Residual: 0.0154

Table 23: Linear mixed effects model fitting TransformerXL similarity between was/were and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). isHIGH corresponds to attachment point that the RC verb agrees with. layer corresponds to the hidden layer in the model.

GPT-2 XL was/were Similarity

| Term          | Coefficient | Standard Error | t-value | p-value |
|---------------|-------------|----------------|---------|---------|
| (Intercept)   | 0.6188***   | 0.0080         | 0.0001  | <0.001 |
| hasIC         | -0.0089     | 0.0095         | 0.0001  | <0.05  |
| NP            | 0.0132**    | 0.0043         | 0.0001  | <0.01  |
| layer         | -0.0003     | 0.0002         | 0.0001  |         |
| isHIGH        | 0.0179      | 0.0095         | 0.0001  |         |
| hasIC:NP      | 0.0056      | 0.0060         | 0.0001  | <0.05  |
| hasIC:layer   | 0.0007*     | 0.0003         | 0.0001  | <0.05  |
| NP:layer      | 0.0010***   | 0.0003         | 0.0001  | <0.001 |
| hasIC:isHIGH  | 0.0055      | 0.0135         | 0.0001  |         |
| NP:isHIGH     | -0.0183**   | 0.0060         | 0.0001  | <0.01  |
| layer:isHIGH  | 0.0016***   | 0.0003         | 0.0001  | <0.001 |
| hasIC:NP:layer| -0.0005*    | 0.0002         | 0.0001  | <0.05  |
| hasIC:NP:isHIGH| -0.0026    | 0.0085         | 0.0001  |         |
| hasIC:layer:isHIGH| 0.0005   | 0.0005         | 0.0001  |         |
| NP:layer:isHIGH| -0.0009*** | 0.0002         | 0.0001  | <0.001 |
| hasIC:NP:layer:isHIGH| -0.0003 | 0.0003         | 0.0001  |         |

AIC: -44874.4527
BIC: -44733.6595
Log Likelihood: 22455.2264
Num. obs: 18432
Num. groups: item: 12
Var: item (Intercept): 0.0002
Var: Residual: 0.0051

Table 24: Linear mixed effects model fitting GPT-2 XL similarity between was/were and possible attachment position stimuli from Rohde et al. (2011). hasIC corresponds to whether the main verb is an object-biased IC verb or not. NP corresponds to possible attachment point (higher vs. lower nominal). isHIGH corresponds to attachment point that the RC verb agrees with. layer corresponds to the hidden layer in the model.
Figure 3: Layer-wise representational similarity for GPT-2 XL between *who* and the higher/lower nominal; stimuli from Rohde et al. (2011) (e.g., *the man admired the agent of the rockers who*). The predicted human-like pattern is given in the rightmost figure. Broken into attachment location (higher noun vs. lower noun) and verb type (object-biased IC verb vs. non-IC verb). Greater similarity corresponds to greater relationship between attachment location and *who*. 
Figure 4: Layer-wise representational similarity between the RC verb (was/were) and the higher/lower nominal; stimuli from Rohde et al. (2011) (e.g., the man admired the agent of the rockers who was/were). Results broken into attachment location (higher noun vs. lower noun) and verb type (object-biased IC verb vs. non-IC verb) are given in a) for stimuli where the RC verb agrees with the higher nominal (e.g., agent of the rockers who was), and in b) for stimuli where the RC verb agrees with the lower nominal (e.g., rockers who were). The explicit agreement should force a particular attachment location to be preferred, with verb IC bias dampening this effect (the predicted human-like pattern is depicted in b) and d)). Greater similarity corresponds to greater relationship between attachment location and was/were.