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

Supplementary Materials

1 Templates

We created seven templates (one for each of the structures we tested) to generate the adaptation and test sets. Each template had seven slots: subject, object of the relative clause, object of the main clause, verb in the relative clause, verb in the main clause, adverb for the main clause and adverb for the relative clause. The adverb arguments were blank strings half the time. The seven templates varied in the order in which they combined these arguments together to form a sentence. Therefore, for a given set of arguments, we were able to generate seven lexically matched sentences with different structures.

We included several sources of noise in our sentence generation process.

- Each noun slot was filled by a plural noun 40% of the time.
- Every noun phrase was modified with an adjective with 50% probability and every adjective was further modified with an intensifier with 40% probability.
- In cases when a verb (in the main clause or relative clause) was modified by an adverb, the adverb occurred pre-verbally or post-verbally with equal probability.

The slots in the templates were filled by 223 verbs, 164 nouns, 24 adverbs and 78 adjectives. In order to ensure semantic plausibility, we created sub-classes of nouns, adverbs and adjectives and manually specified which sub-classes could combine together. For example, the noun subclass "human" consisted of the nouns *friend*, *cousin*, *partner*, *sibling* and *colleague*. This class could serve as subjects for 38 verbs and could be modified by four sub-classes of adjectives. Similarly the verb *congratulated* could take the noun subclass "human" as its subject and the noun subclasses "scienceperson" and "power" and as its object (e.g., *scientist*, *researcher* etc.; *principal*, *manager* etc.). Additionally, it could be modified by adverb subclasses "sad" and "time" (e.g, *sadly*,*gloomily* etc.; *yesterday*, *last month* etc.)

We ensured that there was no lexical overlap between adaptation and test sets, apart from function words (like *the*, *and*, *by*, *that* etc) and intensifiers (like *very*, *rather*, *quite etc*). We also ensured that verbs, nouns, adverbs and adjectives were not repeated within the same sentence.

2 Relationship between $\mathbb{A}(Y \mid X)$ and *Surp*(Y) prior to adaptation

Figure 1

LM formula: $A(Y \mid X) \sim scale(\textit{Surp}(Y), scale = FALSE)$
$\hat{\beta} = 0.061, \textit{SE} = 0.0003, p < 2e-16$

3 Statistical Analyses:

This section contains details about the statistical analyses for all the results described in the main paper. In describing the formula for our mixed effects models we use standard LMER notation.

3.1 Validating AE as a similarity metric

For this analyses we fit a separate LMEM for each of the different structures that models could get adapted to.

LMER formula:
AE $\sim$ structure + (1 | adaptlist) + (1 | clist)

- Structure is a categorical variable coded as 1 if the test structure is the same as the adaptation structure and -1 if it is different.
- adaptlist: Which of the 10 adaptation-test sets we generated was the model adapted to and tested on?
- clist: Which subset of Wikipedia was the model trained on?

Structure adapted to	$\hat{\beta_{structure}}$	*SE*	p-value
Unreduced Object RC	0.256	0.001	$p < 2e-16$
Reduced Object RC	0.171	0.001	$p < 2e-16$
Unreduced Passive RC	0.229	0.001	$p < 2e-16$
Reduced Passive RC	0.100	0.001	$p < 2e-16$
Active Subject RC	0.194	0.001	$p < 2e-16$
Subject coordination	0.147	0.001	$p < 2e-16$
Object coordination	0.145	0.001	$p < 2e-16$

Table 1: Analysis 5.1

3.2 Similarity between sentences with different types of VP coordination

We fit the following mixed effect model on LMs that were adapted to sentences with coordination.

LMER formula:
AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable coded as 1 if the model was tested on sentences with RCs and -1 if the model was tested on sentences with the other type of coordination (e.g, for model adapted to ASRC-matched coordination, testtype was -1 if it was tested on PS/ORC-matched coordination)

$\hat{\beta} = -0.173, \textit{SE} = 0.0007, p < 2e-16$

3.3 Similarity between sentences with different types of RCs

We fit the following mixed effect model on LMs that were adapted to sentences with RCs.

LMER formula:
AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable coded as 1 if the model was tested on sentences with other types RCs (e.g., for a model adapted to unreduced object RC, the value of testttype was 1 when tested on reduced object RC, reduced/unreduced passive RC and active subject RC). It was coded as -1 if the model was tested on sentences with coordination.

$\hat{\beta} = 0.038, \textit{SE} = 0.0004, p < 2e-16$

3.4 Similarity between sentences belonging to different sub-classes of RCs

We fit the following mixed effect model on LMs that were adapted to sentences with object or passive RCs.

LMER formula:
AE $\sim$ testtype + (1 | adaptlist) + (1 | clist)

testtype was a categorical variable with four levels: passive match, reduced match, no match and both match. Since there were four levels, there were three contrasts. Passive match was chosen as the baseline and coded as 0 for all of the contrasts. For each contrast, one of the other levels was coded as 1 — i.e. in each contrast, the mean adaptation effect of passive match was compared to the mean adaptation effect of one of the other conditions.

Contrast	$\hat{\beta_{testtype}}$	*SE*	p-value
Reduced match vs. Passive match	-0.058	0.001	$p < 2e-16$
Both match vs. Passive match	0.171	0.001	$p < 2e-16$
No match vs. Passive match	-0.143	0.001	$p < 2e-16$

Table 2: Analysis 5.4

3.5 What properties of sentences drive the similarity between them?

We a separate mixed effects model for each of the three linguistically interpretable classes discussed in Section 5.5 of the paper. We did not include the baseline models in these analyses.

LMER formula:
$\mathbb{D}(S, \neg S) \sim$ scale(nhid) * scale(csize) + (1 | adaptlist) + (1 | clist)

nhid refers to the number of hidden units (100, 200, 400, 800, 1600) and csize refers to the training corpus size in millions of tokens (2, 10, 20).

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	0.008	0.002	$p = 0.003$
csize	-0.011	0.001	$p = 0.00002$
nhid:csize	-0.012	0.001	$p = 0.00001$

Table 3: $\mathbb{D}(RC, \neg RC)$

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	0.016	0.001	$p < 2e-16$
csize	-0.006	0.001	$p = 0.00004$
nhid:csize	-0.008	0.001	$p < 0.00001$

Table 4: $\mathbb{D}$(*Reduced match*, ¬*Reduced match*)

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
nhid	-0.007	0.002	$p = 0.008$
csize	-0.023	0.002	$p < 2e-16$
nhid:csize	-0.040	0.001	$p < 2e-16$

Table 5: $\mathbb{D}(RC_X, RC \neq X)$

3.6 Does $\mathbb{D}(RC, \neg RC)$ predict agreement prediction accuracy?

We fit a separate linear regression model for LMs adapted to either reduced or unreduced Object RCs.

LM formula:
accuracy $\sim \mathbb{D}$(*RC*, ¬*RC*) + scale(nhid) + scale(csize)

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}(RC, \neg RC)$	-0.007	0.098	$p = 0.947$
nhid	0.057	0.007	$p \ll 0.0000001$
csize	0.001	0.008	$p = 0.879$

Table 6: Models adapted to unreduced object RCs

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}(RC, \neg RC)$	-0.084	0.113	$p = 0.465$
nhid	0.013	0.005	$p = 0.018$
csize	-0.004	0.008	$p = 0.489$

Table 7: Models adapted to reduced object RCs

4 Relationship between $\mathbb{D}(RC, \neg RC)$ and agreement prediction accuracy for other structures

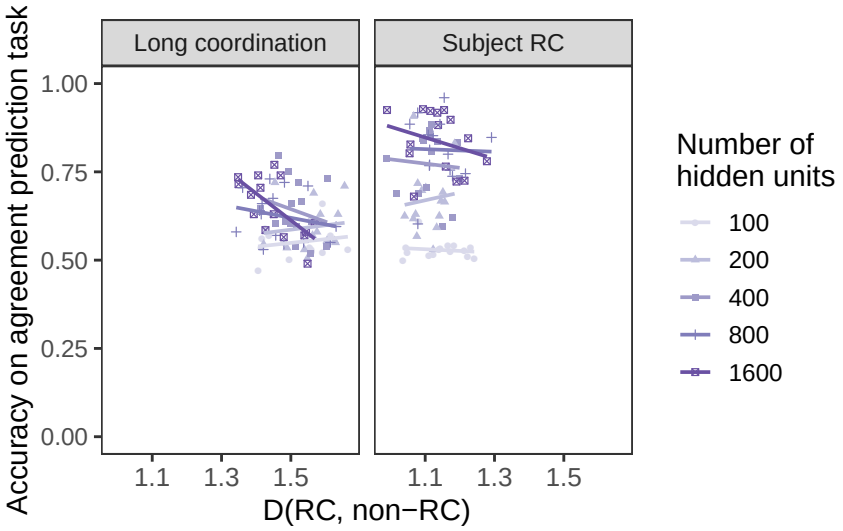

Figure 2

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}(RC, \neg RC)$	-0.215	0.204	$p = 0.297$
nhid	0.089	0.013	$p \ll 0.0000001$
csize	0.016	0.013	$p = 0.211$

Table 8: Models adapted to unreduced active subject RCs

Predictor	$\hat{\beta_{testtype}}$	*SE*	p-value
$\mathbb{D}$(*RC*, ¬*RC*)	-0.125	0.110	$p = 0.259$
nhid	0.023	0.008	$p = 0.014$
csize	0.025	0.008	$p = 0.003$

Table 9: Models adapted to unreduced sentences with long coordinaiton