Supplemental material for "Investigating representations of verb bias in neural language models"

Robert D. Hawkins[*1], **Takateru Yamakoshi**[*1,2], **Thomas L. Griffiths**[1], **Adele E. Goldberg**[1]
[1]Princeton University, [2]University of Tokyo
{rdhawkins,takateru,tomg,adele}@princeton.edu,

Appendix A: Data collection details

There are many possible ways of empirically eliciting acceptability judgements (Marty et al., 2019; Podesva and Sharma, 2014; Langsford et al., 2018). We chose to present pairs of sentences together with a continuous slider to maximize our power to detect gradient preferences. We generated a sentence pair for each verb-theme item by randomly selecting a subject from a list of 8 names (e.g. Juan, Alice), and selecting recipients from a short list corresponding to the given condition (e.g. "him," "her," or "them" for the pronoun condition; "the man," "the woman," "the team" for the short definite condition, etc.) See Table S1 for examples. We implemented our study using jsPsych (De Leeuw, 2015) and paid participants a $1.00 base pay in addition to an additional $1.00 completion bonus.

To ensure data quality, we excluded participants who failed an initial comprehension quiz or either of two attention checks where one of the sentences in the pair was randomly scrambled:

(1) a. The man ate a slice of cake.
b. The man cake of slice ate a.

We also excluded individual trials with response times of < 3 seconds, and all trials from participants who responded this quickly for more than a quarter of their responses, since it was not possible to read the sentences in that time. Due to these exclusions, as well as generic participant dropout on Mechanical Turk, not all sentences received the same number of judgements, but we ensured that at least 5 judgements were collected for each sentence pair.

Appendix B: Regression specifications

To evaluate the binary effect of alternating vs. non-alternating verbs in Section 4.1, we constructed a mixed-effects model predicting human preferences including a dummy-coded fixed effect for the "alternating" vs. "non-alternating" classification from Levin (1993). We also included random intercepts and slopes for each human participant.

To evaluate the effect of information structure in Section 4.1, our mixed-effects model included fixed effects for recipient length, recipient definiteness, and theme definiteness. We included random intercepts and effects of recipient length and definiteness for each participant and verb to control for clustered variance at these levels. See Fig. S1 for the full pattern of results, split by "alternating" and "non-alternating" verbs. Complete regression results are shown in Tables S3 and S4.

Appendix C: Analysis details

For each of three sentence positions of interest investigated in section 5 (after verb, after first argument, and after second argument), we fit a linear regression predicting human judgements from the hidden states. Because of the high dimensionality of these states, we used ridge regression to prevent overfitting[1]. The ridge regression regularization hyper parameter was optimized for each regression model through a log-scale grid search ($\alpha \in [10^0, 10^7]$) on a held-out validation set. As our evaluation metric, we computed R^2, or variance explained. Results were averaged across 10 runs of cross-validation, using random 80/20 splits (see Table S2 for best-performing hyperparameter configurations).

Because the predicted judgements were relative preferences between the two sentences, we concatenated the hidden states of the two sentences together as input. For the 2-layer LSTMs, we used the final hidden state. For the deeper GPT-2 architectures, which are known to represent different information at different layers, we did not know

[1]We used the `sciki-learn` implementation.

a priori which layer would be most appropriate. We thus conducted the regression analysis separately for each layer, and reported the highest performance that was achievable by the model across all layers. In other words, we computed the cross-validated mean performance for each layer and selected the best. This approach has also been used in other recent work (Schrimpf et al., 2020).

References

Joshua R De Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

Steven Langsford, Amy Perfors, Andrew T Hendrickson, Lauren A Kennedy, and Danielle J Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1).

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Paul Marty, Emmanuel Chemla, and Jon Sprouse. 2019. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Manuscript. https://ling. auf. net/lingbuzz/004588*.

Robert J Podesva and Devyani Sharma. 2014. *Research methods in linguistics*. Cambridge University Press.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

DO sentence	**PO sentence**
Michael transported her the food	Michael transported the food to her
Bob recited the woman something	Bob recited something to the woman
Juan took a woman a gift	Juan took a gift to a woman
Alice supplied the man who was from work the news	Alice supplied the news to the man who was from work

Table S1: Example sentence pairs

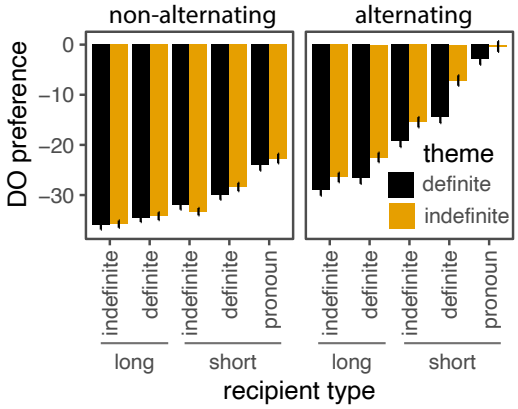

Figure S1: Full pattern of human recipient and theme effects for alternating and non-alternating verbs.

	LSTM	LSTM-large	GPT2	GPT2-large
after verb	$1.3(\pm 0.2)\times 10^2$	$4.3(\pm 0.2)\times 10^2$	$1.0(\pm 0.1)\times 10^4$	$2.4(\pm 0.3)\times 10^4$
after 1st arg.	$1.1(\pm 0.2)\times 10^2$	$2.3(\pm 0.2)\times 10^2$	$6.3(\pm 0.6)\times 10^3$	$1.9(\pm 0.2)\times 10^4$
after 2nd arg.	$1.1(\pm 0.2)\times 10^2$	$1.8(\pm 0.2)\times 10^2$	$1.9(\pm 0.1)\times 10^3$	$3.6(\pm 0.4)\times 10^3$

Table S2: Regularization hyperparameter configuration for each model and task. SEM across cross-validation runs in parentheses.

term	estimate	t statistic	df	p value
(Intercept)	36.44	31.02	229.52	$< 1.0 \times 10^{-32}$
recipient length long vs pronoun	-16.27	-21.15	257.31	$< 1.0 \times 10^{-32}$
recipient length short vs pronoun	-8.00	-18.19	281.29	$< 1.0 \times 10^{-32}$
recipient definite vs. indefinite	-3.91	-14.41	194.96	$< 1.0 \times 10^{-32}$
theme definite vs indefinite	2.23	10.99	46616.18	$< 1.0 \times 10^{-32}$

Table S3: Fixed effect estimates for human mixed-effects regression, including random effects at the verb-level and participant level. Recipient length, recipient definiteness, and theme definiteness are dummy coded.

Figure S2: Histograms of individual slider responses for all 200 verbs. Verbs are ranked from lowest mean preference for DO to highest mean preference for DO. Verbs classified as "non-alternating" by (Levin, 1993) colored red, "alternating" colored blue.

random group	term	estimate
participant	sd(Intercept)	9.18
participant	cor(Intercept, recipient length long vs pronoun)	-0.52
participant	cor(Intercept, recipient length short vs pronoun)	-0.45
participant	cor(Intercept, recipient definite vs. indefinite)	-0.76
participant	sd(recipient length long vs. pronoun)	8.96
participant	cor(recipient length long vs pronoun, short vs. pronoun)	0.84
participant	cor(recipient length long vs pronoun, definite vs indefinite)	0.90
participant	sd(recipient length short vs. pronoun)	6.81
participant	cor(recipient length short vs pronoun, definite vs indefinite)	0.91
participant	sd(recipient definite vs indefinite)	0.52
verb	sd(Intercept)	15.70
verb	cor(Intercept, recipient length long vs pronoun)	-0.93
verb	cor(Intercept, recipient length short vs pronoun)	-0.76
verb	cor(Intercept, recipient definite vs. indefinite)	-0.80
verb	sd(recipient length long vs pronoun)	9.22
verb	cor(recipient length long vs pronoun, short vs. pronoun)	0.92
verb	cor(recipient length long vs pronoun, definite vs indefinite)	0.76
verb	sd(recipient length short vs. pronoun)	3.48
verb	cor(recipient length short vs pronoun, definite vs indefinite)	0.66
verb	sd(recipient definite vs indefinite)	2.19
Residual	sd(observation)	22.25

Table S4: Random-effect estimates for human mixed-effects regression.

model	regression term	estimate	t statistic	df	p value	sig. level
bert	(Intercept)	-2.83	-6.66	118.43	9.22e-10	***
bert	recipient length pronoun vs. long	-6.08	-12.75	99.34	1.23e-22	***
bert	recipient length pronoun vs. short	2.59	8.00	143.07	3.91e-13	***
bert	recipient definite vs. indefinite	-5.68	-18.98	99.08	9.10e-35	***
bert	theme definite vs. indefinite	4.00	19.99	2198.00	7.95e-82	***
gpt2	(Intercept)	1.01	5.59	121.11	1.43e-07	***
gpt2	recipient length pronoun vs. long	-6.43	-29.23	100.52	6.02e-51	***
gpt2	recipient length pronoun vs. short	-2.44	-17.44	202.93	3.09e-42	***
gpt2	recipient definite vs. indefinite	-0.25	-2.00	99.42	4.80e-02	*
gpt2	theme_ypeindef	0.96	11.13	2198.00	5.14e-28	***
gpt2-large	(Intercept)	0.20	1.09	116.31	2.80e-01	n.s.
gpt2-large	recipient length pronoun vs. long	-5.81	-27.91	99.00	1.02e-48	***
gpt2-large	recipient length pronoun vs. short	-1.78	-12.85	99.00	7.96e-23	***
gpt2-large	recipient definite vs. indefinite	-0.57	-4.80	99.00	5.65e-06	***
gpt2-large	theme definite vs. indefinite	1.44	17.00	2099.00	8.04e-61	***
lstm	(Intercept)	-1.85	-9.02	124.11	2.92e-15	***
lstm	recipient length pronoun vs. long	-2.80	-8.80	100.14	4.07e-14	***
lstm	recipient length pronoun vs. short	-0.87	-5.26	219.65	3.44e-07	***
lstm	recipient definite vs. indefinite	-1.33	-12.04	1464.63	7.00e-32	***
lstm	theme definite vs. indefinite	1.61	16.04	2297.00	6.16e-55	***
lstm-large	(Intercept)	-1.19	-3.05	136.46	2.74e-03	**
lstm-large	recipient length pronoun vs. long	-9.38	-20.77	105.00	5.73e-39	***
lstm-large	recipient length pronoun vs. short	-2.30	-6.98	411.84	1.16e-11	***
lstm-large	recipient definite vs. indefinite	-1.02	-3.73	100.60	3.16e-04	***
lstm-large	theme definite vs. indefinite	3.21	14.67	2198.00	1.47e-46	***
ngram	(Intercept)	1.27	13.27	124.45	1.39e-25	***
ngram	recipient length pronoun vs. long	-1.93	-19.59	107.84	2.60e-37	***
ngram	recipient length pronoun vs. short	-1.26	-12.86	107.83	1.68e-23	***
ngram	recipient definite vs. indefinite	-0.04	-0.72	98.99	4.72e-01	n.s.
ngram	theme definite vs. indefinite	0.87	16.59	2197.99	2.59e-58	***

Table S5: Mixed-effects regression results for each model, including random effects at the verb-level. Recipient length, recipient definiteness, and theme definiteness are dummy coded. *** denotes $p < 0.001$, ** denotes $p < 0.01$, * denotes $p < 0.05$, n.s. denotes 'not significant.'