Supplementary material for "A nearly-exhaustive experimental investigation of bridge effects in English"

NICK HUANG
National University of Singapore

DIOGO ALMEIDA
New York University Abu Dhabi

JON SPROUSE
New York University Abu Dhabi

APPENDIX A. VERBS AND DESCRIPTIVE STATISTICS ABOUT ACCEPTABILITY AND PENALTIES (CSV FILE). These figures are derived from our acceptability judgment experiments and data analysis process as reported in Section 3.

APPENDIX B. CORRECTING FOR THE RELIABILITY-BASED ATTENUATION OF R^2. We correct for attenuation in R^2 values with the formula in (1). This formula is derived from the formula for correcting Pearson correlation coefficients, given in (2) (e.g. Spearman 1904; Muchinsky 1996), and the fact that R^2 in simple linear regressions is equivalent to the square of the correlation coefficient between the dependent and independent variables.

(1) $$\text{Corrected } R^2 = \frac{\text{Observed } R^2}{\text{Reliability}_X \times \text{Reliability}_Y}$$

(2) $$\text{Observed correlation}_{X,Y} = \text{True correlation}_{X,Y} \times \sqrt{\text{Reliability}_X \times \text{Reliability}_Y}$$

We obtained reliability estimates through bootstrapping. For each verb, we calculated the long-distance penalty scores for every participant whose responses met our inclusion criteria. For each of the 484 verbs of interest, we created two sets of penalty scores that match in size the original set of scores, by randomly sampling with replacement from the original set. We calculated a mean penalty score for each verb in each set, producing two lists of 484 penalty scores. The Pearson correlation between the two lists was calculated. We repeated this process 5,000 times, taking the mean correlation as the estimate of reliability of long-distance penalties. The reliability of acceptability penalties, in the absence of a dialogue establishing prior context (Section 3.2) is .81, while the reliability of penalties in the presence of such a dialogue (Section 3.3) is .76.

We repeat this analysis for the two backgroundedness measures (True and not-False measures; Section 4.1). The reliability of the True measure is .95, while the reliability of the not-False measure is .85. Note that these estimates are for a total of 482 verbs (the 484 verbs less *bear* and *stand*; *forgive* did not meet our inclusion criteria).

Calculating reliability for the other variables is trickier, since the process presupposes that we can easily obtain new estimates for each measure. This is not feasible for semantic similarity measures, which were derived using computationally intensive methods, nor for measures derived from large, tagged corpora, since there are relatively few of these. For the sake of exposition, we assume perfect reliability (=1) for these measures.

APPENDIX C. REGRESSION RESULTS FOR THE TEMPLATE-BASED PROCESSING THEORY. As described in Section 4.2, we used four different data sets to calculate a set of three semantic similarity measures per data set: a similarity score with *say* as the benchmark, a similarity score with *think* as the benchmark, and a hybrid semantic similarity score that takes whichever score is greater between *say* and *think*. We fitted regression models for each of the twelve similarity measures, but only reported results for four of these measures—the hybrid scores—in the main paper, for space reasons. Tables 1 and 2 present the results for the remaining eight measures.

	No prior context					With prior context				
	b (s.e.)	*t*	*p*	Eff. Size	BF_{10}	*b (s.e.)*	*t*	*p*	Eff. Size	BF_{10}
Similarity to *say*										
LSA/Wikipedia	-0.10 (0.05)	-1.88	.06	0.09	<0.1	-0.15 (0.06)	-2.25	.02	0.13	<0.1
GloVe/Wikipedia	0.20 (0.04)	5.13	<.01	0.24	>100	0.29 (0.05)	5.99	<.01	0.34	>100
GloVe/Gigaword	0.27 (0.04)	6.76	<.01	0.32	>100	0.33 (0.05)	6.70	<.01	0.40	>100
WordNet path-similarity (log)	0.06 (0.04)	1.70	.09	0.07	<0.1	0.07 (0.05)	1.46	.14	0.08	<0.1
Similarity to *think*										
LSA/Wikipedia	-0.13 (0.04)	-3.22	<.01	0.12	0.9	-0.10 (0.05)	-1.88	.06	0.09	<0.1
GloVe/Wikipedia	0.17 (0.04)	4.76	<.01	0.20	>100	0.33 (0.04)	7.52	<.01	0.39	>100
GloVe/Gigaword	0.15 (0.04)	4.12	<.01	0.17	24.5	0.29 (0.05)	6.35	<.01	0.33	>100
WordNet path-similarity (log)	0.10 (0.04)	2.45	.01	0.12	<0.1	0.11 (0.05)	2.21	.03	0.13	<0.1

TABLE 1. Interaction effects between wh-dependency length and semantic similarity scores for models of z-scored acceptability.

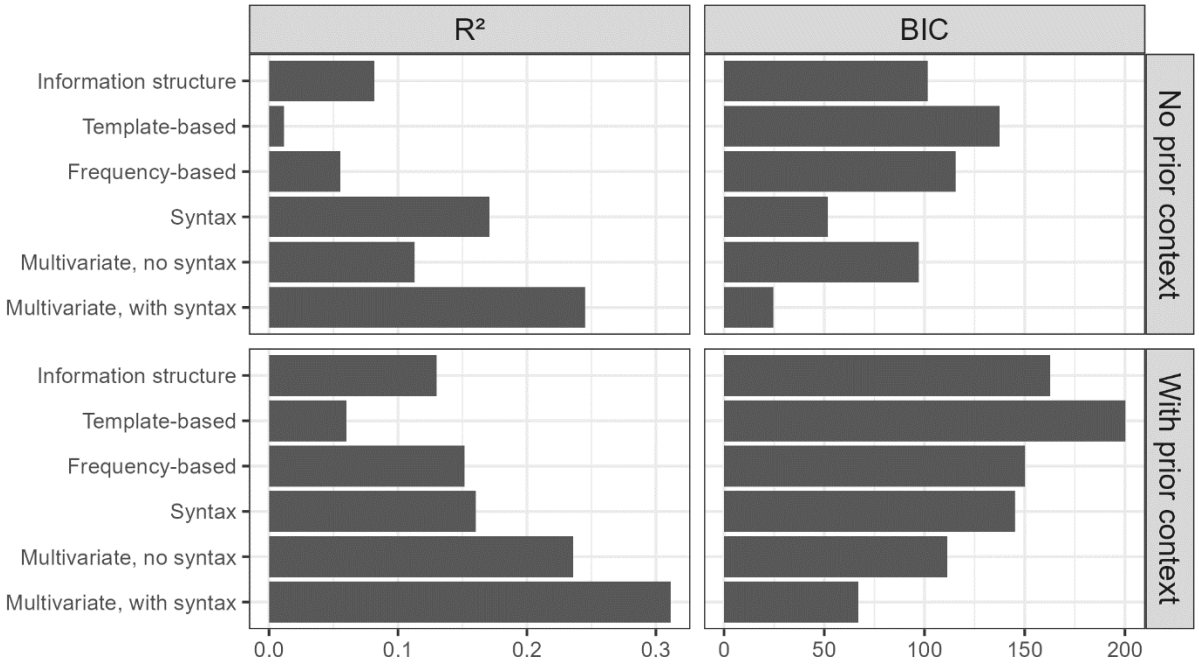
	No prior context						With prior context					
	b (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2	*b* (s.e.)	*t*	Eff. Size	BF_{10}	R^2	Corr. R^2
Similarity to *say*												
LSA/Wikipedia	0.08 (0.10)	0.79	0.07	<0.1	.001	.002	0.14 (0.11)	1.26	0.12	0.1	.004	.005
GloVe/Wikipedia	-0.21 (0.07)	-2.82	0.25	2.5	.018	.022	-0.30 (0.08)	-3.71	0.35	42.3	.030	.040
GloVe/Gigaword	-0.27 (0.08)	-3.62	0.33	30.9	.028	.035	-0.33 (0.08)	-4.12	0.40	>100	.037	.049
WordNet path-similarity (log 10)	-0.06 (0.07)	-0.88	0.07	<0.1	.002	.002	-0.08 (0.08)	-0.99	0.09	<0.1	.002	.003
Similarity to *think*												
LSA/Wikipedia	0.12 (0.08)	1.57	0.11	0.2	.006	.007	0.10 (0.08)	1.13	0.09	<0.1	.003	.004
GloVe/Wikipedia	-0.18 (0.07)	-2.59	0.21	1.3	.015	.019	-0.33 (0.07)	-4.53	0.39	>100	.045	.059
GloVe/Gigaword	-0.16 (0.07)	-2.28	0.18	0.6	.011	.014	-0.29 (0.07)	-3.82	0.33	64.7	.032	.042
WordNet path-similarity (log 10)	-0.11 (0.08)	-1.36	0.12	0.1	.004	.005	-0.11 (0.08)	-1.35	0.13	0.1	.004	.005

TABLE 2. Comparison of models of bridge effects for template-based processing theory.

APPENDIX D. RESULTS FOR LINEAR REGRESSIONS BETWEEN PENALTIES AND BEST-PERFORMING PREDICTORS OF EACH THEORY, FOR ALTERNATIVE VERB EXCLUSION CRITERIA. Sections 5-7 presented analyses for a set of 484 verbs of interest, for which "no context" short wh-dependencies had a z-scored acceptability rating of 0 or greater, on the assumption that verbs with negative ratings do not allow finite clausal complements. As described in Section 8.3, a reviewer expressed concerns over the validity of this criterion, because the presence of the wh-dependency might have also lowered acceptability.

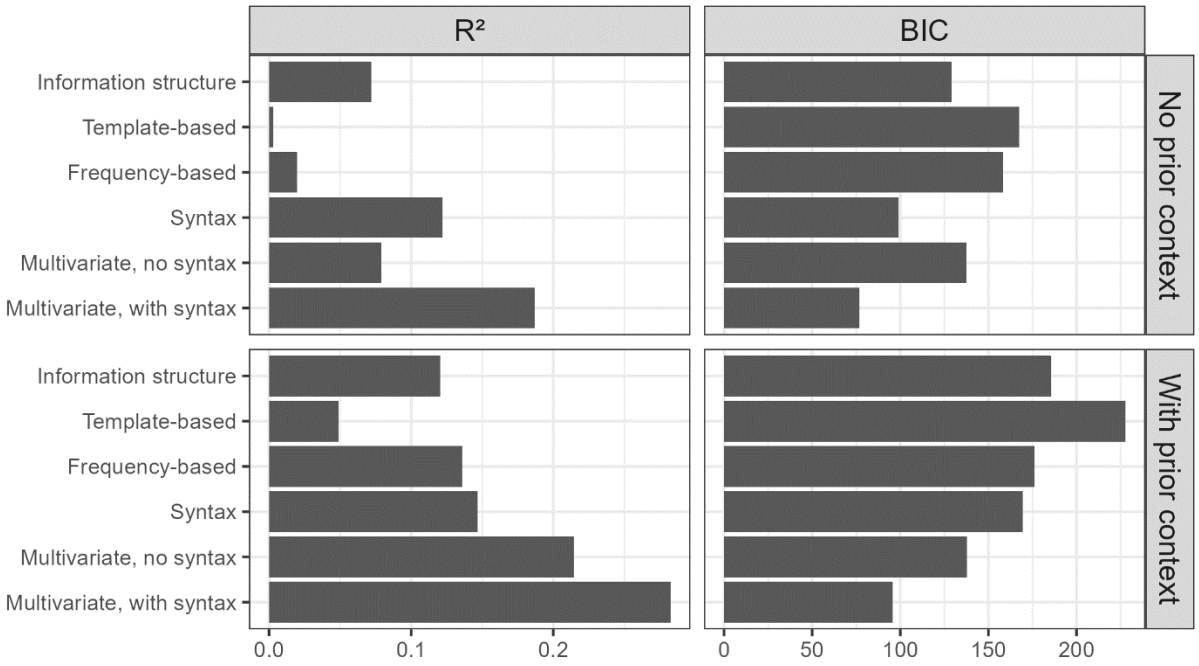
In response to this concern, we re-ran a key analysis—the model fit analysis reported in Figure 9, Section 7.2. This analysis compares model fits for the best-performing predictor for each of the four single-factor theories: %True responses (information structure), GloVe/Wikipedia similarity (template-based), log verb bias (frequency-based), and nonfinite complementation (syntax), and contrasts them with two multivariate models, one that linearly combines all three non-syntactic predictors, and another that linearly combines all four predictors. Crucially, we tried out two different verb exclusion criteria, (i) relaxing the no-context short wh-dependency acceptability threshold to -0.25, and (ii) eliminating it altogether. In both cases, we still required short wh-dependencies to be at least as acceptable as long wh-dependencies. The first criterion (threshold of -.25) yielded a set of 488 verbs with a full set of predictors, and the second (no threshold) a set of 536 verbs with a full set of predictors.

Figures 1 and 2 show model fits (R^2s) and Bayesian Information Criterion (BIC), which balances model fit with a penalty for increased complexity (the lower the BIC, the better). Visually speaking, both figures are very similar to each other (and also to Figure 9 and tables reported in the paper): the syntax-only model is often one of the best single-factor models, with fits comparable to, if not higher than, the information structure-only or frequency-only models. The template-based-only models had the lowest R^2s. The "multivariate, with syntax" models consistently had the highest R^2s and lowest BICs. These results show that our conclusions leading up to and through Section 7 are not sensitive to what verbs were included in our analyses.



Note: The higher the R^2 and the lower the BIC, the better the fit.

FIGURE 1. Model fits for selected models of bridge effects, for the 488 verbs where short wh-dependencies had a z-scored acceptability rating above -0.25 (among other criteria).



Note: The higher the R^2 and the lower the BIC, the better the fit.

FIGURE 2. Model fits for selected models of bridge effects, for a set of 536 verbs where there was no criterion on the acceptability of short wh-dependencies (among other criteria).