Evidence for Composition Operations in Broad-Coverage Sentence Processing

Christian Clark and William Schuler
The Ohio State University

During language comprehension, humans are able to assemble the meanings of incoming words into precise interpretations. This process can be described in terms of composition operations such as filler-gap extraction (Figure 1). Previous work such as [10] has established that these operations influence comprehension difficulty, but such experiments have typically been confined to small-scale, controlled stimuli. Leveraging deep syntactic annotations from a generalized categorial grammar (GCG) [5], we test for more general effects of composition operations using three broad-coverage English corpora: per-word self-paced reading times from the Natural Stories Corpus [2], eye-tracking go-past durations from the Dundee Corpus [3], and fMRI blood oxygen level-dependent (BOLD) measurements from an fMRI version of the Natural Stories Corpus [8]. BOLD measurements are from the language network as defined in [1].

Each corpus was split into exploratory and held-out partitions. For preliminary testing, a set of linear mixed-effects regression (LMER) models was fit to the exploratory partition of each corpus. Each model included an indicator variable marking whether a particular composition operation occurred at each word, plus a set of baseline predictors. Indicator variables were derived from GCG annotations of the Natural Stories and Dundee corpora [9]. The baseline predictors for self-paced reading were word position, word length, and surprisal from GPT-2 [7], a popular Transformer-based language model. The baseline predictors for eye-tracking go-past durations were word position, word length, saccade length, and GPT-2 surprisal; and the baseline predictors for fMRI BOLD were the index of the fMRI sample, the deconvolutional intercept, a flag for sentence-final words, the pause duration after each word, and GPT-2 surprisal. Prior to fitting the LMER models, fMRI predictors were convolved with the canonical hemodynamic response function [4] to temporally align them with BOLD measurements. By-subject random slopes for each fixed effect were initially included in all models, but some were later removed due to convergence errors. For each dataset, an additional LMER model was fit containing only the baseline predictors, and this was used to perform likelihood-ratio tests (LRTs) to determine whether each composition operation contributed to predicting processing difficulty.

Table 1 presents the results of the exploratory testing. All operations show significant effects on at least one dataset. For subsequent evaluation, we focused on extraction due to its relatively low p-values across the three datasets. An additional LRT was run on the held-out partition of each dataset to confirm the observed effect of extraction on comprehension difficulty; Table 2 presents the results. No other predictors were tested on the held-out partitions.

In the held-out data, a facilitation effect from extraction is observed on the two latency-based measures, while extraction is also associated with higher BOLD signal. One potential explanation for the facilitation effect is a filled-gap effect [10]: the processor may be eager to perform extractions as early as possible, which would be consistent with the slowdown reported in [10] when gap-filling is delayed. Note that the GCG-based annotations place extractions right at the verb (Figure 1), consistent with [6], rather than at a gap location. In contrast, the increased BOLD signal could be construed as evidence that some type of structure is being built or stored when extractions are hypothesized, but without delaying processing.

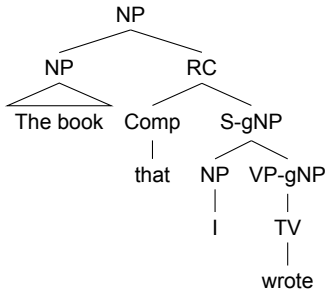

Figure 1: Example analysis of *The book that I wrote*, illustrating filler-gap extraction. An extraction operation occurs at *wrote*, converting it from a transitive verb to a verb phrase with a noun-phrase gap (-gNP).

Predictor	Natural Stories		Dundee		Natural Stories fMRI	
	Effect size ($\log_e$ ms)	p-value	Effect size ($\log_e$ ms)	p-value	Effect size (BOLD)	p-value
Extraction	-9.77E-03	**1.41E-06**	-1.06E-02	1.28E-01	1.90E-01	**1.89E-05**
Preceding argument attachment	3.14E-03	**2.69E-02**	-2.64E-03	5.79E-01	1.11E-01	**3.31E-03**
Succeeding argument attachment	-1.62E-02	**7.50E-41**	-1.35E-02	**3.43E-04**	-5.15E-02	1.47E-01
Preceding conjunct attachment	-1.67E-03	4.54E-01	3.21E-02	**3.72E-06**	-2.00E-01	**2.20E-03**
Succeeding conjunct attachment	2.24E-02	**1.50E-22**	1.01E-02	4.08E-01	-1.78E-01	**1.12E-02**
Preceding modifier attachment	1.79E-02	**4.93E-57**	3.13E-03	3.08E-01	1.95E-02	4.18E-01
Succeeding modifier attachment	7.49E-03	**7.08E-11**	1.77E-02	**4.91E-08**	2.03E-02	4.04E-01

Table 1: Likelihood-ratio test results on the exploratory partitions. Bolded p-values are statistically significant ($p \leq 0.05$).

Predictor	Natural Stories		Dundee		Natural Stories fMRI	
	Effect size ($\log_e$ ms)	p-value	Effect size ($\log_e$ ms)	p-value	Effect size (BOLD)	p-value
Extraction	-1.14E-02	2.69E-06	-1.76E-02	1.03E-02	1.75E-01	1.58E-03
GPT-2 surprisal	6.23E-03		1.55E-02		3.37E-03	

Table 2: Likelihood-ratio test results on the held-out partitions. Only the extraction operation was tested. All p-values are statistically significant ($p \leq 0.05$). Fixed effects of GPT-2 surprisal from the full LMER models are included for comparison. Extraction is about as predictive as one point of surprisal on latency data, and is far more predictive than surprisal on fMRI BOLD.

References

[1] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher. *Journal of Neurophysiology*, 2010.

[2] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. *Language Resources and Evaluation*, 2021.

[3] A. Kennedy, R. Hill, and J. Pynte. In *Proceedings of the 12th European Conference on Eye Movement*, 2003.

[4] M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager. *Neuroimage*, 2009.

[5] L. Nguyen, M. van Schijndel, and W. Schuler. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2012.

[6] M. Pickering and G. Barry. *Language and Cognitive Processes*, 1991.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. *ArXiv*, 2019.

[8] C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, and E. Fedorenko. *Neuropsychologia*, 2019.

[9] C. Shain, M. van Schijndel, and W. Schuler. In *Workshop on Linguistic and Neuro-Cognitive Resources*, 2018.

[10] L. Stowe. *Language and Cognitive Processes*, 1986.