Robust Effects of Working Memory Demand in Language-Selective Cortex

C. Shain[1], I. Blank[2], E. Fedorenko[1], E. Gibson[1], W. Schuler[3] [1]MIT, [2]UCLA, [3]OSU

To understand language, we must infer structured meanings from real-time auditory or visual signals. Researchers have long focused on word-by-word structure building in working memory (WM) as a mechanism that might enable this feat [7, 11, 13]. However, some have argued that language processing does not typically involve rich word-by-word structure building [17, 6], and/or that apparent WM effects are underlyingly driven by surprisal (how predictable a word is in context) [9]. Consistent with this alternative, some recent behavioral studies of naturalistic language processing that control for surprisal have not shown clear WM effects [2, 18, 15].

In this fMRI study, we investigate signatures of word-by-word WM demand during naturalistic language comprehension under rigorous surprisal controls. In addition, we address a related debate about whether the WM mechanisms involved in language comprehension are primarily specialized for language [10, 19, 1] or domain general [8, 16, 4]. To do so, we examine a large public dataset (78 subjects) of fMRI responses to naturalistic story listening [14]. In each participant, we functionally localize (1) the language-selective network [5] and (2) the "multiple-demand" (MD) network [3], which supports WM across domains and is therefore the most plausible candidate domain-general network to support WM for language.

Investigating general WM involvement is challenging because estimators of WM demand for language are highly theory-dependent. To address this, we divide our analysis protocol into exploratory and generalization phases (**Fig 1B**), using continuous-time deconvolutional regression (CDR) [15] for hemodynamic response estimation. In the exploratory phase, we consider a diverse set of 22 WM predictors derived from three existing theories of sentence processing with broad-coverage computational implementations: the Dependency Locality Theory [7], ACT-R theory [11], and left-corner parsing theories [13]. From these, we select those predictors that significantly improve fit to the training data ($p < 0.05$) for follow-up evaluation. In the generalization phase, we statistically evaluate the contribution of predictors selected during the exploratory phase to the fit of pre-trained CDR models on unseen data, using the same 50-50 data split as in [14].

In the language network, the exploratory phase selected two predictors (integration cost and storage cost according to the Dependency Locality Theory [7]) for further evaluation. In the MD network, no WM predictor met criteria for further evaluation. We thus find no evidence of MD involvement in WM for language processing. Results of the generalization phase are shown in **Fig 1A,C**. Integration cost (a measure of syntactic dependency length) is significant in the language network over rigorous surprisal controls, including surprisals derived from a large transformer language model (GPT-2-XL [12]), supporting surprisal-independent effects of memory demand on language network activity. Storage cost is not significant in the language network, and neither WM predictor is significant in the MD network. The language network response to integration cost is significantly larger than the MD network response in direct comparisons.

Results therefore show robust surprisal-independent effects of memory demand in the language network and no effect of memory demand in the multiple-demand network. Our findings thus support the view that language comprehension involves computationally demanding word-by-word structure building operations in working memory, in addition to any prediction-related mechanisms. Further, these memory operations appear to be primarily conducted by the same neural resources that store linguistic knowledge, with no evidence of involvement of brain regions known to support working memory across domains.

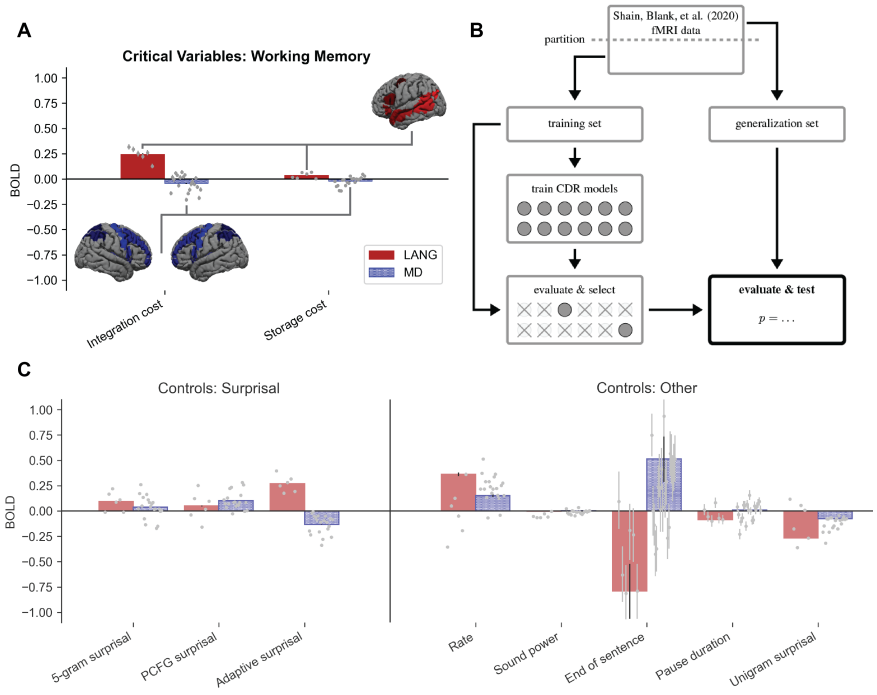

Figure 1: The critical working memory result (**A**), with reference estimates for surprisal variables and other controls shown in **C**. Dots show estimates for specific functional regions of interest, and error bars represent 95% credible intervals. A schematic illustration of the analysis procedure is shown in **B**.

References

[1] Caplan, D. and Waters, G. S. *Behavioral and Brain Sciences*, 1999.
[2] Demberg, V. and Keller, F. *Cognition*, 2008.
[3] Duncan, J. *Trends in Cognitive Sciences*, 2010.
[4] Fedorenko, E., Gibson, E., and Rohde, D. *Journal of Memory and Language*, 2006.
[5] Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. *Journal of Neurophysiology*, 2010.
[6] Frank, S. L. and Bod, R. *Psychological Science*, 2011.
[7] Gibson, E. The Dependency Locality Theory: A distance-based theory of linguistic complexity. 2000.
[8] Just, M. A. and Carpenter, P. A. *Psychological Review*, 1992.
[9] Levy, R. *Cognition*, 2008.
[10] Lewis, R. L. *The Journal of Psycholinguistic Research*, 1996.
[11] Lewis, R. L. and Vasishth, S. *Cognitive Science*, 2005.
[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. *OpenAI Blog*, 2019.
[13] Rasmussen, N. E. and Schuler, W. *Cognitive Science*, 2018.
[14] Shain, C., Blank, I., van Schijndel, M., Schuler, W., and Fedorenko, E. *Neuropsychologia*, 2020.
[15] Shain, C. and Schuler, W. *Cognition*, 2021.
[16] Stowe, L. A., Broere, C. A. J., Paans, A. M. J., Wijers, A. A., Mulder, G., Vaalburg, W., and Zwarts, F. *Neuroreport*, 1998.
[17] Swets, B., Desmet, T., Clifton, C., and Ferreira, F. *Memory and Cognition*, 2008.
[18] van Schijndel, M. and Schuler, W. In *Proceedings of NAACL-HLT 2013*, 2013.
[19] Waters, G. S. and Caplan, D. *Psychological Review*, 1996.