A Rational Neural Network Model of Memory Retrieval for Syntactic Dependency Resolution

William Timkey (wpt2011@nyu.edu)[1], Brian Dillon[2], Tal Linzen[1]

[1]Department of Linguistics, New York University, [2]Department of Linguistics, UMass, Amherst

It is well-established that working memory (WM) is accessed via cue-based retrieval [5]. But how can a system based on surface-level associations rapidly resolve long-distance syntactic dependencies defined by structural relations? One prominent view is that comprehenders retrieve dependency targets using heuristic cues such as number marking [10], semantic compatibility [1], or static syntactic features like `[+subject]` [9]. Although these features clearly affect processing in anomalous sentences, there is surprisingly little evidence that they influence retrieval in typical grammatical sentences [e.g., 4,6,1,7]. An alternative view is that retrieval relies on richer, dynamic syntactic features, but the space of possible features is vast. How can we discover the WM representations subserving dependency resolution? We propose a neural network model of syntactic dependency resolution based on the rational hypothesis that WM representations are optimized for accurate retrieval, and show how its learned representations map onto interpretable mechanistic hypotheses about retrieval cues. Subtle model predictions are verified against reading times in multiple center embedded sentences. More broadly, the model can provide a test bed for generating rational, mechanistic hypotheses about the representational content of WM in sentence processing.

Model: The model is a recurrent neural network arc-eager transition-based dependency parser augmented with an associational WM store (a single QV neural attention head [8]). The WM component stores vector representations of previous words and their dependency structure. When the parser posits a dependency between the current word and a previous one, a retrieval cue vector is generated to probe WM for the dependency target. The retrieved item is composed with the current word (via a feed-forward layer), appended to the WM store, and used to predict the next word and parsing action. On a machine parsed version of the 100m token wikitext-103 dataset, the model is jointly optimized to predict upcoming words and parsing actions, and to retrieve the correct dependency targets. The trained models provide both estimates retrieval interference (measured as the probability of retrieving the correct target from memory), and word surprisal.

Experiments: The model's learned representations can be causally interpreted through activation patching (Fig. 1a). Using simple stimuli from [10], we test hypotheses (Fig 1b) about how the model resolves subject-verb dependencies. The model is unaffected by manipulations to the subject's number or lexical semantics, interference only arises from subject nouns that are not yet headed by a verb. This can explain difficulty at the penultimate verb in sentences with multiple center-embedding [e.g., 3], a rare point at which multiple long distance subject-verb dependencies are unresolved. We tested this prediction on stimuli from [2] (Fig 2a-b), and verify that models predict greater interference as embedding depth increases (Fig 2c-e). The model also correctly predicts that semantic compatibility between the penultimate verb and its subject can facilitate processing, particularly with multiple levels of center embedding, suggesting that semantic cues influence retrieval when syntactic cues are insufficent. Our approach provides a principled method for generating testable hypotheses about WM and retrieval cue representations through rational analysis. Our rational model reveals that learners should generally prioritize highly discriminative structural cues over surface-level item features like number or lexical semantics in forming dependencies.

1 Figures and Tables

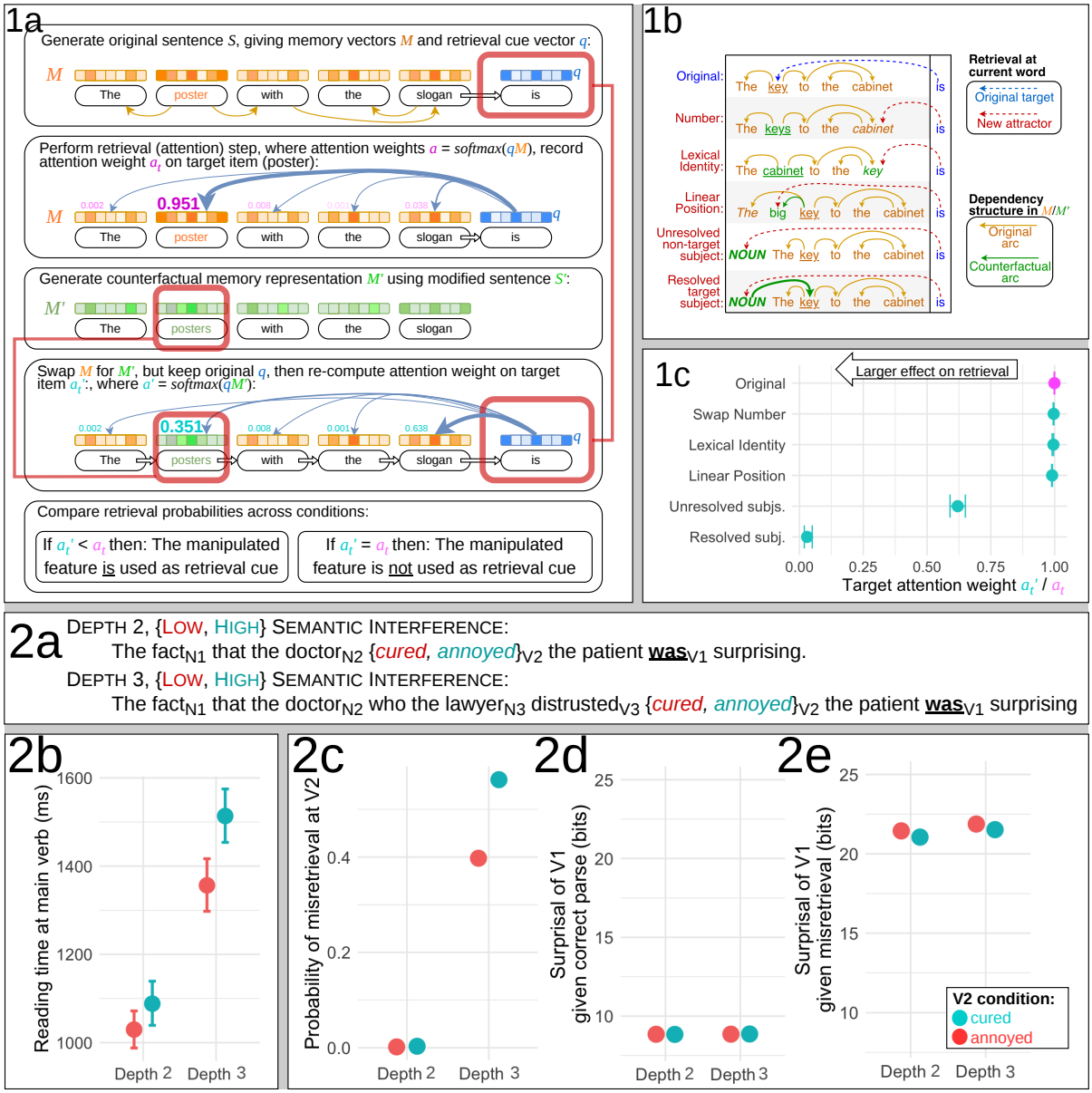

Figure 1: **1a)** An overview of the activation patching method; **1b)** The specific manipulations we evaluated; **1c)** Results of the activation patching experiments, averaged across all grammatical stimuli from [10; Exp. 5]. The baseline retrieval accuracy on the original sentences was nearly perfect. The only manipulations that affected retrieval accuracy were those that added or removed an unresolved subject; **2a-b)** Stimuli (2a) and reading times at V1 (2b) from [2; Exp. 2], omitting some conditions due to space constraints; **2c-e)** Results from our model, measured as: 2c) the probability of retrieving a non-target item at V2; 2d) the surprisal of V1 given that there was no misretrieval; 2e) the surprsial of V1 given that the main subject N1 was erroneously retrieved as the subject of V2.

[1] Cunnings, I. & Sturt, P. (2018). *Journal of Memory & Language* **[2]** Hahn, M. et al. (2022). *PNAS* **[3]** Häussler, J. & Bader, M. (2015). *Frontiers in Psychology* **[4]** Laurinavichyute, A. & von der Malsburg, T. (2024). *Journal of Memory & Language* **[5]** McElree, B. (2006). In *Psychology of Learning and Motivation* **[6]** Schlueter, Z. et al. (2019). *Frontiers in Psychology* **[7]** Schoknecht, P. et al. (2025). *Journal of Memory & Language* **[8]** Timkey, W. & Linzen, T. (2023). *Findings of EMNLP* **[9]** Van Dyke, J. (2007). *Journal of experimental psychology.* **[10]** Wagers, M. et al. (2009). *Journal of Memory & Language*