A transient binding model of interference in sentence processing

Maayan Keshev, Mandy Cartner, Aya Meltzer-Asscher, and Brian Dillon

Sentence comprehension is vulnerable to interference effects. Research into this has focused on interference to the retrieval of memory items rather than to their maintenance [1]. However, the question of how transient item-position associations are encoded is a key issue in working memory research [2-3]. We propose a model of this working memory function during sentence processing that uses a distributed neural architecture to represent the correspondence between lexical items / morphemes and their syntactic position in a sentence. Our proposal extends a neural network model of item-position associations in working memory ('serial-order-in-a-box' and its 'SOB-complex-span' version [4-5]). We show that encoding transient morpheme-position associations accounts for various interference effects in sentence processing.

Model: Based on [3-5], we assume that working memory encodes transient associations between distributed morpheme encodings, in a designated *item* layer, and distributed syntactic position markers, in a designated *position* layer (Figure 1). During encoding, a fully connected weight matrix **W** is learned via a Hebbian update rule (Figure 1A). After encoding item-position pairs, retrieval proceeds by reinstating morpheme information encoded in **W** for a desired syntactic position vector ***p*** (Figure 1B). The reconstructed item ***v'*** is a version of the encoded item with distortions due to overlapping associations between other positions and items in **W**. ***v'*** is then compared against the lexical roots available in memory and the alternative features to determine which morphemes to access (Figure 1C).

We compare the predictions of our model against two empirical datasets of English speakers' responses to a 4AFC task targeting the subject of English sentences as in (1) [6] and (2) [7].

(1) The apprentice of the chef(s) worked diligently.
Who worked diligently? Apprentice, apprentices, chef, chefs

(2) The admirer of the {singer(s) | play(s)} apparently thinks the show was a big success.
Who considered the show a success? Admirer, admirers, singer|play, singers|plays

Results: Our model derives four key features of interference effects (i) ***Agreement distortion:*** A number feature mismatch between the distractor (*chefs* in 1) and the singular target (*apprentice* in 1) results in distortion of the subject's number (Figure 2A, see also [8]). (ii) ***Markedness asymmetry:*** The model also reproduces the finding that plural subjects are less affected by a mismatching distractor and the relatively high error rate with plural subjects (Figure 2B, see [9] for a similar empirical pattern in preamble repetition errors). (iii) ***Independence of agreement and lexical-root errors:*** Our model assumes separate decoding of the lexical root and the number morpheme, predicting independence of target-distractor confusion and agreement distortion. Empirical data from [7] confirms that the matching in agreement features does not modulate the probability of picking a distractor-like noun, and that semantic similarity does not modulate agreement distortion (Figure 2C, but cf [10]). (iv) ***Position similarity effects:*** Similar positions are more confusable in item-position binding. This feature of the model has the potential to account for effects of structural position on interference [11] and attraction [12].

Conclusion: We argue that a model of how feature-position bindings are maintained is required for a theory of sentence processing. We propose a model that can account for encoding interference and representational distortion - two effects that cannot be accounted for within cue-based retrieval. The current proposal also dovetails with work in psychology (STM/LTM interactions in an active, goal directed WM), deep learning (distributed representations), and linguistics (importance of relational information for syntactic dependencies).

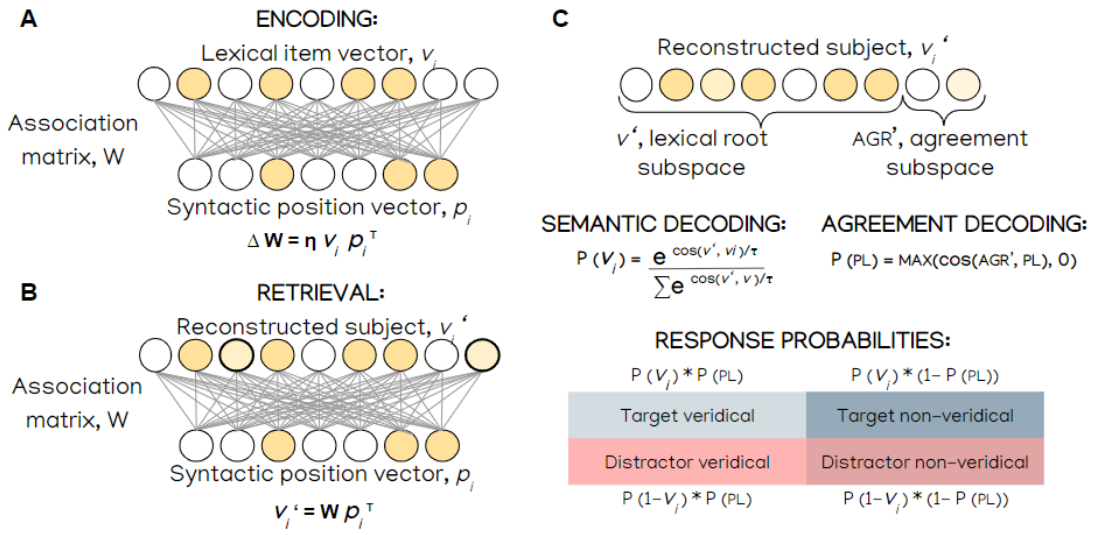

Fig 1. Model implementation. η encoding strength parameter; cos cosine similarity; τ softmax temperature

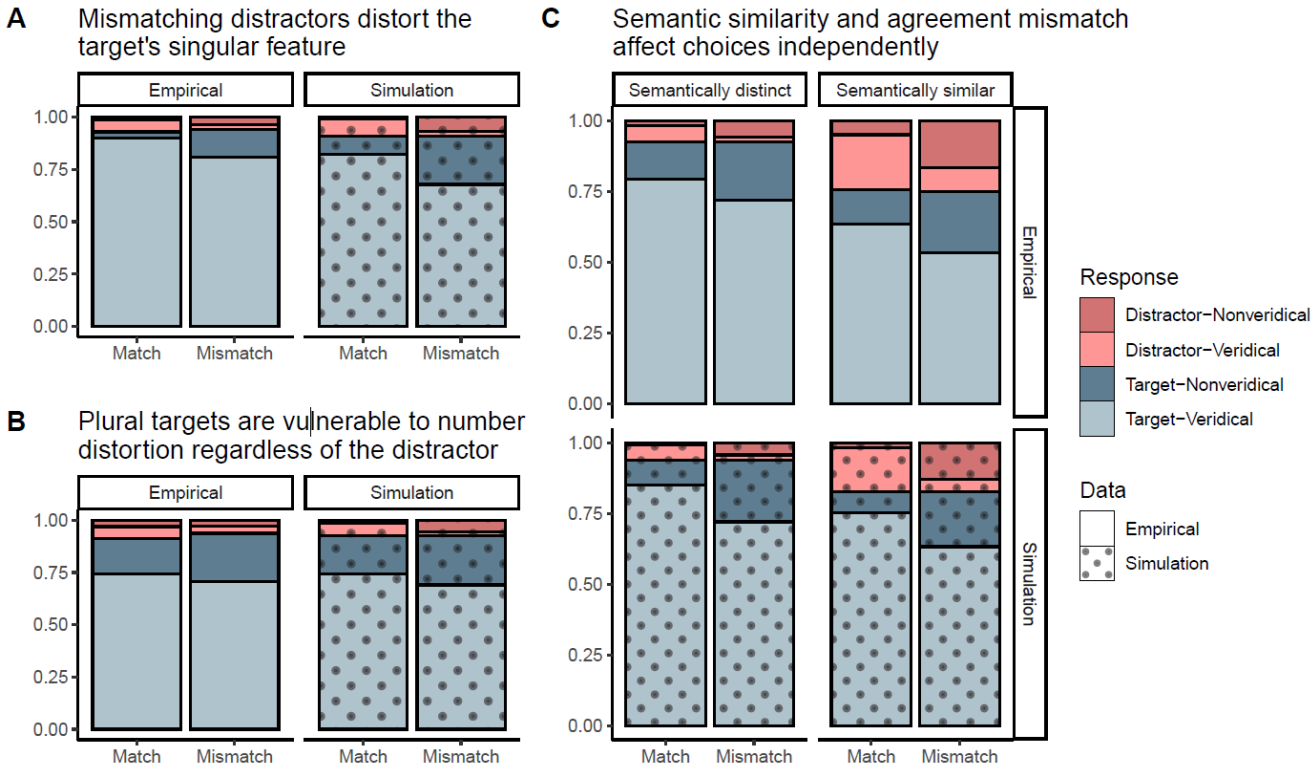

Fig 2. Simulation results vs. empirical data.

[1] Lewis, Vasishth, & Van Dyke (2006). *TiCS*. **[2]** Treisman (1996). *Current Opinion in Neurobiology*. **[3]** Smolensky, Goldrick, & Mathis (2014). *Cognitive Science*. **[4]** Farrell & Lewandowsky (2002). *Psychonomic Bulletin and Review.* **[5]** Oberauer, Lewandowsky, Farrell, Jarrold & Greaves (2012). *Psychonomic Bulletin and Review.* **[6]** Authors (in prep). **[7]** Laurinavichyute & von der Malsburg (2023). *Cognitive Science.* **[8]** Paape, Avetisyan, Lago, & Vasishth (2021). *Cognitive Science.* **[9]** Brehm, Cho, & Smolensky (2022). *Cognitive Science.* **[10]** Villata & Franck (2020). *Journal of Experimental Psychology: Learning, Memory, and Cognition*. **[11]** Van Dyke (2007). *JEP: LMC*. **[12]** Bock & Cutting (1992). *JML*.