A behavioral and computational investigation of referential modification and memory for object features

Caitlin Volante, Evgeniia Diachek, Deon T. Benton, and Sarah Brown-Schmidt
Vanderbilt University

Referential expressions (e.g., *the sleeveless dress*) often include modifiers used to uniquely identify the intended referent. The use of adjective modifiers (e.g., sleeveless) improves memory for named objects (e.g., dress) compared to using nouns only[1]. Evidence from the language processing literature demonstrates that addressees consider multiple candidate referents temporarily consistent with the unfolding noun phrase (i.e., considering both a sleeveless shirt and dress while hearing *sleeveless dress*)[2], and that these temporary considerations improve memory for those candidate referents[3]. The mechanism underlying this observed memory boost may be *referential* activation; alternatively, the modifier may boost memory for the lexicalized *feature* alone. We test three competing hypotheses. Lexicalizing one feature (e.g., *sleeveless dress*) may strengthen its representation, making it more memorable than other features (e.g., dress length). Alternatively, the lexicalized feature may serve as a retrieval cue for the other feature, boosting memory for both. Finally, lexicalizing one feature of an object may not impact memory for individual features. In a preregistered study (N=192, Prolific), participants (Ps) viewed pairs of unrelated images and heard an audio description of one (the target) in English asking them to click on it (**Fig.1**). The description was either a bare noun (e.g. *the dress*) or a pre-nominally modified expression (e.g. *the sleeveless dress*). Critical stimuli were clothing items with two critical features: one named (e.g., sleeveless), and one unnamed (e.g., long). At test, Ps viewed two images: one previously viewed at study, and one new (foil) image. Ps were asked to select the item they saw before and were told to pay attention to details. We manipulated the properties of the foil image such that it *differed* from the target in 1) the critical named feature (i.e., named feature mismatch, unnamed feature match); 2) the unnamed feature (i.e., named feature match, foil image unnamed feature mismatch); or 3) both features. **Results:** A mixed-effects logistic model of the 2AFC memory data compared accuracy across the 3 foil types. In addition to overall better memory when expressions had been modified ($p<.001$), there were significant interactions between modification at study and foil type at test (**Fig.2**). Accuracy was higher when the foil mismatched the target on both named and unnamed features vs. when it mismatched on one feature, an effect that was enhanced with modification ($p < .05$). Critically, for modified expressions, Ps were more accurate at identifying the target when the foil mismatched the target on the named feature vs. unnamed feature ($p < .0001$); this effect was not observed when the target expression was unmodified, resulting in a significant modification × foil type interaction ($p <.001$). This finding supports the hypothesis that lexicalizing one object feature improves memory for that particular feature more than other features of the object. In other words, the previously-demonstrated memory boost[3] is due to better memory for the lexicalized *feature* not the *referent* as a whole. To explore the mechanistic basis of this finding, we built a three-layer autoencoder neural network. These models are trained to recreate the input pattern of activity along the corresponding output groups through an intermediate group of hidden units[4]. We propose the underlying mechanism is based on how strongly named vs. unnamed features are encoded and subsequently decay. We implement this by giving named features increased learning rates (LR) and decreased weight decays (WD), and vice versa for unnamed features. The model was trained on 20 objects, each with a named and unnamed feature (denoted by LR and WD levels). At test, the model was presented with the training objects in addition to new objects that either matched on the named feature, matched on the unnamed feature, or mismatched on both. **Results:** The model replicates the behavioral findings (**Fig.3-4**) through a mechanism by which the lexicalization of one feature increases encoding for that particular feature whilst unnamed features decay at a faster rate compared to named features. **Discussion:** Empirical and computational findings demonstrate a modification-driven memory boost driven by feature-specific encoding and decay, consistent with claims that object features are remembered and forgotten independently[5].

Figure 1. Schematic of the experimental procedure. At study participants see two pictures (e.g., dress, giraffe) and one is named with a modified noun phrase. At test participants see two similar images (two dresses) and must identify which is the target (seen at study). Test trials always present the target and one of 3 foil types: (1) foil matches named feature and mismatches unnamed feature; (2) matches unnamed feature and mismatches named feature; (3) matches neither feature.

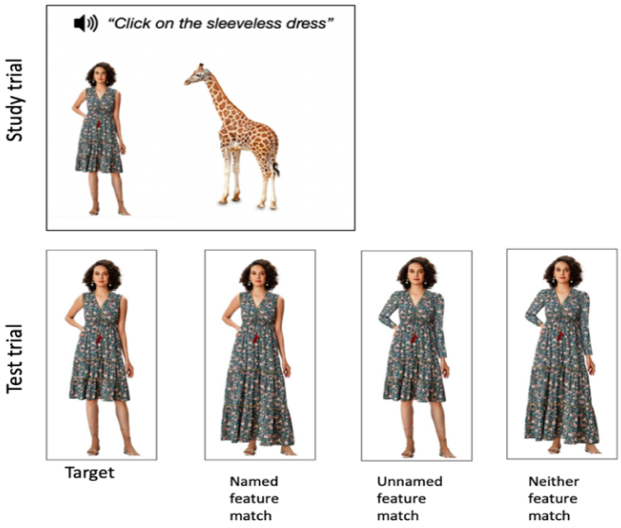
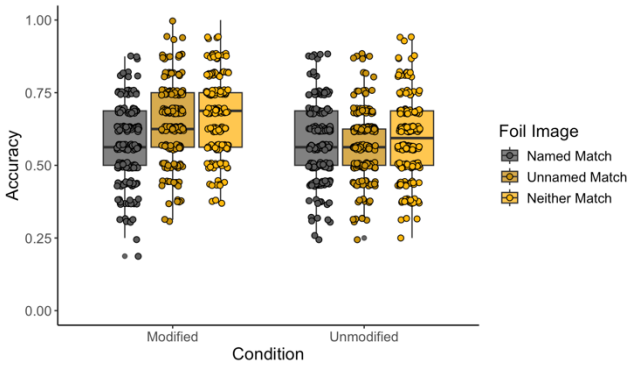


Figure 2. E2: Accuracy in 2AFC memory test. Error bars represent by-participant standard error of the mean. Data points represent mean accuracies for each participant.



References: 1. Yoon, S. O., et al., (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition, 154,* 102-117. **2.** Eberhard, et al. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, *24*, 409-436. **3.** Lord, K. & Brown-Schmidt, S. (2022). Temporary ambiguity and memory for the context of spoken language. *Psychonomic Bulletin & Review,* 29, 1440–1450. **4.** Mareschal, D. (2003). Connectionist models of learning and development in infancy. In P. T. Quinlan (Ed.), … *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp. 43–82). Psychology Press. **5.** Brady, et al. (2013). Real-world objects are not represented as bound units: independent forgetting of different object details from visual memory. *JEP: General*, *142*(3), 791.

Figure 3. Model output for the modified condition over 9 different combinations of learning rates, weight decays, and epoch lists. Models were trained for different numbers of epochs (i.e., weight updates) to ensure that the main findings were not idiosyncratic to particular training durations.

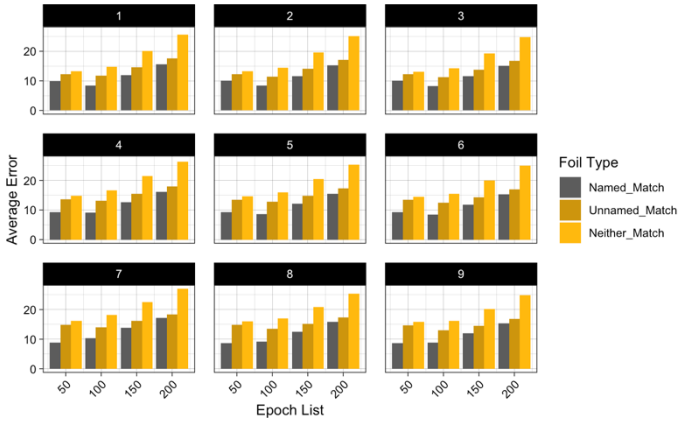


Figure 4. Model output for the unmodified condition over 9 different combinations of learning rates, weight decays, and epoch lists.

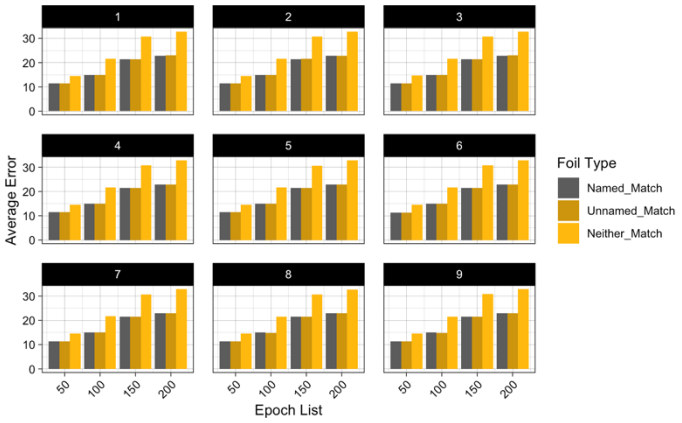


Figure 5. Schematic of the model.