Feature distortion and memory updating: Experimental and modeling evidence

Maayan Keshev and Brian Dillon

Many sentence processing theories assume that the parsing is executed over veridical memory representations. Under such approaches errors in linguistic dependency formation arise only through access to a wrong memory item [1]. In contrast, rational inference models propose that nonveridical representations are considered to overcome production errors [2] and noisy memory encodings [3-4]. We test whether memory representations are strictly veridical or whether they can be distorted and updated. We use two offline comprehension datasets from two distinct dependencies, and contrast the predictions of cue-based retrieval [1] with models that allow nonveridical memory representations and rational inference [3-4].

Datasets: In a new SUBJECT-VERB experiment (80 participants, 40 sets), we use sentences as (1). We focus on grammatical sentences to ensure that comprehension errors do not reflect an ad-hoc strategy recruited for interpretation of illicit dependencies [5]. In addition to examining target distractor feature matching and availability of agreement cues at the verb, we also manipulate the feature of the target to investigate number markedness effects.

(1) The apprentice of the chef(s) {worked | works} diligently on the souffles.
Who worked diligently? The apprentice, the apprentices, the chef, the chefs

Cue-based retrieval [1] predicts that the distractor will be occasionally interpreted as the target, when the verb carries agreement features matching the distractor (i.e. *chef works*, see Figure 1A). Noisy encoding models predict that a distractor mismatching the target (*chefs*) occasionally distorts features of the target [6] (leading to the nonveridical *apprentices*). These models can be supplemented by a rational updating process whereby features from the verb are used to reduce uncertainty about the representation of the subject [4]. This predicts that feature distortion would mostly arise when the verb doesn't carry agreement (i.e. *worked*), which could reduce uncertainty about the features of the target.

In addition, we examine a FILLER-GAP dataset from [7] with Hebrew relative clauses (1). In this dataset, the salient feature is grammatical gender rather than number, and the availability of agreement cues is manipulated through an optional resumptive pronoun.

(2) Hebrew: The manager.M hates the cashier.F that the customers like { __ | her}.
Who did customers like? The cashier.M, the cashier.F, the manager.M, the manager.F

Modeling: We implement multinomial processing trees in Stan. We fit the models separately for data from singular subject-verb dependencies (Figure 1), plural subject-verb dependencies, and filler-gap dependencies. For each case, we compare the models' predictive fit using *k*-fold cross-validation. Despite overall differences between the datasets (Figure 2), we find that the best fitting model (Table 1) is the one allowing feature distortion (due to the distractor) and updating (at an agreeing verb), in unmarked (singular) subject-verb dependencies and in filler-gap dependencies. For marked (plural) subjects, a distortion+updating model performed similarly to a simple distortion model, but better than a cue-based retrieval model.

Discussion: Our data and data from [7] are better accounted for by models where memory representations can be distorted and updated. This extends previous findings of nonveridical agreement representations in final interpretation [8-10]. A recurring finding in these studies is that ungrammatical verbs distort the memory of the target subject, independently of the distractor. We provide evidence that this arises in grammatical dependencies. Following [4], we propose that the pattern of comprehension errors is best captured as a combination of feature distortion models of attraction [6] and rational inference whereby comprehenders use additional agreement sites (where such sites are available) to offset memory interference.

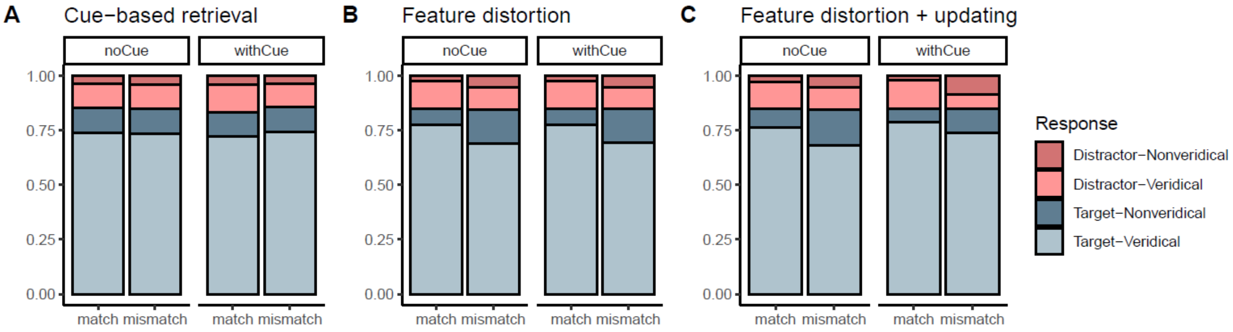

Figure 1. Posterior-based model predictions. The predicted response patterns are derived from models fitted to the singular subject-verb dataset. Similar results are produced for other model fits.

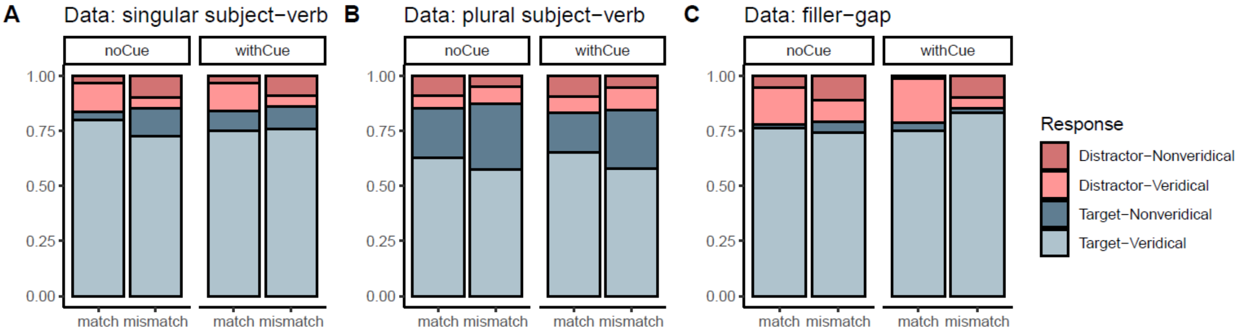

Figure 2. Empirical data from interpretation of the current subject-verb comprehension experiment and the relative clause comprehension from [5].

	Subject-verb dependencies		Filler-gap dependencies
	Singular subjects	Plural subjects	
Distortion+updating vs. Distortion only:	**Δeldp=21.3, SE=7.4**	Δeldp=3.0, SE=1.8	**Δeldp=45.4, SE=10.07**
Distortion+updating vs. Cue-based retrieval:	**Δeldp=42.1, SE =11.4**	**Δeldp=11.9, SE=5.45**	**Δeldp=53.6, SE=13.8**

Table 1. Results k-fold cross-validation. Δelpd is the difference between the models' predictive fit, measured by the deviation between predictions of the fitted model and the heldout data (expected log pointwise predictive density). Positive values reflect an advantage of the distortion+updating model. We interpret an advantage as reliable when the Δelpd value is over 2 SE (standard errors). Reliable contrasts are marked in the table in boldface.

References: **[1]** Lewis, & Vasishth (2005). *CogSci.* **[2]** Gibson, Bergen, & Piantadosi (2013). *PNAS*. **[3]** Futrell, Gibson, & Levy, (2020). *CogSci*. **[4]** Keshev & Meltzer-Asscher (2024). *JML.* **[5]** Molinaro, Kim, Vespignani, & Job (2008). *Cognition* **[6]**. Eberhard, Cutting, & Bock, (2005). Psychological Review. **[7]** Koesterich, Keshev, Shamai, & Meltzer-Asscher (2021). *HSP talk.* **[8]** Paape, Avetisyan, Lago, & Vasishth (2021). *CogSci*. **[9]** Brehm, Jackson, & Miller (2019). *Quarterly Journal of Experimental Psychology*. **[10]** Patson & Husband (2016). *Quarterly Journal of Experimental Psychology*.