Cue-based retrieval mechanism cannot capture illusions of plausibility

Novak (Xingyu) Shi (U of Oxford), Colin Phillips (U of Oxford; U of Maryland College Park)

Recent studies have revealed a striking parallel between agreement attraction in ungrammatical sentences [1] and semantic attraction in implausible sentences (as in (1)), which has been attributed to a unified cue-based retrieval account [2-4]. However, for the latter, the assumptions of binary and fixed categorisations of verb-based retrieval cues like [+tippable] (parallel to [+plural]) are not clearly scalable accounts of memory encoding and retrieval for semantic memory. In view of this challenge, this study probed into the causes of illusions of plausibility by testing a finer gradation of the relation between verbs and potential semantic attractors: (a) probable distractors, i.e., plausible and probable arguments of the verb (e.g., “waitress”); (b) plausible distractors, i.e., plausible but improbable (e.g., “singer”); (c) partial feature-matching distractors, which respect some key features, e.g., *animacy*, but not all features required by the verb (e.g., “drinker”). Using a speeded forced-choice comprehension task, we found that only the probable distractors and the plausible distractors induced attraction in implausible sentences, with effects that were comparable regardless of probability. By contrast, partial feature-matching distractors patterned together with the implausible baseline distractors, suggesting that plausibility really was the key factor rather than feature-matching, which argues against a cue-based mechanism where plausibility as a relational notion cannot be encoded.

Participants were English native speakers recruited via Prolific (N = 58). We followed prior work to focus on verb-object relations in a filler-gap dependency, manipulating the plausibility (plausible vs. implausible target) and distractor type (probable vs. plausible vs. partial feature-matching vs. implausible) using 48 item sets (Table 1). Probable distractors were constructed by selecting the modal response (i.e. the one with the highest cloze probability) extracted from the RoBERTa Large model, with sentence preambles minimally different to the experimental sentences. All distractors and targets were then verified in an independent plausibility rating study (N = 48) (Table 2). In the main experiment, sentences were shown word-by-word in RSVP, with the critical verb always appearing at the end and turning green. Participants were required to judge whether the green word was a good/sensible continuation by responding yes or no. Accuracy data were analyzed in logistic linear mixed effects models using simple contrast coding, with the implausible distractor in the implausible sentence as the reference level.

Results revealed a significant main effect of Plausibility on accuracy rates, with implausible sentences being judged less accurately than plausible sentences (p < .001). Importantly, this effect of Plausibility differed across Attractor types: it was stronger for the probable distractors (*Est.* = 1.82, z = 4.3, p < .001) and for the plausible distractors (*Est.* = 0.96, z = 2.4, p < .05) than the implausible baseline distractors, but it did not differ between the baseline and the partial feature-matching distractors (p = .5) (Fig. 1). This suggests that the partial feature-matching distractors patterned together with the implausible baseline to contrast with the other two types, inducing no attraction effect. Post-hoc analysis revealed a non-significant difference between the probable distractors and the plausible distractors, indicating that the probability difference did not modulate the degree of attraction (p = .98) (Fig. 1).

Taken together, the results argue against a cue-based retrieval account of illusions of plausibility, because (1) even animacy match failed to induce attractions, and (2) plausibility as a relational notion cannot be inherently stored during memory encoding. Moreover, the comparable attraction effects by the probable distractors and the plausible distractors ruled out a simple form-to-form retrieval process that operates solely under verb-based probability.

(1) ?The pub owner noticed the *beer* that the customer near the *waitress* had generously **tipped.**

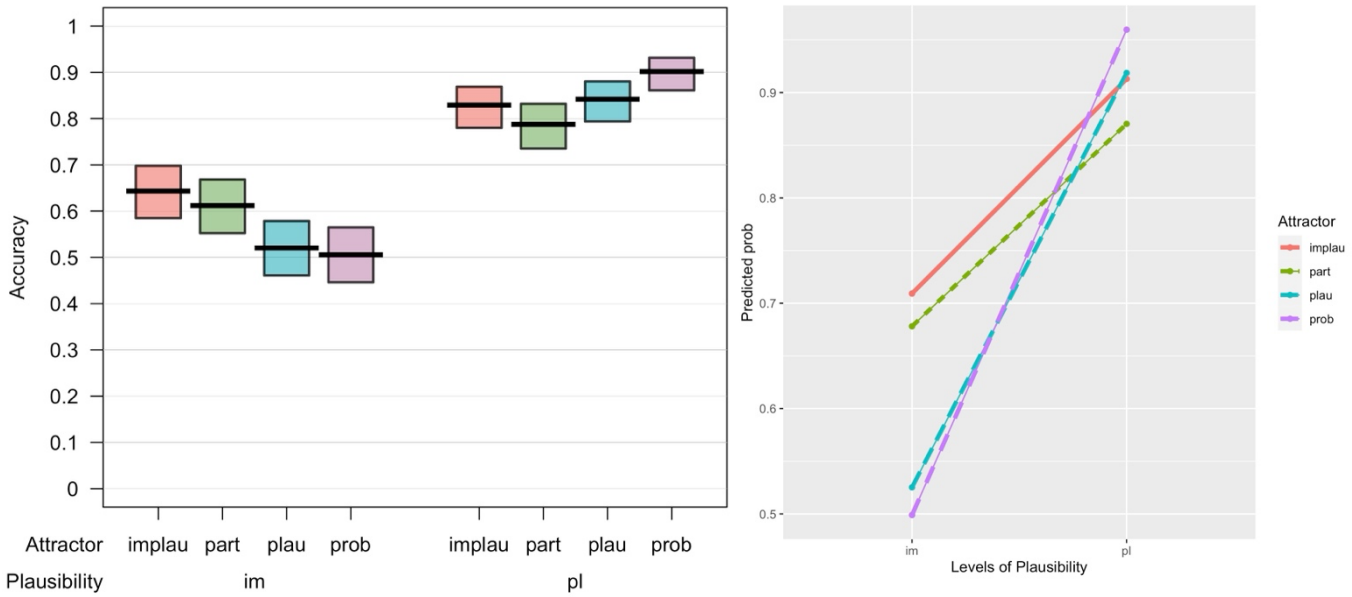
Table 1. Critical conditions

Conditions	Sample sentence
Plausible vs. Implausible target, Probable distractor	The pub owner was at the pub. She noticed the *bartender (vs. beer)* that the customer near the *waitress* had generously **tipped**
Plausible vs. Implausible target, Plausible distractor	The pub owner was at the pub. She noticed the *bartender (vs. beer)* that the customer near the *singer* had generously **tipped**
Plausible vs. Implausible target, Partial feature-matching distractor	The pub owner was at the pub. She noticed *the bartender (vs. beer)* that the customer near the *drinker* had generously **tipped**
Plausible vs. Implausible target, Implausible distractor	The pub owner was at the pub. She noticed the *bartender (vs. beer)* that the customer near the *chair* had generously **tipped**

Table 2. Item statistics

Item type	Example	Plausibility (Mean)	Cloze probability (Mean)
Probable (prob)_distractor	*waitress*	*6.30*	33.3%
Plausible (plau)_distractor	*singer*	*5.95*	<1%
Partial feature-matching (part)_distractor	*drinker*	*4.03*	<1%
Implausible (implau)_distractor	*chair*	*1.93*	<1%
Targets (implausible/plausible)	*beer/bartender*	*2.0 / 6.27*	/

Figure 1. Judgement accuracy rates (left) and interaction plot (right)



References: [1] Wagers, Lau, & Phillips. (2009). Agreement attraction in comprehension: Representations and processes. *JML.* [2] Cunnings & Sturt. (2018). Retrieval interference and semantic interpretation. *JML.* [3] Fujita, & Cunnings. (2022). Interference and filler-gap dependency formation in native and non-native language comprehension: JEP: LMC. [4] Laurinavichyute & von der Malsburg. (2022). Semantic attraction in sentence comprehension. *Cognitive Science*.