CONTRASTING FACILITATION PROFILES FOR AGREEMENT AND REFLEXIVES REVISITED: A LARGE-SCALE EMPIRICAL EVALUATION OF THE CUE-BASED RETRIEVAL MODEL

Lena A. Jäger (University of Potsdam), Daniela Mertzen (University of Potsdam), Julie Van Dyke (Haskins Laboratories), & Shravan Vasishth (University of Potsdam)
lena.jaeger@uni-potsdam.de

Even though both (1a) and (1b) are ungrammatical, reading studies consistently report faster reading times at the auxiliary verb *were* in (1b) vs. (1a).

1a. *The bodybuilder who worked with the personal trainer were competitive.
1b. *The bodybuilder who worked with the personal trainers were competitive.

The Lewis & Vasishth 2005 cue-based retrieval model of sentence processing (LV05) [1] correctly predicts faster reading times at the auxiliary verb *were* in (1b) vs. (1a): The plural marking on the auxiliary verb *were* triggers a retrieval of a plural-marked subject, but occasionally the plural-marked distractor noun *trainers* is misretrieved in (1b) vs. (1a). The LV05 model predicts a facilitation effect of approximately -26 ms; constrained variation of parameters can lead to mean predicted effects ranging from -10 ms to -57 ms (see Fig. 2 in [4]).

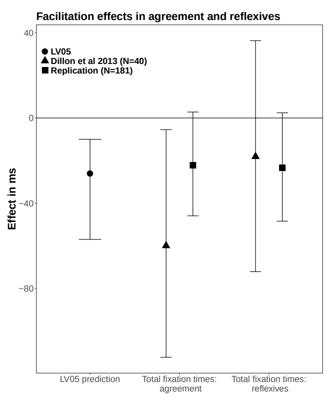
In an eyetracking study, Dillon et al. [2] showed that faster total reading times are seen at the auxiliary in (1b) vs. (1a), as predicted by the LV05 model (see figure below). A reanalysis of [2]'s data using a maximal Bayesian linear mixed model shows that the estimated mean facilitation in their data is -60 ms, with a 95% probability that the facilitation effect lies between -112 and -5 ms (this is the so-called 95% credible interval).

Interestingly, Dillon and colleagues also showed that a similar configuration, antecedent-reflexive dependencies, for which the LV05 model predicts similar facilitation effects as for subject-verb agreement, shows no facilitation effects at all at the reflexive *themselves*: our Bayesian linear mixed model reanalysis showed a mean total reading time of -18 ms, 95% credible interval $[-72, 36]$.

2a. *The bodybuilder who worked with the personal trainer injured themselves.
2b. *The bodybuilder who worked with the personal trainers injured themselves.

Dillon et al. argue that reflexives are immune to misretrieval effects because binding theory's Principle A acts as a filter, allowing misretrieval-free and deterministic access to the antecedent. However, [2] had a relatively small sample size (N=40). Using the largest LV05-predicted effect size (-57 ms) and the standard error estimate from [2], the probability of detecting an effect correctly with 40 participants is 30%. When power is this low, many null results will be found and any statistically significant estimate (e.g., the facilitation in (1b) vs. (1a)) will *always* be exaggerated [3,5]. This is because the standard error is so large that under repeated sampling, the effect estimates will fluctuate, hence any estimate close to the true mean will not cross the significance threshold [3]. Crucially, both the agreement and reflexive effects in [2] have such wide uncertainty intervals that the LV05 model's predictions are fully compatible with them (see figure).

Accurate estimates with narrower credible intervals can only be obtained with larger sample studies [3,5]. We therefore conducted a direct replication of [2]'s eyetracking study, but with a larger participant sample size (N=181); for a predicted effect of -57 ms, power is now 88% (we also had grammatical controls in the experiment, as in [2], but these are not discussed here due to space constraints). Here, both the agreement (1a,b) and reflexives (2a,b) show similar facilitation effect estimates in total reading times, closer to the magnitude predicted by the LV05 model: agreement: -22 ms $[-46, 3]$; reflexives -23 ms $[-48, 3]$.



In sum, both agreement and reflexive dependencies seem to show similar facilitation profiles, consistent with the predictions of the LV05 model. More generally, this work demonstrates the importance of conducting larger-sample studies in order to obtain more precise estimates for evaluating predictions of quantitative models. **References**: [1] Lewis and Vasishth, 2005, Cog Sci. [2] Dillon et al, 2013, JML. [3] Gelman and Carlin, 2014, PPS. [4] Engelmann et al. 2018, https://osf.io/b56qv/ [5] Vasishth et al, 2018, https://osf.io/p9baz/