Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study

Lena A. Jäger
Department of Linguistics and Institute for Computer Science, University of Potsdam, Germany

Daniela Mertzen
Department of Linguistics, University of Potsdam, Germany

Julie A. Van Dyke
Haskins Laboratories, New Haven, CT, United States

Shravan Vasishth
Department of Linguistics, University of Potsdam, Germany

April 11, 2019

Abstract

Cue-based retrieval theories in sentence processing predict two classes of interference effect: (i) *Inhibitory interference* is predicted when multiple items match a retrieval cue: cue-overloading leads to a decrease in the probability of retrieving the correct target; and (ii) *Facilitatory interference* arises when a retrieval target as well as a distractor only partially match the retrieval cues; this partial matching leads to an overall speedup in retrieval time. Inhibitory interference effects are widely observed, but facilitatory interference apparently has an exception: reflexives have been claimed to show no facilitatory interference effects. Because the finding is based on underpowered studies, we conducted a large-sample experiment that investigated both facilitatory and inhibitory interference. In contrast to previous studies, we find facilitatory interference effects in reflexives. We also present a comprehensive quantitative evaluation of a cue-based retrieval model (Engelmann et al., 2019, available as a Shiny App: https://engelmann.shinyapps.io/inter-act/) with respect to the large-sample data.

Keywords: cue-based retrieval; sentence processing; similarity-based interference; reflexives; agreement; Bayesian data analysis; replication

Please send correspondence to lena.jaeger@uni-potsdam.de.

Introduction

What are the constraints on linguistic dependency formation in online sentence comprehension? This has been a central theoretical question in psycholinguistics. Inspired by research in cognitive psychology, constraints on working memory have been invoked to explain how the human sentence parsing system works out who did what to whom. For example, when a verb is read or heard, what mechanism does the parsing system use to identify the subject and object of the verb? A widely accepted view (Lewis, Vasishth, & Van Dyke, 2006; McElree, 2003; Van Dyke & Lewis, 2003) is that a cue-based retrieval mechanism drives this dependency completion process. When a dependency needs to be completed, the cue-based retrieval account assumes that certain features (*retrieval cues*) are used to retrieve the co-dependent item, the *retrieval target*, from memory. An important consequence of such a cue-based retrieval mechanism is that whenever other items, called *distractors*, also match some or all of the retrieval cues, *similarity-based interference* can arise.

As an example of similarity-based interference, consider the subject-verb dependency shown below in 1. This set of sentences is taken from Dillon, Mishler, Sloggett, and Phillips (2013). Following the convention in Engelmann, Jäger, and Vasishth (2019), we show retrieval cues in curly braces, and binary-valued features on nouns that match or mismatch the retrieval cues.

(1) a. *Agreement; grammatical; interference*
The amateur bodybuilder$^{+singular}_{+local\ subject}$ who worked with the personal trainer$^{+singular}_{-local\ subject}$ amazingly was$\{^{singular}_{local\ subject}\}$ competitive for the gold medal.

b. *Agreement; grammatical; no interference*
The amateur bodybuilder$^{+singular}_{+local\ subject}$ who worked with the personal trainers$^{-singular}_{-local\ subject}$ amazingly was$\{^{singular}_{local\ subject}\}$ competitive for the gold medal.

In these sentences, the dependency of interest is the one between the main clause verb *was* and its subject *the amateur bodybuilder*. Consistent with evidence suggesting that focal attention is highly limited (e.g., McElree, 2006), the distal subject must be retrieved from memory when the verb is encountered. Simplifying somewhat, we assume that the verb uses two cues, number and local-subject status, to search for the retrieval target (i.e., the subject). Because of the perfect match between the retrieval cues and the target, the sentences are grammatical.

In 1a, one of these retrieval cues, the singular number feature, matches not only with the main-clause singular subject but also with the distractor, the singular noun inside the relative clause, *the personal trainer*. By contrast, in 1b, this distractor noun is plural-marked (*the personal trainers*) and so does not match the number retrieval cue. The situation in 1a, where both the target and the distractor noun (partially) match the retrieval cues, is referred to as *cue overload*. This cue overload leads to interference, which is expressed as a slowdown at the verb (where the subject must be retrieved) in reading time in self-paced reading and eyetracking experiments (Van Dyke, 2007; Van Dyke & Lewis, 2003; Van Dyke & McElree, 2011). Following Dillon (2011), we will refer to this slowdown as *inhibitory interference*.

Interference due to cue-overload is a key prediction of cue-based retrieval models of sentence processing (Lewis & Vasishth, 2005; McElree, 2000; Van Dyke, 2007; Van Dyke

& Lewis, 2003; Van Dyke & McElree, 2011). A computationally implemented model that predicts such inhibitory interference effects is the cue-based retrieval model of Lewis and Vasishth (2005) (henceforth LV05).[1] This model was developed within the general cognitive architecture, Adaptive Control of Thought-Rational (ACT-R, Anderson et al. 2004). Cue-based retrieval models can explain interference effects (Dillon et al., 2013; Jäger, Engelmann, & Vasishth, 2015; Kush & Phillips, 2014; Nicenboim, Logačev, Gattei, & Vasishth, 2016; Nicenboim, Vasishth, Engelmann, & Suckow, 2018; Parker & Phillips, 2016, 2017; Patil, Vasishth, & Lewis, 2016; Vasishth, Bruessow, Lewis, & Drenhaus, 2008), but they have also been invoked in connection with a range of other issues in sentence processing: the interaction between predictive processing and memory (Boston, Hale, Vasishth, & Kliegl, 2011), impairments in individuals with aphasia (Mätzig, Vasishth, Engelmann, Caplan, & Burchert, 2018; Patil, Hanne, Burchert, Bleser, & Vasishth, 2016), the interaction between oculomotor control and sentence comprehension (Dotlačil, 2018; Engelmann, Vasishth, Engbert, & Kliegl, 2013), the processing of ellipsis (Martin & McElree, 2009; Parker, 2018), the effect of working memory capacity differences on underspecification and "good-enough" processing (Engelmann, 2016; von der Malsburg & Vasishth, 2013), and the interaction between discourse/semantic processes and cognition (Brasoveanu & Dotlačil, 2019). The source code of the model is available from https://github.com/felixengelmann/inter-act; and quantitative predictions can be derived graphically using the Shiny App available from https://engelmann.shinyapps.io/inter-act/.

Inhibitory interference arises in the LV05 model as a consequence of the spreading activation assumption inherent in the ACT-R architecture: multiple items (e.g., the target noun and the distractor noun in 1a above) match a retrieval cue, leading to an activation penalty on each item, increasing average retrieval time. The linguistic context that leads to inhibitory interference is illustrated schematically in the upper part of Figure 1.

In addition to inhibitory interference, cue-based retrieval also predicts a so-called facilitatory interference effect in specific situations: when no retrieval candidate fully matches the retrieval cues, and a distractor is present that partially matches the retrieval cues, a race situation arises and leads to an overall speedup in reading time (Engelmann et al., 2019; Logačev & Vasishth, 2015). The assumption in ACT-R is that when a retrieval attempt is initiated, all partial matches become candidates for retrieval, and the item which happens to have a higher activation in a particular trial gets retrieved. Whenever such a race condition holds, average reading times will be as fast or faster compared to when no race condition occurs. When the finishing times of both processes are similar, the average finishing time will be faster; and when one process has a much faster finishing time than the other, the average finishing time will follow the distribution of the faster process. This is illustrated in Figure 2. Also see Logačev and Vasishth (2015) for a detailed exposition.

There is considerable evidence for this kind of facilitatory interference effect in sentence processing. For example, Dillon et al. (2013) showed that in sentences like 2a vs. 2b, mean reading time at the main clause verb *were* was faster by -119 ms (95% confidence interval of [-205, -33] ms). These sentences are ungrammatical because the subject does not match the

[1] We derive specific predictions from the LV05 model, as its computational implementation (code available from: https://github.com/felixengelmann/inter-act) allows us to quantitatively compare model predictions with empirical data. However, in principle a variety of implementations of this theory are possible, and the LV05 model represents only one of these.

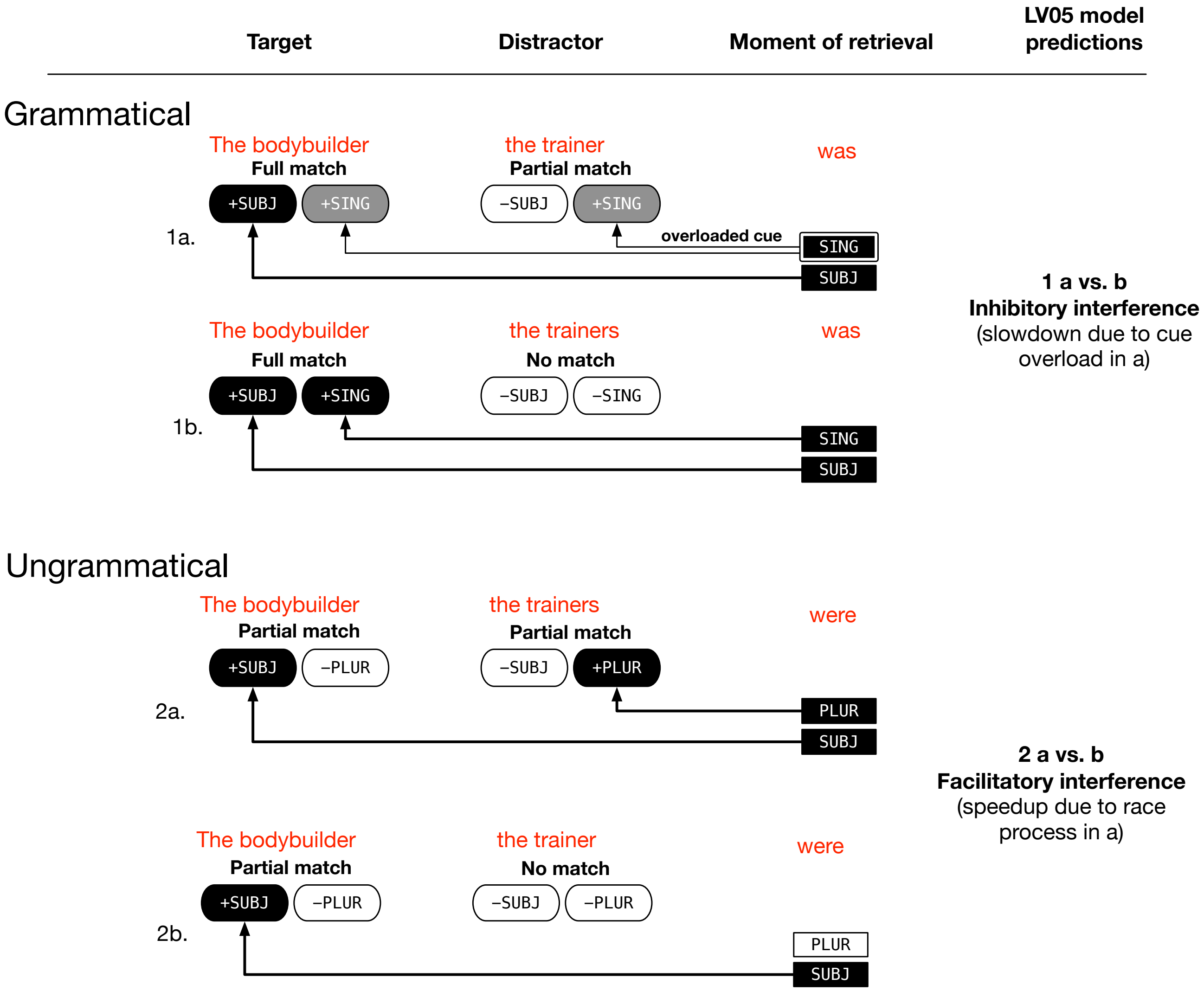


Figure 1. A schematic figure illustrating inhibitory and facilitatory interference in the Lewis & Vasishth, 2005 cue-based retrieval model. The figure is adapted from Engelmann, Jäger, & Vasishth, 2018.

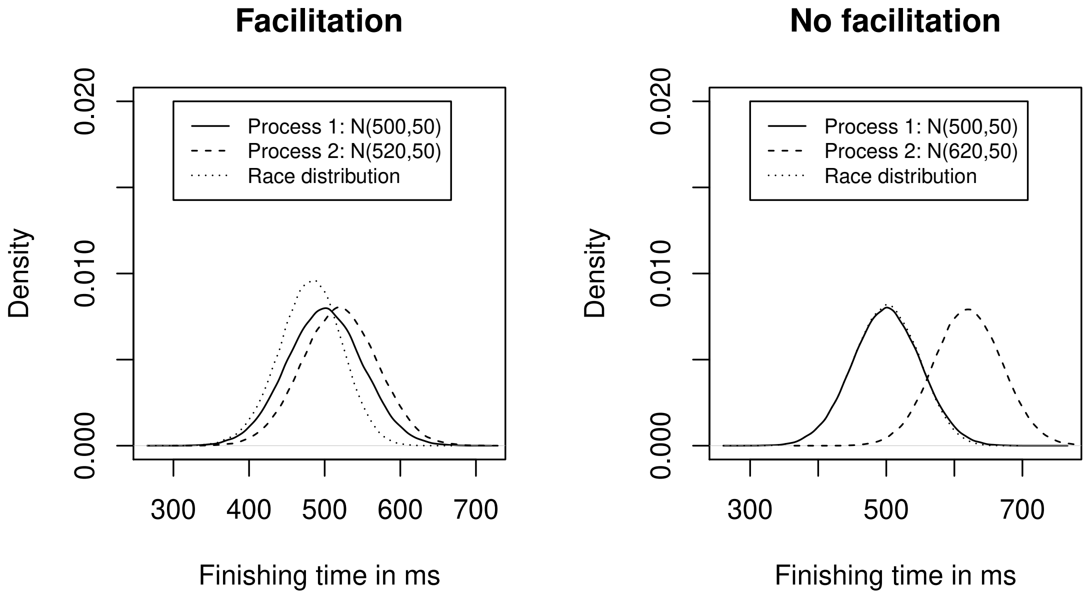


Figure 2. An illustration of a race process involving two distributions. When the two distributions have similar means (left-hand side figure), the distribution of the values from a race process will lead to an overall facilitation. When one distribution has a much smaller mean (right-hand side figure), the distribution of the race process will have the same distribution as the distribution with the smaller mean.

matrix verb's number marking; under ACT-R assumptions, the race situation arises in 2a because a distractor noun phrase matches the number marking on the verb.[2]

(2) a. *Agreement; ungrammatical; interference*
*The amateur bodybuilder$^{-plural}_{+local\ subject}$ who worked with the personal trainers$^{+plural}_{-local\ subject}$ amazingly were$\{^{plural}_{local\ subject}\}$ competitive for the gold medal.

b. *Agreement; ungrammatical; no interference*
*The amateur bodybuilder$^{-plural}_{+local\ subject}$ who worked with the personal trainer$^{-plural}_{-local\ subject}$ amazingly were$\{^{plural}_{local\ subject}\}$ competitive for the gold medal.

Such facilitatory effects have been found in self-paced studies on subject-verb number agreement; for English, see Wagers et al. (2009), and for Spanish, see Lago et al. (2015). In eyetracking data, semantic plausibility manipulations (Cunnings & Sturt, 2018) also show a facilitatory effect in total fixation time that can be explained in terms of a race process.

Although the bulk of research in the cue-based retrieval tradition supports the predictions of inhibitory and faciltatory interference, there is one apparent counterexample. Consider the sentences shown in 3a vs. 3b. The sentence 3a has the same characteristics as the ungrammatical subject-verb construction 2a discussed earlier: the subject (*the amateur bodybuilder*) matches only some of the retrieval cues on the reflexive; it does not match the number cue. In dependencies such as subject-verb agreement, non-structural cues like number are assumed to be used in addition to syntactic cues. Consequently, the phrase *the personal trainers* is a distractor and will be retrieved in some proportion of trials, leading to a race condition, according to the LV05 model. However, as we discuss below in detail, Dillon et al. (2013) only found facilitatory interference in subject-verb dependencies; they did not find evidence for facilitatory interference in antecedent-reflexive constructions. Their explanation for this asymmetry between the two dependency types is based on a proposal by Sturt (2003) according to which, in reflexives, Principle A of the binding theory (Chomsky, 1981) is used exclusively for seeking out the antecedent.[3] Thus, Principle A acts as a filter that allows the parser to unerringly identify the antecedent even if distractors are present.

(3) a. *Reflexive; ungrammatical; interference*
*The amateur bodybuilder$^{-plural}_{+\ c\text{-}com}$ who worked with the personal trainers$^{+plural}_{-\ c\text{-}com}$ amazingly injured themselves$\{^{plural}_{c\text{-}com}\}$ on the lightest weights.

b. *Reflexive; ungrammatical; no interference*
*The amateur bodybuilder$^{-plural}_{+\ c\text{-}com}$ who worked with the personal trainer$^{-plural}_{-\ c\text{-}com}$ amazingly injured themselves$\{^{plural}_{c\text{-}com}\}$ on the lightest weights.

Dillon et al. (2013)'s conclusion about an asymmetry between subject-verb dependencies and antecedent-reflexive dependencies has important theoretical consequences because it

[2] There are other explanations for these facilitatory effects; see, for example, Wagers, Lau, and Phillips (2009) and Lago, Shalom, Sigman, Lau, and Phillips (2015) for further discussion.

[3] In the original proposal by Sturt (2003), a distinction was made between early and late processes: the privileged role of the grammatical constraint was assumed to apply only in early measures in eyetracking data. However, this early-late distinction seems to have been abandoned in later work. We return to this point in the General Discussion.

implies that fundamentally different memory operations may be associated with particular linguistic contexts. Dillon et al. (2013)'s conclusions are supported by Kush (2013) and Cunnings and Sturt (2014), who argue that structural cues are weighted higher than non-structural cues in reflexive-antecedent dependencies. A related finding was made by Van Dyke and McElree (2011), who observed that in (non-agreement) subject-verb dependencies as well, structural cues such as subjecthood are weighted higher than semantic cues. Dillon et al. (2013)'s Experiment 1 is unique in that it is the only within-participants sentence comprehension study that directly compares the two dependency types.

Because of this theoretical significance of Dillon et al. (2013)'s conclusions, we felt that it is important to establish a strong empirical basis for the associated claims. The total fixation time results from Experiment 1 of Dillon et al. (2013) had a number of statistical issues, which we explain next. These issues motivated us to attempt a direct replication of their study.

The first issue was the possibly low statistical power in the Dillon et al. Experiment 1. Taking the quantitative predictions of the LV05 cue-based retrieval model as a guide, we see that the study had prospective power ranging from 7 to 43% for agreement, and from 5 to 51% for reflexives (see Appendix A). Low power has two adverse consequences: as discussed in Hoenig and Heisey (2001) and Gelman and Carlin (2014), null results will often be found even when the null hypothesis is false, and statistically significant results will have exaggerated estimates (so-called Type M error) or even have the wrong sign (so-called Type S error). As discussed in Jäger, Engelmann, and Vasishth (2017), low power has been a common problem in previous studies on interference. It would therefore be valuable and informative to run as high-powered a replication attempt as logistically feasible of Dillon et al.'s Experiment 1. A second worry in the total fixation time results of Dillon and colleagues is that a dependency × interference interaction must be shown in order to argue for a difference in the interference profiles of the two dependency types. Statistically, it is not sufficient to show that a significant facilitatory interference effect is seen in subject-verb dependencies and no significant facilitatory interference effect (i.e., a null result) is seen in reflexive conditions. Such an interaction was tested for the two dependency types in Dillon et al. (2013), but no significant interaction was found. This issue—not establishing that an interaction exists—is apparently a common problem in published work in psychology and related areas. For example, Nieuwenhuis, Forstmann, and Wagenmakers (2011) reviewed 513 neuroscience articles published in top-ranking journals and showed that the authors of more than half of these studies argued for a difference between two pairs of conditions without demonstrating that an interaction holds. Given these concerns, in order to evaluate the predictions of cue-based retrieval theory and to obtain accurate estimates of facilitatory interference effects (if any) in subject-verb dependencies vs. antecedent-reflexive dependencies, it seems vitally important to conduct a higher-power direct replication attempt of the central claims in the Dillon et al. (2013) paper.

We had two related goals in this paper. First, we wanted to establish whether in ungrammatical configurations, a principled difference between reflexives and agreement can be observed in total fixation times such that agreement shows the predicted facilitation whereas reflexives show no sensitivity to the interference manipulation, as was claimed by Dillon et al. (2013). Second, we were interested in comparing cue-based retrieval theory's predictions with the total fixation time data in Dillon et al. (2013)'s original study and in our replication attempt. We were specifically interested in comparing the model predictions to the observed interference patterns in grammatical and ungrammatical conditions.

Towards this end, we begin by presenting quantitative predictions of the Lewis and Vasishth (2005) model. Then, we explain how these predictions will be evaluated against data. Finally, we re-analyze the original data of Dillon et al. (2013)'s Experiment 1 as well as our large-sample replication data to obtain quantitative estimates of interference effects and their interaction with dependency type.[4]

Deriving quantitative predictions from the Lewis and Vasishth (2005) model

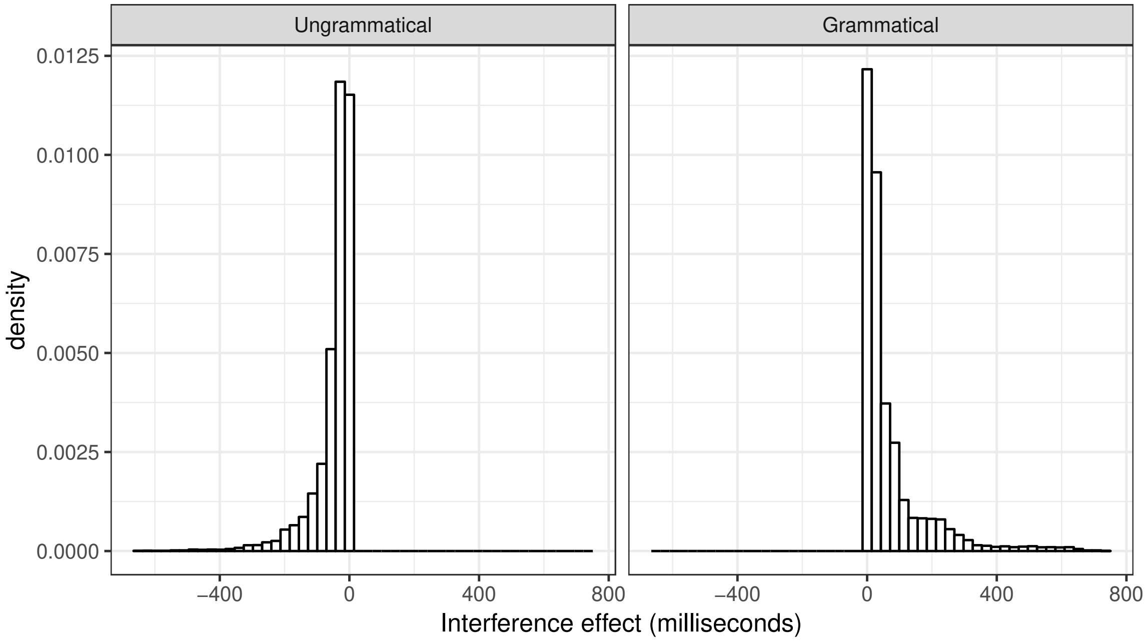


Figure 3. Predictions of the Lewis & Vasishth, 2005 ACT-R cue-based retrieval model for interference effects caused by a cue-matching distractor in sentences with a fully matching target (i.e., grammatical sentences) as in conditions a,b of Example 1, and an only partially cue-matching target (i.e., ungrammatical sentences) as in conditions a,b of Example 2. The simulations are from Engelmann, Jäger, & Vasishth, 2018. The histograms show the distributions of the predicted interference effects in grammatical vs. ungrammatical conditions, for different parameter combinations. The parameter ranges were as follows: latency factor F $\in \{0.05, 0.06, ..., 0.6\}$; the noise parameter ANS $\in \{0.1, 0.2, 0.3\}$; maximum associative strength MAS $\in \{1, 2, 3, 4\}$; mismatch penalty MP $\in \{0, 1, 2\}$; and the retrieval threshold $\theta \in \{-2, -1.5, ..., 0\}$.

We derive quantitative predictions of the LV05 cue-based retrieval model from simulations conducted by Engelmann et al. (2019). They provide a detailed investigation of the range of predictions the model makes for exactly the kind of interference experiment designs shown in Examples 1 and 2. Engelmann et al. simulate mean interference effect

[4] We thank Brian Dillon for generously sharing his original data with us.

sizes for grammatical and ungrammatical interference configurations using a grid of different parameter configurations (see Figure 2 of Engelmann et al. 2019). Here, we assume equal weighting of all cues. Figure 3 summarizes the results of these simulations.

In order to derive reasonably constrained model predictions, we computed the median together with the first and the third quartile of all interference effect sizes that are predicted under any specific parameter combination considered. For grammatical conditions, which correspond to cases where the target fully matches the retrieval cues, the median of the effects predicted by the model is 28 ms, the first quartile is 10 ms and the third quartile is 80 ms. For ungrammatical conditions, which correspond to the cases where a race situation arises, the median of the predicted effects is -26 ms, the first quartile is -57 ms and the third quartile is -10 ms. The reader can independently reproduce model predictions under different parameter settings using this shiny app: https://engelmann.shinyapps.io/inter-act/.

How can we evaluate this range of model predictions against empirical data? We turn to this question below.

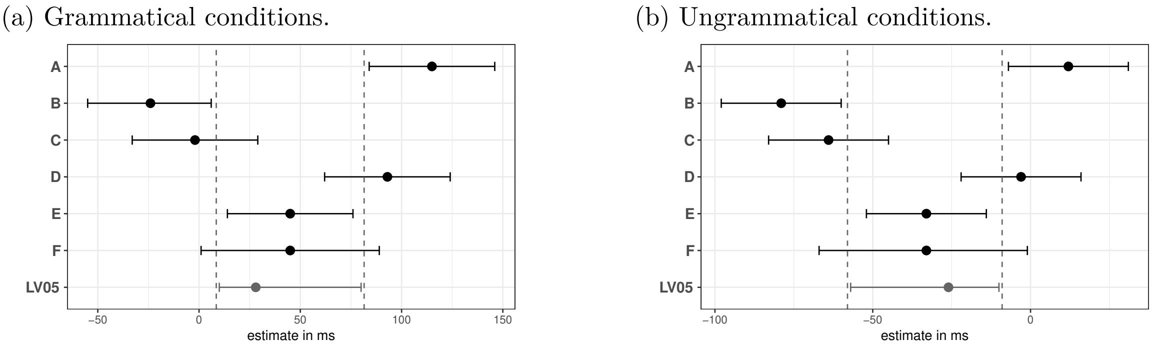
Model evaluation

We adopt the so-called region of practical equivalence (ROPE) approach for model evaluation (Freedman, Lowe, & Macaskill, 1984; Hobbs & Carlin, 2008; Kruschke, 2015; Spiegelhalter, Freedman, & Parmar, 1994). The ROPE approach has the advantage that it places the focus on the uncertainty of the data against the uncertainty of the model's predictions. This implements the proposal in the classic article by Roberts and Pashler (2000) on model evaluation, which states that both model and data uncertainty should be considered when assessing the quality of a model fit.

In the ROPE approach, a range of effect sizes that are predicted by the theory is defined. Such a range can be obtained by (constrained) variation of the free parameters of a model, as we did above. Then, the data are collected with as much precision as is logistically and financially feasible; the goal is to obtain a Bayesian 95% credible interval of the effect of interest such that it is either smaller than or as small as the width of the predicted range from the model. The 95% credible interval demarcates the range over which we can be 95% certain that the true value of the parameter of interest lies, given the data and the statistical model. This is very different from the interpretation of the frequentist confidence interval (Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

The data will be interpreted as validating the theory whenever the 95% credible interval of the effect of interest falls within the bounds of the range of predicted effects. This is illustrated in scenario E in Figure 4. By contrast, the data will be interpreted as falsifying the theory whenever the credible interval lies completely outside of the range of model predictions; these are scenarios A and B in Figure 4. The intermediate outcomes occur when the credible interval and the range of model predictions overlap; these will be interpreted as equivocal evidence; see scenarios C and D in Figure 4. Scenario F represents a situation where the credible interval from the data is wider than the predicted range from the model; in this case, more data should be collected before conclusions can be drawn.

This method can also be used to evaluate whether an effect is effectively zero or not; here, it is necessary to define the range that counts as no effect. An example using this approach is presented in Vasishth, Mertzen, Jäger, and Gelman (2018). A related method has been proposed by Matthews (2019) as a replacement for null hypothesis significance testing.


(a) Grammatical conditions.

(b) Ungrammatical conditions.

Figure 4. Possible outcomes when interpreting the empirical data against the range of predictions of the Lewis and Vasishth (2005) cue-based retrieval model. Panel (a) shows the interference effect in grammatical conditions and panel (b) for ungrammatical conditions. The range of ACT-R predictions is shown at the bottom of each panel (see Section Deriving quantitative predictions from the Lewis and Vasishth (2005) model for details); A-F represent the 95% credible intervals of hypothetical experimental outcomes. Outcomes A and B falsify the model, outcomes C and D are equivocal outcomes, and E would be strong support for the model. Outcome F is uninformative and can only occur when the data does not have sufficient precision given the range of model predictions. The figure is adapted from Spiegelhalter et al. (1994, p. 369).

Since our model evaluation procedure depends on conducting Bayesian analyses, we explain our data-analytical methodology next.

Bayesian parameter estimation

In order to use the region-of-practical-equivalence approach for model evaluation, we need the marginal posterior distribution of the interference effect. A posterior distribution is a probability distribution over possible effect estimates given the data and the statistical model. The posterior distribution thus displays plausible values of the effect given the data and model. Bayes' rule allows this computation: Given a vector y containing data, a joint prior probability density function $p(\theta)$ on the parameters θ, and a likelihood function $p(y \mid \theta)$, we can compute, using Markov chain Monte Carlo methods, the joint posterior conditional density of the parameters given the data, $p(\theta \mid y)$. The computation uses the fact that we can approximate the posterior density up to proportionality via Bayes' rule, which states that the posterior density is proportional to the likelihood multiplied by the prior. Formally, we would write: $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$. Here, $p(\theta \mid y)$ is the posterior density given the data, $p(y \mid \theta)$ is the likelihood, and $p(\theta)$ represents the joint prior distribution of all the parameters in the model. From this joint posterior distribution of the parameters θ, the marginal distribution of each parameter can easily be computed. For extended tutorial introductions to Bayesian data analysis specifically addressed to cognitive scientists, see Nicenboim and Vasishth (2016), Sorensen, Hohenstein, and Vasishth (2016) and Vasishth, Nicenboim, Beckman, Li, and Kong (2018). Textbook-length treatments are provided by

Kruschke (2015), McElreath (2016) and Lambert (2018).

Because psycholinguistics generally uses repeated measures factorial designs, the likelihood function is a complex hierarchical linear model with many variance components. The prior distributions for the parameters are typically chosen so that they have a regularizing effect on the posterior distributions to avoid overfitting (these are sometimes referred to as *weakly informative* priors). In the analyses presented in this paper, we limit ourselves to such regularizing priors. These priors have the effect that the so-called maximal linear mixed model (Barr, Levy, Scheepers, & Tily, 2013) will always converge even when data are relatively sparse; when there is insufficient data, the posterior estimate of each parameter will be determined largely by the regularizing prior. By contrast, maximal models in the frequentist paradigm will fail to converge when there is insufficient data; even if it appears that the maximal model converged, the parameter estimates of the variance components can be very unrealistic and/or can lead to degenerate variance-covariance matrices (Bates, Kliegl, Vasishth, & Baayen, 2015; Vasishth, Nicenboim, et al., 2018). Fitting maximal models using Bayesian methods gives the most conservative estimates of the effects, and allows us to take all potential sources of variance into account, as Barr et al. (2013) recommend.

The Bayesian approach is also more informative than the frequentist one as it is not limited to merely falsifying a point null hypothesis (although this can be done with Bayes factors, Jeffreys 1998), but rather provides direct information about the plausibility of different effect estimates given the data and the model. For an extended discussion of this point in the context of psycholinguistics, see Nicenboim and Vasishth (2016). Based on the posterior distributions, it is possible to make quantitative statements about the probability that an effect lies within a certain range. Thus, we can calculate the 95% credible interval for plausible values of an effect.[5]

Next, we carry out a Bayesian data analysis of the Dillon et al. (2013) study, and of our large-sample replication attempt. We begin by reanalyzing the original data of Dillon et al. (2013)'s Experiment 1; this allows us to directly compare the original results with our replication attempt.

Reanalysis of the Dillon et al. 2013 Experiment 1 data

Recall that Dillon et al. (2013) concluded that the processing of the different syntactic dependencies differs with respect to whether all available retrieval cues are weighted equally and are used for retrieval, or whether exclusively structural cues are used. Specifically, they argue that in the processing of subject-verb agreement, morphosyntactic cues, such as the number feature, are used whereas in reflexives, which are subject to Principle A of the binding theory (Chomsky, 1981), only structural cues are deployed to access the antecedent. This claim is based on their Experiment 1, which directly compared interference effects in subject-verb agreement and in reflexives. The main finding was that in total fixation times, facilitatory interference is seen only in subject-verb agreement sentences but not in reflexive conditions.

In the following, we will first summarize the method and materials used by Dillon et al. (2013) in their Experiment 1, and then present a Bayesian analysis of their data.

[5] There are infinitely many intervals in which 95% of the probability lies; the credible interval is defined as a symmetric interval such that the same amount of probability mass lies to its left and to its right. Alternatively, one could report a 95% interval around, for example, the median of the posterior.

Method and materials of Dillon et al. 2013

In a reading experiment using eyetracking, Dillon et al. collected data from 40 native speakers of American English in the US who were presented with 48 experimental items. There were eight experimental conditions (shown in Example 4), which were presented in a Latin square design, interspersed with 152 fillers. The grammatical to ungrammatical ratio was 4.6 to 1. Items a-d relate to the subject-verb agreement conditions, and e-h to the reflexives.

(4) a. *Agreement; grammatical; interference*
The amateur bodybuilder$^{+singular}_{+local\ subject}$ who worked with the personal trainer$^{+singular}_{-local\ subject}$ amazingly was$\{^{singular}_{local\ subject}\}$ competitive for the gold medal.

b. *Agreement; grammatical; no interference*
The amateur bodybuilder$^{+singular}_{+local\ subject}$ who worked with the personal trainers$^{-singular}_{-local\ subject}$ amazingly was$\{^{singular}_{local\ subject}\}$ competitive for the gold medal.

c. *Agreement; ungrammatical; no interference*
*The amateur bodybuilder$^{-plural}_{+local\ subject}$ who worked with the personal trainer$^{-plural}_{-local\ subject}$ amazingly were$\{^{plural}_{local\ subject}\}$ competitive for the gold medal.

d. *Agreement; ungrammatical; interference*
*The amateur bodybuilder$^{-plural}_{+local\ subject}$ who worked with the personal trainers$^{+plural}_{-local\ subject}$ amazingly were$\{^{plural}_{local\ subject}\}$ competitive for the gold medal.

e. *Reflexive; grammatical; interference*
The amateur bodybuilder$^{+singular}_{+\ c\text{-}com}$ who worked with the personal trainer$^{+singular}_{-\ c\text{-}com}$ amazingly injured himself$\{^{singular}_{c\text{-}com}\}$ on the lightest weights.

f. *Reflexive; grammatical; no interference*
The amateur bodybuilder$^{+singular}_{+\ c\text{-}com}$ who worked with the personal trainers$^{-singular}_{-\ c\text{-}com}$ amazingly injured himself$\{^{sinular}_{c\text{-}com}\}$ on the lightest weights.

g. *Reflexive; ungrammatical; no interference*
*The amateur bodybuilder$^{-plural}_{+\ c\text{-}com}$ who worked with the personal trainer$^{-plural}_{-\ c\text{-}com}$ amazingly injured themselves$\{^{plural}_{c\text{-}com}\}$ on the lightest weights.

h. *Reflexive; ungrammatical; interference*
*The amateur bodybuilder$^{-plural}_{+\ c\text{-}com}$ who worked with the personal trainers$^{+plural}_{-\ c\text{-}com}$ amazingly injured themselves$\{^{plural}_{c\text{-}com}\}$ on the lightest weights.

All eight conditions in one set of items started with the same singular subject noun phrase (NP), which was the target for retrieval (*The amateur bodybuilder* in Example 4). This target NP was modified by a subject-relative clause containing a distractor NP (*the personal trainer/s* in Example 1) whose match with the number feature on the matrix verb (agreement conditions a-d) or the reflexive (conditions e-h) was manipulated; we refer to the number manipulation on the distractor NP as the *interference* factor.

For agreement conditions, the matrix clause verb (*was/were* in Example 4) that triggered the critical retrieval was followed by an adjective. For reflexive conditions, the antecedent of the reflexive (*himself/themselves* in Example 4) was the sentence-initial noun phrase. The grammaticality of the sentences was manipulated by having the number feature of the reflexive or the matrix verb match or mismatch the singular target NP. Hence, conditions with a plural matrix verb or a plural reflexive were ungrammatical.

Bayesian re-analysis of the Dillon et al. data

Our primary analysis focused on total fixation times (i.e., the sum of all fixation durations on a region), in both the re-analysis of the original data as well as in the analysis of the replication experiment. This is because the conclusions of Dillon et al. (2013) were based on total fixation time. As they explain (Dillon et al., 2013, p. 92), they investigated a higher order interaction between grammaticality, interference, and dependency type. This interaction did not reach significance in the minF′ calculation (F(1,81)=2.66, p=0.11), but was considered to furnish evidence for a difference in interference patterns in the two dependency types. The other measures they report (first-pass regressions and first-pass reading time) did not show any evidence for an interaction either.

We used the same critical region as Dillon et al. (2013): in agreement conditions, the critical region was the main clause verb and the following adjective, and in the reflexive conditions, it was the reflexive and the following preposition. We only analyzed the critical region as Dillon and colleagues' conclusions were based on this region.

All data analyses were carried out in the R (version 3.5.1) programming environment (R Core Team, 2016). The Bayesian hierarchical models were fit using Stan (Carpenter et al., 2017), via the R package RStan, version 2.18.1 (Stan Development Team, 2017a) and the R package brms (Bürkner, 2017) for the calculation of Bayes factors.

We fit two hierarchical linear models in order to unpack the main effects and interactions, and the nested effects of interest. Model 1 tests for an interaction between dependency and interference separately within ungrammatical and grammatical conditions. If dependency type does not matter, no interactions involving dependency type are expected. Model 2 investigates interference effects separately in agreement and reflexive constructions; these interference effects are nested within the grammatical and ungrammatical conditions. Both models include main effects of dependency, grammaticality, the interaction between grammaticality and dependency in order to fully account for the factorial structure of the experiment. The contrast coding of all comparisons included in the models is summarized in Table 1.

All interference effects were coded such that a positive coefficient means inhibitory interference, i.e., a slowdown in reading times in the interference conditions. A positive coefficient for the main effect of grammaticality means that the ungrammatical conditions are read more slowly and a positive coefficient for dependency means that agreement conditions take longer to read than reflexive conditions. All contrasts were coded as ±0.5, such that the estimated model parameters would reflect the predicted effect, i.e., the difference between the relevant condition means.

For our first research question, namely whether there is a difference between the dependency types with respect to the interference effect in ungrammatical conditions as claimed by Dillon et al. (2013), the relevant comparison is the two-way interaction between dependency and interference within ungrammatical sentences in Model 1.

For our second research question, namely a quantitative evaluation of the predictions of cue-based retrieval theory, the relevant comparisons are the interference effects within

		Experimental condition							
		Agreement				Reflexives			
		gram		ungram		gram		ungram	
		int	no int	int	no int	int	no int	int	no int
Models 1, 2	Dependency	0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	-0.5
	Grammaticality	-0.5	-0.5	0.5	0.5	-0.5	-0.5	0.5	0.5
	Dependency×Grammaticality	-0.5	-0.5	0.5	0.5	0.5	0.5	-0.5	-0.5
Model 1	Interference [grammatical]	0.5	-0.5	0	0	0.5	-0.5	0	0
	Interference [ungrammatical]	0	0	-0.5	0.5	0	0	-0.5	0.5
	Dependency×Interference [grammatical]	0.5	-0.5	0	0	-0.5	0.5	0	0
	Dependency×Interference [ungrammatical]	0	0	-0.5	0.5	0	0	0.5	-0.5
Model 2	Interference [grammatical] [reflexives]	0	0	0	0	0.5	-0.5	0	0
	Interference [grammatical] [agreement]	0.5	-0.5	0	0	0	0	0	0
	Interference [ungrammatical] [reflexives]	0	0	0	0	0	0	-0.5	0.5
	Interference [ungrammatical] [agreement]	0	0	-0.5	0.5	0	0	0	0

Table 1
Contrast coding of the independent variables. For the analysis of total fixation times of the original Dillon et al. (2013, Experiment 1) data as well as of the replication data, two models were fit, as described in the main text. Here, gram refers to grammatical, ungram refers to ungrammatical, and int refers to interference.

grammatical and ungrammatical conditions of Model 1. From the model's perspective, the overall effect collapsed over the two dependency types is the relevant effect; however, in order to account for the possibility of a difference between the dependency types, we will also evaluate the model against each of the grammatical and ungrammatical interference effects nested within dependency type, which are included in Model 2. None of the other fixed effects are of theoretical interest to our research questions and are only included to reflect the factorial design of the experiment.

We used a hierarchical LogNormal likelihood function to model the raw values in milliseconds (ms); this is equivalent to fitting a hierarchical linear model with a Normal likelihood on log-transformed values. One important motivation for using a LogNormal generative distribution instead of the standardly used Normal is that the latter leads to unrealistic posterior predictive data; see Appendix B, and Vasishth, Mertzen, et al. (2018); Vasishth and Nicenboim (2015). All models had a full random effects structure. That is, we assumed correlated varying intercepts and slopes for items and for subjects. As prior distributions, we used a standard normal distribution $N(0, 1)$ for all fixed effects (except the intercept, which had a N(0,10) prior), and a standard normal distribution truncated at 0 for the standard deviation parameters. These are regularizing, mildly informative priors (Gelman et al., 2014), as discussed earlier. Within the variance-covariance matrices of the by-subject and by-items random effects, priors were defined for the correlation matrices using a so-called LKJ prior (Lewandowski, Kurowicka, & Joe, 2009). This prior has a parameter, η; setting the parameter to 2.0 has a regularizing effect by disfavoring extreme values (± 1) for the correlations; for details see Stan Development Team (2017b). For each of the models, we sampled from the joint posterior distribution by running four MCMC chains at 2000 iterations each. The first half of the samples was discarded as warm-up samples. Convergence was checked using the R-hat convergence diagnostic and by visual inspection of the chains (Gelman et al., 2014).

Results

The results of our Bayesian analysis of the original data from Dillon et al. (2013)'s Experiment 1 are summarized in Table 2, which shows the mean of the posterior distribution of each parameter of interest (backtransformed to ms), together with a 95% credible interval (CrI), i.e., a range of plausible values of the effect given the data and the model. A detailed comparison of our analysis with the original analysis performed by Dillon et al. (2013) is provided in Appendix B.

We will present the results focusing on the two research questions we set out to answer: (i) is there a principled difference between the facilitation profiles observed in ungrammatical subject-verb agreement and reflexives, as claimed by Dillon et al. (2013), and (ii) are the predictions of cue-based retrieval theory consistent with the empirical data?

The interference effect in ungrammatical conditions across dependency types. Model 1 investigates the two-way interaction between dependency and interference within ungrammatical sentences, and does not show any evidence for a difference between the dependency types: the mean of the posterior is -21 ms with CrI [-56, 12] ms.

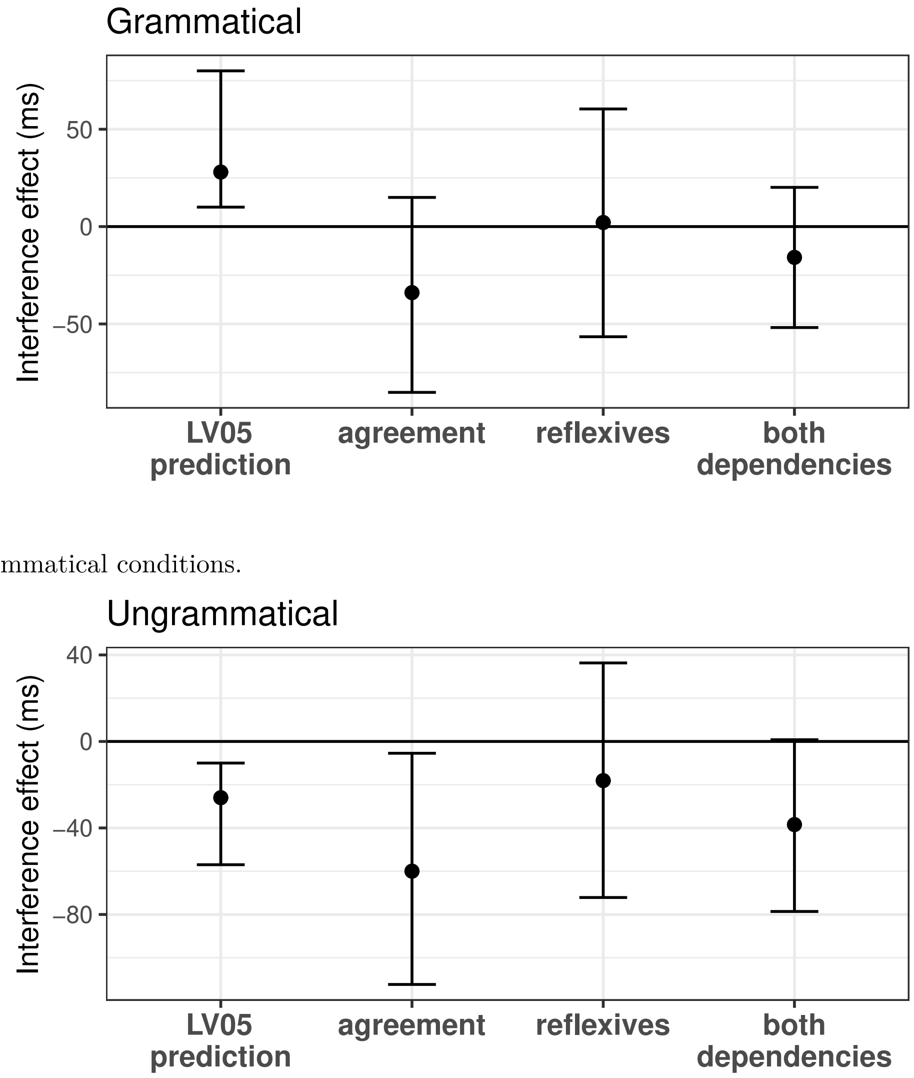
Model 2 reveals that the interference effect in ungrammatical conditions is numerically much larger in subject-verb agreement than in reflexives. In fact, given the reading (eyetracking and self-paced reading) literature on interference effects (for an overview, see Jäger et al. 2017), the mean of the posterior distribution of the interference effect in ungrammatical agreement conditions is surprisingly large: -60 ms, CrI [-112, -5] ms (cf. the Jäger et al. (2017) meta-analysis estimate of -22 ms [-36, -9]). In reflexives, the mean of the posterior is much smaller: -18 ms, CrI [-72, 36] ms.

Comparison of the empirical estimates with model predictions. Figure 5 compares the range of effect sizes predicted by the LV05 cue-based retrieval model with the empirical estimates obtained from Dillon et al. (2013)'s Experiment 1 total fixation time data. The figure shows the estimated interference effects observed in total fixation times within each level of grammaticality for the two dependency types separately (i.e., the estimates obtained from Model 2), and collapsed over the two dependency types (Model 1).

In grammatical conditions, the credible interval of the interference effect is [-57, 60] ms in reflexives and [-85, 15] ms in agreement conditions. When collapsing the two dependencies, the credible interval of the overall interference effect is [-52, 20] ms. When comparing these estimates with the model predictions of [10, 80] ms, we see that the range of model predictions and the empirical estimates overlap on one side, resulting in scenario C of Figure 4. The data for both agreement and reflexives are equivocal; the patterns certainly can't be seen as a strong validation of the cue-based retrieval model.

In the ungrammatical conditions, the credible interval of the interference effect is [-72, 36] ms in reflexives and [-112, -5] ms in agreement conditions. The credible interval of the

(a) Grammatical conditions.



(b) Ungrammatical conditions.

Figure 5. Evaluation of the ACT-R predictions (see Section Deriving quantitative predictions from the Lewis and Vasishth (2005) model for details) against the corresponding estimates (posterior means and credible intervals) obtained from total fixation times of the empirical data of Dillon et al. (2013).

	Effect	Posterior mean (ms)
Models 1, 2	Dependency	119 [71, 169]
	Grammaticality	100 [69, 134]
	Dependency×Grammaticality	9 [-18, 36]
Model 1	Interference [grammatical]	-16 [-52, 20]
	Interference [ungrammatical]	-38 [-79, 1]
	Dependency×Interference [grammatical]	-17 [-56, 19]
	Dependency×Interference [ungrammatical]	-21 [-56, 12]
Model 2	Interference [grammatical] [reflexives]	2 [-57, 60]
	Interference [grammatical] [agreement]	-34 [-85, 15]
	Interference [ungrammatical] [reflexives]	-18 [-72, 36]
	Interference [ungrammatical] [agreement]	-60 [-112, -5]

Table 2

Bayesian analysis of Dillon et al. (2013)'s Experiment 1. The table shows the mean of all fixed effects' posterior distributions together with 95% Bayesian credible intervals of total fixation times at the critical region. All models were fit on the log-scale; all numbers in this table are back-transformed to ms for easier interpretability. For more details about the model specification and the contrast coding of the fixed effects, see Section Bayesian re-analysis of the Dillon et al. data.

overall interference effect across the two dependencies is [-79, 1] ms. These credible intervals are so large that we end up in scenario F of Figure 4: the data's credible interval goes beyond the model predictions of [-57, -10] ms on both sides. We therefore conclude that these data are uninformative for a quantitative evaluation of the Lewis and Vasishth (2005) cue-based retrieval model. A higher-power study is necessary to adequately evaluate these predictions (for detailed discussion on power estimation see Cohen (1988) and Gelman and Carlin (2014).

Discussion

The results of our analysis of Dillon et al. (2013)'s Experiment 1 data show that it is difficult to argue for a dependency × interference interaction (within grammatical or within ungrammatical conditions). In particular, the data do not support the claim that agreement conditions show facilitatory interference effects in ungrammatical sentences but reflexives do not; the large uncertainty associated with the interference effect in reflexives does not allow any conclusions about the absence or presence of an effect.[6]

We now turn to the large-sample replication attempt.

[6] The first four conditions of Dillon et al. (2013)'s second experiment are identical to the reflexives conditions of their Experiment 1. In a post-hoc analysis suggested by Dillon, we combined these data in order to increase statistical power of the analysis. The posterior mean of the dependency×interference interaction within ungrammatical sentences is 3 ms with a credible interval of [-29, 34] ms; within ungrammatical agreement conditions, the posterior mean of the interference effect is -59 ms with a credible interval of [-113, -7] ms, and in ungrammatical reflexives conditions, it is 0 ms with a credible interval of [-33, 33] ms.

Replication experiment

Due to concerns about the statistical power of Dillon et al. (2013)'s Experiment 1, we carried out a direct replication attempt with a larger participant sample. The power analysis presented in Appendix A shows that the prospective power of the replication attempt has an upper bound of 99% when the largest predicted effect size is considered; this contrasts with the upper bound of 51% power in the original experiment.

Method

Participants. Over a period of four years (2014-2018), eye-movement data of 190 participants were collected at Haskins Laboratories (New Haven, CT). All participants had normal or corrected-to-normal vision and no previous diagnoses of reading or language disability. For their participation, each participant received 20USD. Data from 9 participants were excluded due to poor calibration.

Materials. We followed the experiment design by Dillon et al. (2013) and used the same 48 experimental items as the original Experiment 1 (see Example 1). For more details about the experimental stimuli see Section Method and materials of Dillon et al. 2013. We selected 128 of the 152 original filler items; to reduce the length of the overall procedure, we removed a filler experiment that Dillon et al. (2013) had used. Half of all experimental and filler items were followed by a yes/no comprehension question which targeted different parts of the sentences.

Procedure. The 48 experimental items were presented in a Latin Square design. Experimental and filler items were randomized within each Latin Square list. Sentences were presented in one line on the screen in Times New Roman font (size 20). For some very long sentences, the non-critical end of the sentence was displayed in a second line. A 21-inch monitor with a resolution of 1680×1050 pixels was used to display the sentences.

After giving informed consent, participants were seated in front of the display monitor. A chin rest and forehead rest were used to avoid head movements. The eye-to-screen distance was approximately 98 cm. An Eyelink 1000 eye-tracker with a desktop-mounted camera was used for monocular tracking at a sampling rate of 1000 Hz.

After setting up the camera, a 9-point calibration was performed. Testing began after a short practice session of 4 trials. Comprehension questions were answered by pressing a button on a gamepad. A break was offered to participants halfway through the session, and additional breaks were given when necessary. Re-calibrations were performed after the break, and whenever deemed necessary.

The collection of eye-movement data, including setup, calibrations, re-calibrations and breaks, took approximately 45 minutes.

Results

Question response accuracies. Overall response accuracy on experimental trials followed by a comprehension question was 88%. Mean accuracies for each experimental condition are provided in Table 3. Accuracy on filler items was 91%.

	Agreement				**Reflexives**			
	grammatical		ungrammatical		grammatical		ungrammatical	
	int	no int	int	no int	inter	no int	int	no int
Accuracy (%)	88	88	89	89	86	87	90	89

Table 3
Mean response accuracies of the trials followed by a comprehension question for each experimental condition (int refers to the interference conditions, and no int to the no-interference conditions.

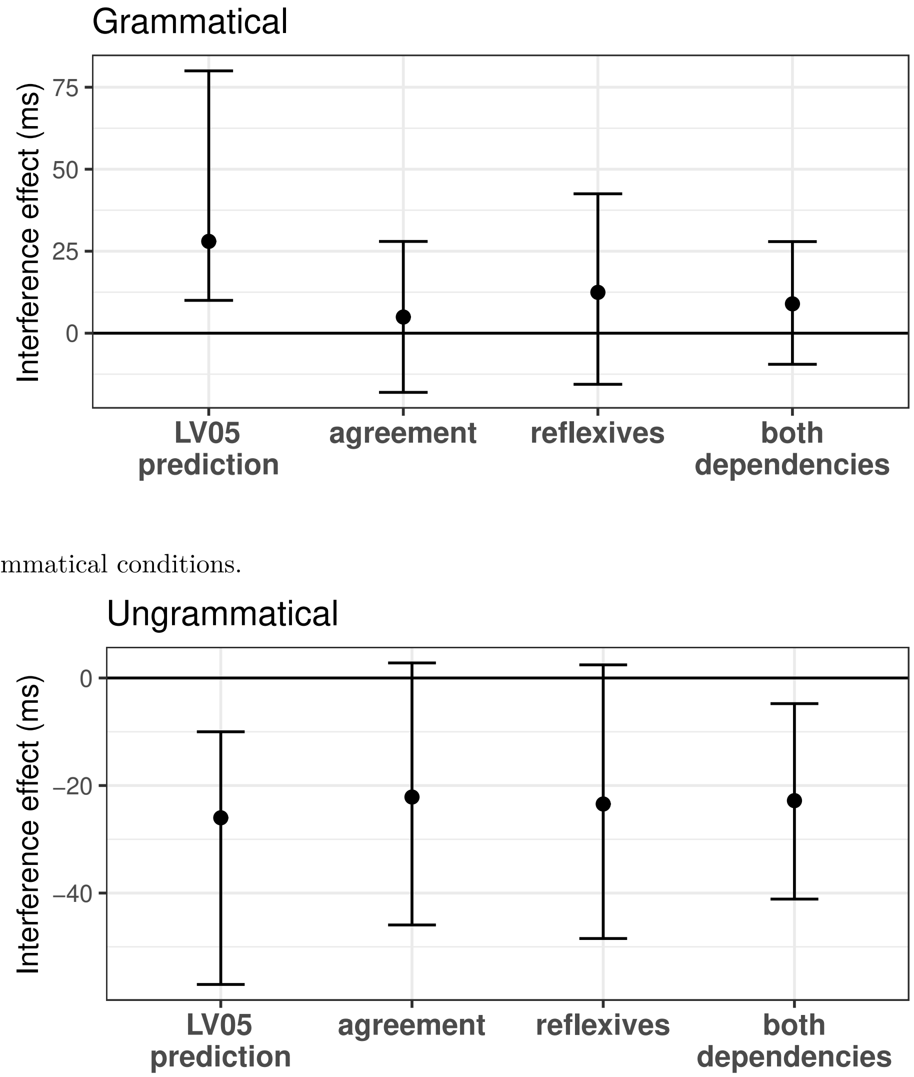
Primary analysis based on total fixation time. Following the Dillon et al. (2013) analysis, we collapsed the main clause verb or the reflexive with the subsequent word to form the critical region. Recall that Dillon et al. (2013) only found an effect in total fixation time; first-pass regressions and first-pass reading time showed no evidence for differing interference profiles in the two dependency types. Accordingly, we restricted our primary analysis to total fixation time, and we fit the same two Bayesian hierarchical models discussed in the section entitled Bayesian re-analysis of the Dillon et al. data.

The total fixation time results are summarized in Table 4, which shows the posterior mean of each fixed effect together with a 95% credible interval. Overall, the posterior distributions of the parameters obtained from the replication experiment have a much higher precision (i.e., tighter credible intervals) than the posteriors computed from the original data, as was expected given the much larger sample size.

Similar to the presentation of the results of the Bayesian analysis of the original data, we will focus on the effects that are relevant for evaluating (i) the Dillon et al. (2013) claim that in ungrammatical sentences, interference from the number feature affects only subject-verb agreement but not reflexives, and (ii) the quantitative predictions of the Lewis and Vasishth (2005) cue-based retrieval model.

The interference effect in ungrammatical conditions across dependency types. The two-way interaction between dependency and interference within ungrammatical conditions (Model 1) is effectively centered around zero (1 [-17, 18] ms). Hence, from the replication data, it is difficult to argue for a difference between the dependency types in ungrammatical conditions as claimed by Dillon et al. (2013).

(a) Grammatical conditions.



(b) Ungrammatical conditions.

Figure 6. Evaluation of the ACT-R predictions (see Section Deriving quantitative predictions from the Lewis and Vasishth (2005) model for details) against the corresponding estimates (posterior means and credible intervals) obtained from total fixation times of the replication experiment.

	Effect	Posterior mean (ms)
M 1, 2	Dependency	141 [100, 184]
	Grammaticality	121 [100, 141]
	Dependency×Grammaticality	-17 [-30, -5]
M 1	Interference [grammatical]	9 [-9, 28]
	Interference [ungrammatical]	-23 [-41, -5]
	Dependency×Interference [grammatical]	-4 [-21, 13]
	Dependency×Interference [ungrammatical]	1 [-17, 18]
M 2	Interference [grammatical] [reflexives]	12 [-16, 43]
	Interference [grammatical] [agreement]	5 [-18, 28]
	Interference [ungrammatical] [reflexives]	-23 [-48, 2]
	Interference [ungrammatical] [agreement]	-22 [-46, 3]

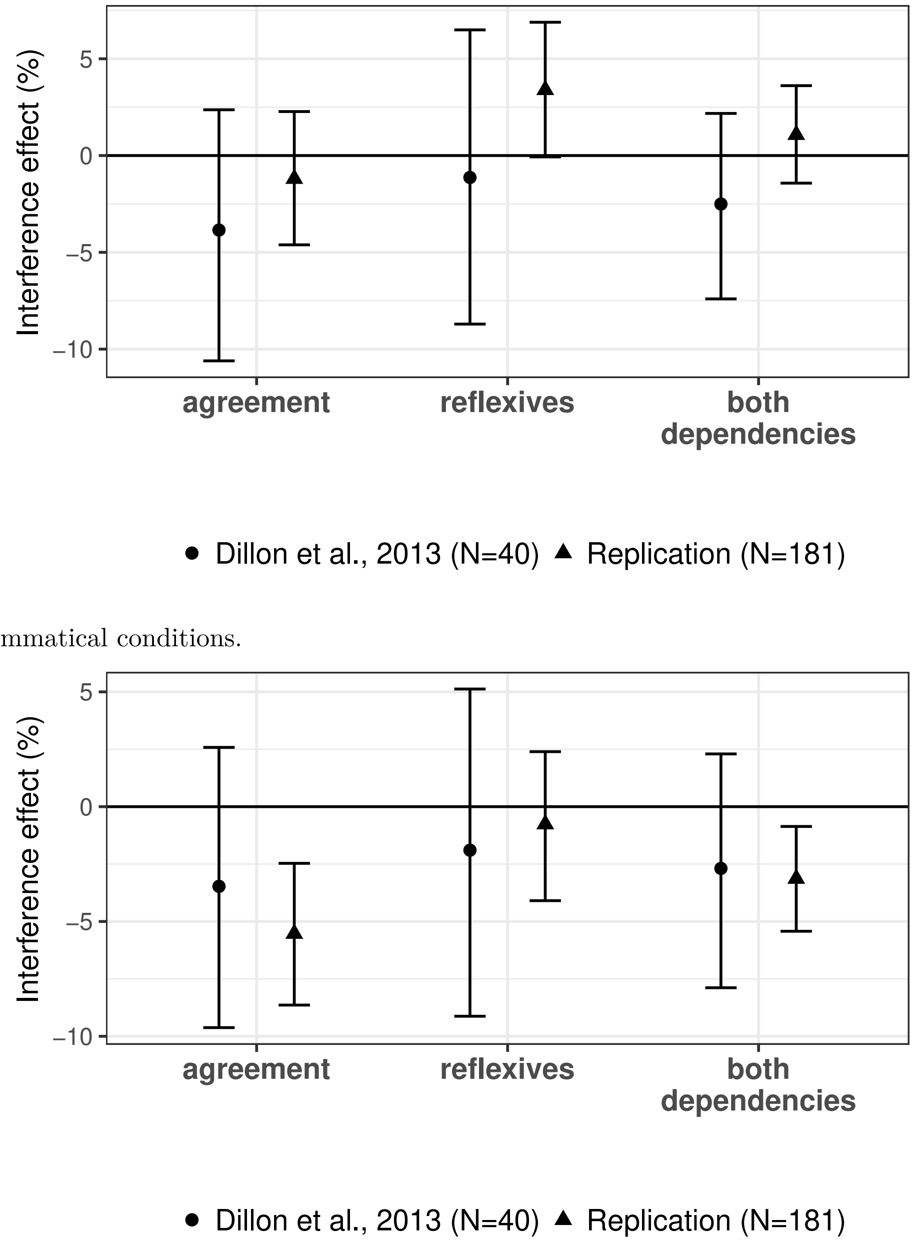
Table 4

Bayesian analysis of total fixation time in the replication experiment. The table shows the posterior means of the fixed effects together with 95% Bayesian credible intervals of total fixation times at the critical region. All models (M1 and M2) were fit on the log-scale; all numbers in this table are back-transformed to milliseconds for easier interpretability. For more details about the model specification and the contrast coding of the fixed effects, see Section Bayesian re-analysis of the Dillon et al. data and Table 1.

Comparison of the empirical estimates with model predictions. Figure 6 shows the range of model predictions from the Lewis and Vasishth (2005) cue-based ACT-R model of sentence processing (see Section Deriving quantitative predictions from the Lewis and Vasishth (2005) model) together with the empirical estimates obtained from the replication experiment. The figure shows the estimated interference effects observed in total fixation times within each level of grammaticality for the two dependency types separately (i.e., the estimates obtained from Model 2), as well as the overall interference effect across the two dependency types (Model 1).

In grammatical conditions, the credible interval of the overall interference effect across dependency types (Model 1) is [-9, 28] ms. In reflexives alone, the credible interval of the interference effect is [-16, 43] ms. In agreement conditions, the credible interval is [-18, 28] ms. The range of model predictions and the empirical estimates overlap on one side, resulting in the theoretical scenario C of Figure 4. Under this scenario, the evidence is equivocal: about half of the data's credible interval falls within the range of predicted effects.

In ungrammatical conditions, the credible interval of the overall interference effect across dependency types is [-41, -5] ms. The credible intervals of the interference effects nested within dependency type are strikingly similar: [-48, 2] ms for reflexives and [-46, 3] ms for subject-verb agreement. There is a much larger overlap of the model predictions and the empirical data — but still the estimates' credible interval do not completely fall inside the range of plausible model predictions, resulting in the theoretical scenario D of Figure 4. When collapsing over dependency types, the credible interval does not include zero, but the credible interval still goes beyond the model predictions (see Figure 6b), meaning that it is plausible that the true facilitation is even smaller than the lower boundary of the range of model predictions.



(a) Grammatical conditions.

(b) Ungrammatical conditions.

Figure 7. Analysis of first-pass regressions. Interference effects in grammatical (panel a) and ungrammatical conditions (panel b). The figure shows the posterior means together with 95% credible intervals of the interference effects in the percentage of first-pass regressions. These estimates were obtained from the Bayesian analysis of the original data of Dillon et al. (2013), and from our replication data. Separate effect estimates for each dependency type as well as the overall effect obtained when collapsing over dependencies are presented. Quantitative predictions of the cue-based retrieval model are not shown alongside the regression probabilities because there is no obvious linking function that quantitatively maps retrieval time in cue-based retrieval theory to first-pass regressions.

Secondary analysis based on first-pass regressions. As mentioned earlier, our primary, confirmatory analysis of total fixation times was based on the fact that Dillon et al. (2013) found effects (facilitatory interference in subject-verb agreement but not in reflexives) only in this dependent measure in their Experiment 1. In first-pass reading times and first-pass regressions, they found no evidence for the dependency × interference interaction. Dillon et al. (2013, p. 88) expected slower reading times or a lower proportion of regressions in ungrammatical interference manipulations in both agreement and reflexive constructions. This is not a very focused prediction: it subsumes all eye-tracking dependent measures. One common problem that eye-tracking researchers face is that often a predicted effect cannot a priori be attributed to a specific reading measure or a limited set of measures. As a consequence, a frequently used procedure is to analyze several measures (often at multiple regions, such as the critical and post-critical region)[7] These multiple comparisons, however, lead to a greatly inflated Type I error probability in eyetracking data (von der Malsburg & Angele, 2017). Because of this multiple testing problem, it is possible that the original Dillon et al. pattern found in total fixation time is indeed a false positive. A further issue is that such exploratory analyses of the data render hypothesis testing—including significance testing in the frequentist framework—invalid (De Groot, 1956/2014; Nicenboim et al., 2018).

Despite these issues, it can be useful to conduct exploratory analyses to generate new hypotheses that can be tested in future experiments. Hence, we additionally analyzed two further dependent measures: first-pass reading times and first-pass regressions. Should the effect of interest be observed in (either of) these two measures, it will be necessary to validate such an exploratory finding in a future confirmatory study.

First-pass reading times did not show any indication of contrasting interference profiles in our data; no interference effects were found at all in this measure in subject-verb agreement dependencies nor in antecedent-reflexive dependencies. We therefore don't discuss this measure any further.

The patterns seen in first-pass regressions are more interesting: in grammatical conditions, reflexives show some indication of the predicted inhibitory interference, whereas no interference is observed in agreement conditions.[8] The posterior mean of the interaction between dependency type and interference within grammatical conditions is -2.3%, with a 95% credible interval of [-4.7, 0.1]%. In ungrammatical conditions, by contrast, agreement conditions show a clear indication of the predicted facilitation, whereas reflexives do not show any interference. The posterior distribution of the dependency × interference interaction within ungrammatical conditions is highly similar to the one in grammatical conditions, it has a mean of -2.4%, and a credible interval of [-4.7, 0.0]%. Figure 7 shows the interference effects in first-pass regressions estimated from the replication data together with the ones obtained from the original data of Dillon et al. (2013). Quantitative predictions of the cue-based retrieval model are not shown because there is no obvious linking function that quantitatively maps retrieval time in cue-based retrieval theory to first-pass regressions.

[7] One way to alleviate this issue is to conduct a confirmatory analysis on a measure that showed the effect in a previous experiment, as we do in the current paper, or to run an exploratory pilot experiment, followed by a higher power confirmatory study in which the primary analysis tests the findings of the previous experiment (Nicenboim et al., 2018).

[8] The statistical model here was a Bayesian logistic mixed effects regression with a binomial link function, always fit with a full variance-covariance matrix assumed for subject and item random effects.

Discussion

As in the total fixation time analysis of the original data, total fixation time in the replication data do not show any indication for a difference between the interference profiles of the two dependency types. Indeed, the estimates for the facilitation in ungrammatical sentences in both reflexives and subject-verb agreement dependencies are almost identical; they both show facilitation. This is not consistent with the idea, suggested by Dillon and colleagues, that there are contrasting interference profiles for agreement vs. reflexives. Furthermore, the estimated facilitation in the agreement conditions is much smaller for our large-sample replication attempt than the estimate obtained from the original data. The smaller estimates in our replication data, along with their much tighter credible intervals relative to the original study, suggest that the effect estimate for agreement conditions in the original data may be an overestimate, a so-called Type M error (Gelman & Carlin, 2014). Type M errors can occur when statistical power is low (for a discussion of Type M errors in psycholinguistics, see Vasishth, Mertzen, et al. 2018).

Turning next to the analysis of first-pass regressions, in grammatical conditions, reflexives show inhibitory interference. This is an uncontroversial finding as it is predicted by cue-based retrieval theories in terms of cue-overload. By contrast, the subject-verb agreement dependency appears to be insensitive to the interference manipulation. The absence of an inhibitory interference effect in grammatical subject-verb agreement conditions is consistent with previous studies' findings, as summarized in the literature review by Jäger et al. (2017). As Nicenboim et al. (2018) suggest, it may be difficult to detect this effect even in large sample-size studies due to a smaller effect size in grammatical subject-verb agreement configurations. An alternative explanation for the absence of the effect is suggested by Wagers et al. (2009): cue-based retrieval is only triggered in ungrammatical subject-verb agreement constructions, where a mismatch is detected between the subject and verb's number feature. If no retrieval occurs in grammatical subject-verb agreement constructions, no interference would occur.

By contrast, in ungrammatical conditions we see a contrasting interference profile in subject-verb agreement vs. reflexives, a pattern consistent with the Dillon et al. (2013) proposal. Here, agreement conditions show facilitation, whereas reflexives seem to be immune to interference. The reflexives result is also consistent with the original Sturt (2003) proposal, which stated that reflexives are immune to interference only in the early moments of processing. If first-pass regressions index early processing, then our finding would be consistent with Sturt's original account.

It is difficult to draw a strong conclusion from first-pass regressions in the ungrammatical conditions without a fresh confirmatory analysis with a large sample of data. There are several reasons for being skeptical. First, the original study by Dillon and colleagues—the only published study that investigated that the dependency × interference interaction—does not show any evidence at all for interference effects in first-pass regressions; crucially, the study does not even show the uncontroversial facilitatory interference effect in subject-verb dependencies. This total absence of any interference effect in first-pass regressions in Dillon and colleagues' study could simply be due to low power; but it may also suggest that the observed interference in our replication data is an accidental outcome. Second, very few reading studies show interference effects in first-pass regressions: the literature review in

Jäger et al. (2017) shows that only 4 out of 22 comparisons found significant interference effects in first-pass regressions, whereas 12 that had total fixation time as the dependent measure (see Appendix A of Jäger et al. 2017). Open-access data from studies published more recently, such as Cunnings and Sturt (2018) or Parker and Phillips (2017), which investigate facilitatory interference effects, do not show interference effects in first-pass regressions (we discuss the Parker & Phillips, 2017 data below in more detail). Third, there is only weak evidence for this dependency × interference interaction effect in our replication data. Evidence for or against an effect being present requires a hypothesis test, which can be computed using Bayes factors.[9] Comparing the full model which contains all the contrasts with a null model that does not have the relevant interaction term in it (i.e., when the null hypothesis is that the interaction term is 0) shows that in first-pass regressions, the evidence for a dependency × interference interaction is between 0.6 and 2 in grammatical conditions and between 1.5 and 5 in ungrammatical conditions. A Bayes factor of larger than 10 in favor of the full model is generally considered to be strong evidence for an effect being present (Jeffreys, 1998). Thus, the evidence from first-pass regressions for an interaction is quite weak. The results of the Bayes factor analyses are shown in Table 5.

Nevertheless, it is possible that the first-pass regression patterns we observed are replicable and robust; if this turns out to be the case, this would be a strong validation of the Sturt (2003) and Dillon et al. (2013) proposal. For this reason, a very informative future line of research would be to conduct a new large-sample replication attempt, i.e., a confirmatory study, that investigates the effect in first-pass regressions, as well as in total fixation times. If the same pattern as in our replication can be found, that would validate the original Sturt (2003) proposal that reflexives should be immune to interference in only early measures (in the Sturt paper, this was first-pass reading time); in late measures (for Sturt, this was re-reading time), interference effects should be observed. The patterns we found could be consistent with this proposal, provided that two sets of effects from first-pass regressions and total reading times can be replicated.

Returning to the confirmatory analysis involving total fixation times, we can conclude the following. The total fixation times show nearly identical facilitatory interference effects in both dependency types, suggesting a similar retrieval mechanism. Our conclusion that different dependencies might have a similar retrieval mechanism is also supported by a recent paper by Cunnings and Sturt (2018), which also found facilitatory interference effects in total fixation time in non-agreement subject-verb dependencies. Of course, large-scale replication attempts should be made to replicate the findings that we report here; in that sense, our conclusions should be regarded as conditional on the effects replicating in future work.

With respect to the predictions of the cue-based retrieval model, and assuming equal cue-weighting, in ungrammatical conditions, the replication data are consistent with the

[9] Bayes factors are the Bayesian analog of frequentist likelihood ratio tests, aka ANOVA. Bayes factors calculations provide the odds that the full model vs. a null model is compatible with the data (Jeffreys, 1998). In Bayes factor analyses, it is important to calculate a range of Bayes factors values given increasingly informative priors on the parameter that is being tested (Lee & Wagenmakers, 2014; Nicenboim, Vasishth, & Rösler, 2019). This is because uninformative priors can heavily favor the null model when the effect is small, as is generally the case in eyetracking studies. For discussion, see Lee and Wagenmakers (2014). Our analyses were carried out using the bridge-sampling approach (Gronau et al., 2017) implemented in the R package brms (Bürkner, 2017), a front-end for Stan.

	Grammatical conditions	
	Prior on Dep × Int effect	Bayes factor in favor of alternative
1	Normal(0,1)	0.57
2	Normal(0,0.8)	0.71
3	Normal(0,0.6)	0.95
4	Normal(0,0.4)	1.36
5	Normal(0,0.2)	1.94
	Ungrammatical conditions	
	Prior on Dep × Int effect	Bayes factor in favor of alternative
1	Normal(0,1)	1.54
2	Normal(0,0.8)	1.97
3	Normal(0,0.6)	2.54
4	Normal(0,0.4)	3.54
5	Normal(0,0.2)	5.31

Table 5

Bayes factor analysis of first-pass regressions in our replication data of the dependency × intererence interaction, in grammatical and ungrammatical conditions. Shown are increasingly informative priors on the parameter representing the interaction term in the model; for example, Normal(0,1) means a normal distribution with mean 0 and standard deviation 1. We consider a range of priors here because of the well-known sensitivity of the Bayes factor to prior specification. The Bayes factor analysis shows the evidence in favor of the interaction term being present in the model; a value smaller than 1 favors the null model, and a value larger than 1 favors the full model including the interaction term. A value of larger than 10 is generally considered to be strong evidence for the effect of interest being present (Jeffreys, 1998).

quantitative model predictions. However, the data also suggest that the facilitatory effect in ungrammatical sentences might be even smaller than the model's predicted range of effects. In grammatical conditions, by contrast, the data really are equivocal: they neither falsify nor validate the model. Here, too, the data indicate that the true interference effect might be smaller than the model's predictions.

General Discussion

In this work, we investigated an important open theoretical question originally raised by Sturt (2003): do configurational cues, such as those triggered by Principle A of the binding theory, serve as a filter for memory search in sentence processing? Such a proposal implies that certain cues may have a higher weighting than others. Dillon and colleagues have suggested that the answer is yes: in their Experiment 1, total fixation times show that ungrammatical subject-verb agreement constructions exhibit facilitatory interference effects but ungrammatical reflexive constructions do not. An absence of facilitatory interference effects in antecedent-reflexive conditions would imply that Principle A renders this dependency type immune from interference effects.

We wanted to replicate this finding from total fixation time in a higher-powered study. We were also interested in comparing the estimates from the higher-powered replication attempt with the quantitative predictions of cue-based retrieval theory, under the assumption that equal weighting is given to all cues.

The results of our investigations of the total fixation time data are summarized in Figure 8, which also shows the estimates from the data in the original study by Dillon and colleagues. The figure shows the interference effects for reflexives and subject-verb agreement separately, and the overall interference effect computed across both dependency types (recall that the relevant effect from the model's perspective is the overall effect collapsed over the two dependency types). The quantitative predictions of the Lewis and Vasishth (2005) ACT-R cue-based retrieval model are also shown in the figure. These predictions are generated with the assumption that syntactic cues do not have a privileged position when resolving dependencies of either type.

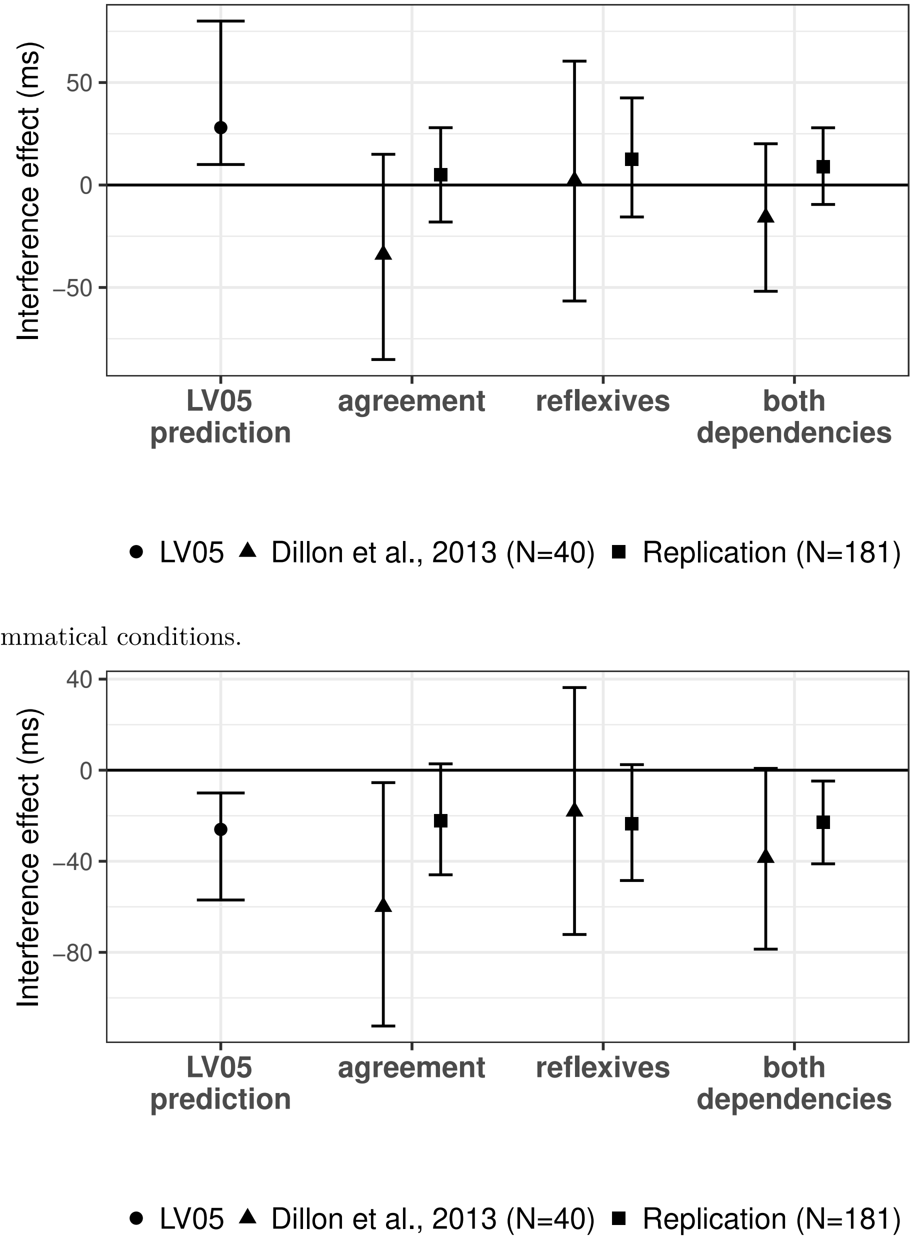
Regarding the difference in the interference profiles in ungrammatical agreement and reflexive conditions, our data analyses of total fixation times using mildly informative priors show that neither the original data collected by Dillon et al. (2013) nor the replication data provide any indication for a difference between the two dependency types. A comparison of the quantitative model predictions with the replication estimates shows that the model's predictions for ungrammatical sentences are consistent with the observed estimates in the replication data. Thus, the replication data seem to be consistent with the view that tree-configurational and non-configurational cues have equal weight.

Regarding the evaluation of the model predictions, for the grammatical conditions the conclusion based on the large-sample replication attempt is equivocal: there is some overlap between the range of predicted values from the model and the 95% credible intervals from the data. For the ungrammatical conditions, the estimates from the replication data seem to be consistent with the LV05 model: the credible intervals from the data largely overlap with the range of values predicted by the model. Of course, it would be very informative to carry out an even larger sample-size study than ours to obtain tighter credible intervals; a strong validation of the LV05 prediction would require that the credible intervals from the data fall entirely within the predicted range. We leave such a larger study for future work.

Arguments in favor of a privileged position for configurational cues

The broader theoretical question relates to whether there is a privileged position for configurational cues. There are in fact good a priori reasons to assume that syntactic cues in general may have some priority. For example, in inhibitory interference settings manipulating syntactic and semantic cues, Glaser, Martin, Van Dyke, Hamilton, and Tan (2013); Tan, Martin, and Van Dyke (2017); Van Dyke (2007) have shown that syntactic interference effects were observed earlier than semantic interference effects. Moreover, garden-path constructions such as reduced relatives (*The evidence/lawyer examined by the judge...*) have been argued to be insensitive to semantic cues such as animacy (Ferreira & Clifton, 1986, but cf. Trueswell, Tanenhaus, & Garnsey, 1994 for evidence against this view).

In the specific case of reflexives, Sturt (2003) has proposed that Principle A could play a dominant role in the earliest moments of processing. Sturt's argument for an early immunity of reflexives to interference was based on the absence of interference effects in first fixation durations. Thus, the fact that our replication study shows facilitatory interference effects in reflexives in the late measure total fixation time may still be consistent with Sturt's original proposal of the priority of configurational syntax.


(a) Grammatical conditions.

(b) Ungrammatical conditions.

Figure 8. Interference effects in grammatical (8a) ungrammatical conditions (8b). The figure shows the posterior means together with 95% credible intervals of the interference effects in total fixation times. These estimates were obtained from the Bayesian analysis of the original data of Dillon et al. (2013), and from our replication data. Separate effect estimates for each dependency type as well as the overall effect obtained when collapsing over dependencies are presented. The left most line of each plot shows the range of predictions of the Lewis and Vasishth (2005) ACT-R cue-based retrieval model (see Section Deriving quantitative predictions from the Lewis and Vasishth (2005) model for details).

However, one would then expect to see facilitatory interference effects in ungrammatical subject-verb agreement constructions in early measures, and specifically in first-pass reading times; this is because in agreement dependencies, no immunity from interference is assumed in the early stages (Dillon et al., 2013). We do not see any interference effects at all in first-pass reading times in the original data from Dillon et al. (2013) or in our replication data. First-pass regressions also didn't show any interference effects in the original Dillon et al. (2013) data. However, in our replication data, we do see some evidence for the Sturt proposal in our exploratory analysis of first-pass regressions: in ungrammatical conditions, subject-verb agreement shows indications of facilitatory interference but reflexives do not. To our knowledge, ours is the first study to find such a dependency × interference interaction in first-pass regressions. If this effect can be robustly replicated in future studies, this would indicate that the original Sturt proposal, which made a distinction between early and late processes, was on the right track: contrasting interference profiles would be expected in the first-pass regressions but not in total fixation time.

Here, it is worth mentioning that an intermediate position is also tenable: it is possible that the configurational cues do not render reflexives *completely* immune to interference (even in the early moments of processing), but are weighted higher than non-configurational cues. This proposal was investigated by Parker and Phillips (2017) empirically and with computational modeling. They derived quantitative predictions using an implementation of the Lewis and Vasishth (2005) model to show that configurational cues have a higher weight than non-configurational ones (they refer to these cues as structural vs. morphological), but that the weighting of non-configurational cues is not zero. This proposal seems quite reasonable: Parker and Phillips found evidence showing larger facilitatory interference effects in eyetracking reading data as a consequence of increasing the number of feature mismatches (one vs. two mismatches) in ungrammatical subject-verb agreement and antecedent-reflexive conditions. They did not report first-pass regressions or total fixation time results; however, the authors shared their data with us, which allowed us to evaluate the evidence in these measures for their claim. First-pass regressions showed no effects at all in two out of their three experiments; in their Experiment 3, first-pass regressions showed fewer regressions when there was a two-feature match with the distractor.[10] By contrast, total fixation times show evidence supporting their conclusion in all of their three experiments. Our reanalyses are available from https://osf.io/reavs/. Thus, Parker and Phillips' total fixation time data furnish good evidence in favor of such an intermediate position as a possible resolution to the theoretical question. As an aside, we note that because their first-pass regressions don't show clear evidence for facilitatory interference, this further reduces our confidence in the dependency × interference interaction we found in our replication data.

In sum, there are good reasons to assume that syntactic cues more generally, and configurational cues in particular, may have a higher weighting compared to other cues. Dillon and colleagues adopted an extreme version of this hypothesis: that configuarational cues have all the weighting, leading to complete immunity in reflexive constructions. However, both our current data and the data from Parker and Phillips cast doubt on this extreme position. We have demonstrated that our total fixation time data are consistent with an

[10] In their Experiment 3, the lme4 model did not converge, but a Bayesian linear mixed model with regularizing priors of the sort used in the present paper showed the expected effect, a 7% (CrIs: -15,-1%) reduction in regression probability, in the two-feature mismatch case. See https://osf.io/reavs/ for details.

implementation of cue-based retrieval that assumes equal weighting for all cues.

Stepping back from the details of this particular experiment and the theoretical issues we have addressed, this work also contributes to the rapidly expanding body of work that re-examines claims in psycholinguistics by conducting higher-powered studies (Nicenboim et al., 2018, 2019; Nieuwland et al., 2018; Vasishth, Mertzen, et al., 2018). In the early days of psycholinguistics, studies on phenomena like garden-path effects needed relatively small sample sizes to robustly demonstrate that an effect can be observed. But when researchers investigate relatively subtle effects, relying on conventional sample sizes of 30-40 participants will lead to either a high proportion of null results, or exaggerated estimates (due to Type M error) that cannot be replicated (Vasishth, Mertzen, et al., 2018). This implies that higher power studies and direct replications are necessary when investigating subtle effects. Prospective power analyses as shown in Appendix A should be carried out routinely before starting a study to ensure that estimates are accurate and have higher precision.

Acknowledgements

We are grateful to Brian Dillon for his magnanimous help with this project, his detailed and constructive comments, and for generously releasing data and code. He also engaged deeply with an earlier version of this paper, leading to significant improvements in the presentation of results. We are also very appreciative of the release of published data and data-analysis code by Ian Cunnings and Patrick Sturt, and by Daniel Parker and Colin Phillips. We are grateful to Morgan Bontrager and Heidi Sarles-Whittlesey for assistance with data collection. The research reported here was partly funded by: the University of Potsdam; the Volkswagen Foundation through grant 89 953; the Deutsche Forschungsgemeinschaft (German Science Foundation), Collaborative Research Center - SFB 1287, project number 317633480 (*Limits of Variability in Language*) through project Q (PIs: Shravan Vasishth and Ralf Engbert), which partly funded Lena Jäger; and the SFB 1287 project B03 (PIs: Ralf Engbert and Shravan Vasishth), which partly funded Daniela Mertzen; and by an Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH/NICHD) grant to Haskins Laboratories: R01 HD-073288 (PI: Julie Van Dyke).

Appendix A

Prospective power analysis of Experiment 1 in Dillon et al., 2013 and of our replication experiment

Given theoretical quantitative estimates of an effect size, we can obtain estimates of power for a particular experimental design such as Experiment 1 of Dillon et al. (2013) by generating simulated data repeatedly to determine true discovery rate, i.e., the proportion of cases where the absolute t-value of the effect of interest exceeds 2.

We carried out this true discovery rate or power analysis as follows. Reproducible code and the associated data are available from https://osf.io/reavs/. We focused on the facilitatory interference effect in ungrammatical subject-verb agreement, and the corresponding effect in ungrammatical reflexives, as these are the important data points for the present paper.

The method for calculating power is as follows:

1. For a given experimental design and given an existing data-set, a maximal linear mixed model with full variance-covariance matrices for subjects and items (without correlation parameters in the variance-covariance matrices) is fit using lme4 to compute estimates of all parameters. The estimates from this model are stored.

2. Then, fake data are generated repeatedly 100 times. For data generation, parameter estimates are taken from the preceding step above. Each fake data-set is analyzed using a maximal linear mixed model (without correlation coefficients). Convergence failures (which were below 3% in the present case, are discarded).

3. True discovery rates are computed for agreement and reflexives in ungrammatical conditions by computing the proportion of cases where the absolute t-value of the effect of interest is greater than 2.

Table A1 summarizes the power analyses using this method. The power analyses shown are for the original Dillon et al. (2013) Experiment 1 and for our replication attempt.

Experiment	Dependency	Effect sizes (ms)		
		-57	-26	-10
		Power estimates (%age)		
Dillon et al., 2013 E1 (n=40)	Agreement	43	26	7
	Reflexive	51	27	5
Our replication (n=181)	Agreement	97	84	11
	Reflexive	99	91	8

Table A1
Prospective power analysis of Experiment 1 of Dillon et al. (2013), and of our large-sample replication attempt. Shown are a range of effect sizes derived from the quantitative predictions of the cue-based retrieval model of Lewis & Vasishth, 2005, and the corresponding power estimates (the probability of correctly detecting the effect under repeated sampling, expressed as a percentage) for the two experiments. The number of items was 48 in both experiments.

Appendix B

Comparison of our analysis with the original analysis by Dillon et al. 2013

In addition to using the Bayesian paradigm for our analysis, there are several other important differences between our analysis and the original analysis conducted by Dillon et al. (2013). First, Dillon and colleagues used repeated measures ANOVAs rather than linear mixed models to analyze total fixation times. To use ANOVA, one has to aggregate over items or participants, which artificially eliminates one source of variance in the data. Using repeated measures ANOVA will generally lead to artificially low p-values in the F1 (by subjects) and F2 (by items) ANOVA because the variance components due to subjects and items are not taken into account simultaneously as in hierarchical linear models with crossed random subjects and items.

Second, the contrast coding of Dillon et al. differs from ours: Dillon et al. coded the interference effect as the effect of the interfering noun (i.e., *personal trainer/s* in Example 1) being plural compared to being singular. Specifically, the interference effect was calculated by subtracting the reading time for conditions with a singular distractor from those with a plural distractor (see Dillon et al. 2013, p. 92). For ungrammatical conditions, this results in the same coding as the one we are applying. However, for grammatical conditions, the sign of the interference effect is opposite in the Dillon et al. coding compared to ours. That is, in the original analysis by Dillon et al., a facilitatory interference effect would have a negative sign in ungrammatical conditions, but a positive sign in grammatical conditions.

Third, Dillon et al. analyzed total fixation times on the ms scale, treating reading times as coming from a Normal distribution. By contrast, we assumed a LogNormal distribution. There are several important statistical reasons for using the LogNormal for analyzing reading time data (Gelman & Hill, 2007). One adverse consequence of not using the LogNormal is shown in Figure B1: the posterior predictive distribution from the Normal-distribution model is very unrealistic compared to the data. For example, the Normal-distribution model predicts negative reading times. Compare this to the posterior predictive distribution from the LogNormal-distribution model. Statistical inferences are not valid when the distributional assumption of the residuals is not approximately satisfied.

From a quantitative perspective, our results differ substantially from those presented by Dillon et al.; see Table B1 for an overview. Most importantly, the interference effect within ungrammatical agreement conditions has been estimated to be -119 ms with a confidence interval of [-205, -33] ms by Dillon et al. (2013), whereas the mean of the posterior distribution obtained from our analysis is -60 ms with a credible interval of [-112, -5] ms. This considerable difference in the estimated effect size is due to the logarithmic transformation we applied to the total fixation times which reduced the impact of extremely large values. In ungrammatical conditions of reflexives, by contrast, we get a larger effect size (-18 ms with a credible interval of [-72, 36] ms) compared to the original analysis (-8 ms [-67, 51] ms, see Dillon et al. 2013, p. 92). In the grammatical conditions, the estimates are also different, albeit less substantially. Whereas Dillon reports an interference effect of -43 ms [-106, 20] ms for agreement and 10 ms [-48, 68] ms for reflexives (see Dillon et al. 2013, p. 92), the mean and credible intervals of the posterior distribution obtained in our analysis are -34 ms [-85, 15] ms and 2 ms [-57, 60] ms, respectively. All of these differences are explained by the log-transformation we have applied and, to a much smaller extent, to the random effects structure of our models: whereas Dillon et al. report by-subject estimates aggregated over items, the hierarchical models we have used account for both between-items

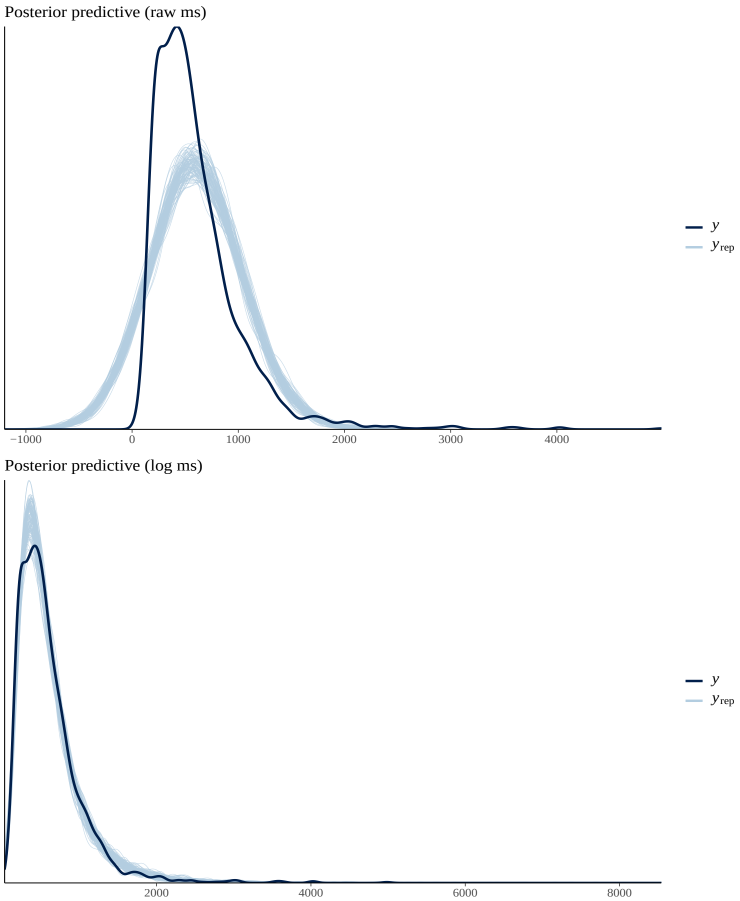


Figure B1. Posterior predictive distributions from 100 samples generated from a model that assumes a Normal distribution as generating reading times in milliseconds, and a LogNormal. The posterior predictive distributions from the Normal-distribution model generate unrealistic values; by contrast, the LogNormal-distribution model generates more reasonable predicted values.

and between-subjects variance simultaneously. This is more conservative than aggregating over subjects and over items.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–60.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models.* (Unpublished manuscript)

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01690965.2010.492228

		Dillon et al., 2013 (ms)	**Our reanalysis (ms)**
Grammatical	Reflexives	$10[-48, 68]$	2 [-57, 60]
	Agreement	$-43[-106, 20]$	-34 [-85, 15]
Ungrammatical	Reflexives	$-8[-67, 51]$	-18 [-72, 36]
	Agreement	$-119[-205, -33]$	-60 [-112, -5]

Table B1

Comparison of our Bayesian estimates of the interference effects nested within grammaticality and dependency type computed from log-transformed data with the estimates reported by Dillon et al. (2013, p. 92) which were computed from untransformed data. For the Bayesian estimates, we show the posterior mean and 95% credible interval, backtransformed to the ms scale; the estimates taken from Dillon et al. (2013) represent means and 95% confidence intervals. We have adjusted the sign of Dillon et al.'s estimates to match our contrast coding.

Brasoveanu, A., & Dotlačil, J. (2019). *Formal linguistics and cognitive architecture.* New York: Springer-Verlag.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht, The Netherlands: Foris.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

Cunnings, I., & Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, *102*, 16-27.

De Groot, A. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Mar1 Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica*, *148*, 188–194. doi: 10.1016/j.actpsy.2014.02.001

Dillon, B. (2011). *Structured access in sentence comprehension* (PhD thesis). University of Maryland, College Park, MD.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103.

Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. *Topics in cognitive science*, *10*(1), 144–160.

Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* (Doctoral thesis). Universität Potsdam, Germany.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). *The effect of prominence and cue association in retrieval processes: A computational account.* Retrieved from https://osf.io/b56qv/ (Manuscript submitted to Cognitive Science)

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*(3), 452-474.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, *40*(3), 575–586.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). Boca Raton, FL: Chapman and Hall/CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., & Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain and language*, *126*(3), 314–326.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80 - 97. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249617300640 doi: https://doi.org/10.1016/j.jmp.2017.09.005

Hobbs, B. P., & Carlin, B. P. (2008). Practical bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 1–8.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19–24.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(617). doi: 10.3389/fpsyg.2015.00617

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316-339. doi: https://doi.org/10.1016/j.jml.2017.01.004

Jeffreys, H. (1998). *The theory of probability.* Oxford University Press. (Original work published 1939)

Kruschke, J. (2015). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, The Netherlands: Academic Press.

Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (PhD thesis). University of Maryland, College Park, MD.

Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*(1252).

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Lambert, B. (2018). *A student's guide to Bayesian statistics.* Sage.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, *100*(9), 1989–2001.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Logačev, P., & Vasishth, S. (2015). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, 1–33.

Martin, A. E., & McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1231.

Matthews, R. A. (2019). Moving towards the post $p < 0.05$ era via the analysis of credibility. *The American Statistician*, *73*(sup1), 202–212.

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, *10*(1), 161–174.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Chapman and Hall/CRC Press.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

McElree, B. (2003). Accessing recent events. *Psychology of Learning and Motivation*, *46*, 155–200.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 155–200). San Diego, CA: Elsevier.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory differences in unbounded dependencies. *Frontiers in Psychology*, *7*(280). (Special Issue on Encoding and Navigating Linguistic Representations in Memory) doi: 10.3389/fpsyg.2016.00280

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass*, *10*(11), 591–613.

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, *42*. doi: 10.1111/cogs.12589

Nicenboim, B., Vasishth, S., & Rösler, F. (2019). *Are words pre-activated probabilistically during sentence comprehension? evidence from new data and a bayesian random-effects meta-analysis using publicly available data.*

Nieuwenhuis, S., Forstmann, B., & Wagenmakers, E. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . others (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468.

Parker, D. (2018). A memory-based explanation of antecedent-ellipsis mismatches: New insights from computational modeling. *Glossa: a journal of general linguistics*, *3*(1).

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321–339.

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

Patil, U., Hanne, S., Burchert, F., Bleser, R. D., & Vasishth, S. (2016). A computational evaluation of sentence comprehension deficits in aphasia. *Cognitive Science*, *40*, 5–50. doi: 10.1111/cogs.12250

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*. (Special Issue on Encoding and Navigating Linguistic Representations in Memory) doi: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00329/full

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Roberts, S., & Pashler, H. (2000, April). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*(3), 175-200. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/statistics/BayesLMMs.html

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357–416.

Stan Development Team. (2017a). *RStan: the R interface to Stan. R package version 2.16.2.* Retrieved from http://mc-stan.org

Stan Development Team. (2017b). Stan modeling language users guide and reference manual, version 2.17.0 [Computer software manual]. Retrieved from http://mc-stan.org

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562.

Tan, Y., Martin, R. C., & Van Dyke, J. A. (2017). Semantic and syntactic interference in sentence comprehension: a comparison of working memory models. *Frontiers in psychology*, *8*, 198.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, *33*, 285–285.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 407–430.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4).

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151-175. Retrieved from https://osf.io/eyphj/ doi: https://doi.org/10.1016/j.jml.2018.07.004

Vasishth, S., & Nicenboim, B. (2015). Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass*, *10*(8), 349–369.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 141-161. Retrieved from https://osf.io/g4zpv/ doi: 10.1016/j.wocn.2018.07.008

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133.

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545-1578. Retrieved from https://doi.org/10.1080/01690965.2012.728232 doi: 10.1080/01690965.2012.728232

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237.