Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German

Abstract

Cue-based retrieval accounts of sentence processing postulate that at a verb, retrieval cues are generated to complete a dependency with the verb's argument(s); for example, the dependency between the subject and the verb must be completed. If these retrieval cues match with not only the subject but also with those on other nouns in the sentence, then processing difficulty arises at the verb. This difficulty in identifying the correct dependent is called similarity-based interference. We present large-sample self-paced reading and event-related potentials experiments using a well-established design to investigate interference due to syntactic and semantic cues in German. In this design, the syntactic cue {+subject} and the semantic cue {+animate} are manipulated. Bayes factors analyses showed evidence for a semantic interference effect in both experiments. Surprisingly, Bayes factors provided evidence against interference due to the syntactic cue {+grammatical subject} in this particular design in both experiments. This finding contradicts the predictions of the standard implementations of cue-based retrieval theory, which (implicitly) assume that both syntactic and semantic cues play an equal role in retrieval. We show through computational modeling that cue-

based retrieval will also shows no syntactic interference in the present design if the parser is assumed to keep track of which clause the subject occurs in. Thus, if syntactic retrieval cues include hierarchical syntactic information (is the noun in the same clause as the verb?), the predictions of the cue-based retrieval model would be consistent with the observed patterns in our data.

Introduction

It is well-established in sentence processing that completing linguistic dependencies requires memory retrieval (see e.g., Lewis et al., 2006). In fact, a key assumption in psycholinguistics is that the same constraints on memory retrieval that have been identified within memory research in cognitive psychology (e.g., Anderson et al., 2004) may be applicable in sentence processing as well. An example of a key construct from cognitive psychology that is relevant for sentence processing is that spreading activation leads to difficulty in identifying the correct item to retrieve from memory. The present work investigates whether and how such independently posited, general constraints on memory retrieval impact sentence processing. As a concrete example of general memory constraints impacting sentence comprehension, consider (1).

(1) The worker was surprised that the resident who was living near the dangerous warehouse was complaining about the investigation. (Van Dyke, 2007)

There is clear evidence in the psycholinguistic literature (e.g., Jäger et al., 2020; Nicenboim et al., 2018; Van Dyke and McElree, 2006; Van Dyke and Lewis, 2003) that, at the verb *was complaining*, it is necessary to retrieve the subject *the resident* from memory in order to interpret who did what to whom. A widely-held assumption is that such memory retrievals during sentence comprehension are guided by retrieval cues (Lewis and Vasishth, 2005; McElree, 2000).

Here, we consider two kinds of cues used in retrieval: semantic and syntactic cues. In a sentence like (1), the retrieval cues generated at the verb are

standardly assumed (e.g., Van Dyke, 2007) to include at least the semantic cue {+animate} and the syntactic cue {+grammatical subject}. These cues pick out the target noun *the resident* as the subject. The distractor *warehouse*, intervening between subject and verb, does not match either of these retrieval cues. However, if the distractor were *neighbor* instead of *warehouse*, then it would match the {+animate} cue at the verb. This overlap in cues between the target noun and the distractor is called cue-overload in cue-based retrieval accounts, and is predicted to trigger semantic interference, which expresses itself as greater processing time at the verb compared to the baseline condition (1).

By contrast, syntactic interference is expected to be triggered if the distractor matches the {+grammatical subject} cue. This situation can occur when the relative clause in (1) that modifies *the resident* were to be changed to *who said that the warehouse was dangerous*. Now, the distractor matches the {+grammatical subject} cue.

Thus, a distractor which matches a semantic cue, here, {+animate}, is predicted to induce semantic interference and a distractor which matches a syntactic cue, here, {+grammatical subject}, is predicted to induce syntactic interference. More generally, interference that is induced by such an overlap between retrieval cues and features of items in memory is called retrieval interference.

Previous work has shown clear evidence for this kind of retrieval interference (Van Dyke, 2007; Van Dyke and McElree, 2006, 2011; Nicenboim et al., 2018; Van Dyke and Lewis, 2003; Jäger et al., 2017). However, two important questions remain unanswered. First, as discussed in Jäger et al.

(2017), almost all the published studies on interference effects are severely underpowered. Low power has the consequence that published estimates will tend to be overestimates or misestimates of the effect size (Vasishth et al., 2018). Over-/misestimates have important implications for theory evaluation. One example is discussed in Vasishth et al. (2018), where a key prediction of surprisal theory (Levy and Keller, 2013) could not be validated when a larger-sample study was conducted. By contrast, the original small-sample study (Levy and Keller, 2013) showed suggestive evidence consistent with the theory. Another example is the recent heated debate (Nieuwland et al., 2018) on prediction in sentence comprehension, which was largely focused on the results of low-powered studies, leading to potentially misleading conclusions (Nicenboim et al., 2020). Second, although it has been established (e.g., Van Dyke, 2007; Mertzen et al., 2023) that syntactic and semantic cues are involved in sentence processing, it is not obvious which cues are relevant for a particular retrieval event. For example, Dillon et al. (2013) argued, following Sturt (2003), that the gender cue was either downweighted or not used at all when an antecedent-reflexive dependency needs to be build. Instead, the antecedent of a reflexive could be identified using the syntactic principle A of the binding theory, which correctly identifies the target for retrieval. Using computational modeling, Dillon et al. (2013) suggested that antecedent-reflexive dependencies were immune to interference, and that the reason for this is that principle A of the binding theory is the primary driver of the cue-based retrieval process (cf. Jäger et al., 2020; Yadav et al., 2022). This was a surprising discovery because the default assumption in cue-based retrieval accounts was that all cues were equally weighted, regardless of the

dependency type; this had the consequence that retrieval interference was expected whenever a cue-overload configuration was present in a sentence. More broadly, an important insight in the work of Dillon et al. (2013), among others, is that the precise details of which cues are used during retrieval can affect processing difficulty in possibly surprising ways. We build on this idea in the present paper.

The present study

Reading studies have been the dominant method used to investigate retrieval interference (see the review in Jäger et al., 2017). We aimed to investigate the role of syntactic and semantic cues during sentence comprehension with higher-powered experiments using self-paced reading (SPR) as well as electroencephalography (EEG), which to our knowledge has never been used to investigate interference during subject-verb dependency resolution.

All experiments used the 2 $\times$ 2 design developed by Van Dyke (2007); the design crosses syntactic and semantic interference.

Materials

As the same materials were used for all our experiments, we describe them here once before we describe the specifics of the experiments, separately. The studies had 120 items with four conditions in a Latin square design shown in (2). The items were constructed following Van Dyke's (2007) items and partially re-used Mertzen et al.'s (2023) items.

(2) Example item (critical word in bold, distractor in italics) of the present study:

a. High syntactic interference with high / low semantic interference:

Die Nachbarin glaubte, dass der Witwer, der erzählt
The$_{\text{FEM}}$ neighbor$_{\text{FEM}}$ believed that the widower who told
hatte, dass der *Einbrecher* / *Verlust* schrecklich war,
had that the burglar / loss awful was
abends regelmäßig **trank**, um zu vergessen.
in.the.evening regularly drank in.order to forget

'The neighbor believed that the widower, who had told her that the burglar / loss was awful, drank regularly in the evenings to forget.'

b. Low syntactic interference with high / low semantic interference:

Die Nachbarin glaubte, dass der Witwer, der von dem
The$_{\text{FEM}}$ neighbor$_{\text{FEM}}$ believed that the widower who about the
schrecklichen *Einbrecher* / *Verlust* erzählt hatte, abends
awful burglar / loss told had in.the.evening
regelmäßig **trank**, um zu vergessen.
regularly drank in.order to forget

'The neighbor believed that the widower, who had told her about the awful burglar / loss, drank regularly in the evenings to forget.'

The comprehension of the sentences in (2) requires a memory retrieval operation at the verb *trank* ('drank') to retrieve its non-adjacent subject *Witwer* ('widower'). During this retrieval, the distractor *Einbrecher* ('burglar') / *Verlust* ('loss') might cause interference and impede comprehension.

Syntactic interference is manipulated via the syntactic status of the distractor, which is either the subject (2a) of an embedded clause, or part of a prepositional phrase (2b) and hence a non-subject. Following Van Dyke

(2007), the assumption here is that during the retrieval of the subject, a distractor in *any* subject position causes more syntactic interference than a distractor inside a prepositional phrase – even if the distractor occurs in an embedded clause (this assumption stands in contrast to findings on so-called agreement attraction, where syntactic hierarchy plays a crucial role, see e.g., Franck et al., 2002, 2006; Franck, 2011; Franck and Wagers, 2020).

Semantic interference is manipulated via the animacy of the distractor. During the retrieval of the subject of the verb *trank* ('drank'), the animate distractor *Einbrecher* ('burglar') is assumed to cause high semantic interference, while the inanimate distractor *Verlust* ('loss') is assumed to cause low semantic interference. The animacy of the distractor is used as a stand-in feature for semantic features relevant for the specific retrieval event which might be more fine-grained. In 2, the semantic retrieval cue might be {+animate} or {+human} but it might as well be {+can drink}. While more generic features like {+animate} or {+human} (all animate distractors were human) might be created during memory encoding, more specific features like {+can drink} are probably only activated when relevant. The present study is not designed to provide insight into the specific semantic features used during memory encoding and retrieval. For experimental work on highly specific retrieval cues like {+sailable} and {+shatterable}, see Van Dyke and McElree (2006) and Cunnings and Sturt (2018). For a principled approach to feature selection using word embeddings, see Smith and Vasishth (2020). A plausibility norming experiment (see page 12) was conducted to ensure that the semantic manipulation did work, i.e., that the semantic requirements of the verb matched the features of the subject and the animate distractor, but

not the features of the inanimate distractor.

Regions of interest in the design. For both our SPR and EEG experiments, our principal focus was on the processing of the verb (*trank*, 'drank'). In the SPR experiment, the reading times of the immediately following spill-over region and of the pre-critical region were also of interest. Previous work using the present design (Van Dyke, 2007; Mertzen et al., 2023) has consistently found effects in the pre-critical region (we return to this point in the discussion sections, see Figure 14 A). A further issue with German embedded clauses (also see Mertzen et al., 2023) was that, without an appropriate intervening phrase, the clause embedding would lead to two verbs appearing one after another (*erzählt hatte, trank*, literally 'told, drank'). This is because in German, the verb is clause-final in embedded clauses. Such verb sequences are inherently difficult to process (Vasishth et al., 2011; Bach et al., 1986).

For these reasons, we added two adverbs before the critical verb. Including these intervening regions had a further advantage: It is well-known (e.g., Rayner et al., 2000) that the last word in a clause-final position causes a slowdown in reading; such a slowdown has the potential to spill over to the subsequent regions, possibly affecting the estimated effects at the critical region. The spill-over of longer reading times can be problematic because the standard deviation increases as well (Wagenmakers and Brown, 2007); such an increase in standard deviation would drastically lower statistical power. Inserting two adverbs before the critical verb thus served to attenuate spill-over from the clause-final word in the embedded clause preceding the critical region. This way, the critical region should be less affected by the clause-boundary slowdown. This design also has the advantage that any potential

effects seen at the pre-critical region (the adverb immediately preceding the verb) are less likely to be due to spill-over from the preceding region. Thus, our focus in the SPR experiment was on the processing difficulty observed at the critical verb and the surrounding regions.

In the EEG experiment, the region of interest was the critical verb. Due to the direct measurement of brain activity with excellent temporal resolution event-related potential (ERP) effects do not tend to spill-over into later regions. We focused specifically on the N400 elicited by the verb. The N400 is a negative deflection in the ERP signal that peaks approximately 400 ms after any word was encountered. Traditionally, it is associated with lexical-semantic processing, but the finding of multi-modal N400 effects suggests that it rather indexes differences in the processing of meaningful stimuli in general (Kutas and Federmeier, 2011). The N400 amplitude is greater for words that deviate from the meaning activated by its context compared to words that fit its context (see e.g., Kutas and Hillyard, 1980; Kutas and Iragui, 1998). Consequently, the amplitude of the N400 to a specific word is a negative function of its predictability within the given context (see e.g., Kutas and Hillyard, 1984; Frank et al., 2015a). Additionally, the N400 elicited by isolated words is negatively correlated with their frequency (Kutas and Federmeier, 2000). Therefore, the N400 amplitude is thought to index a) the increased difficulty to integrate a semantically anomalous or simply unexpected word into the sentence context (e.g., Hagoort et al., 2004) and/or b) the necessity to retrieve lexical information from memory which was not already pre-activated before the stimulus was perceived (e.g., Brouwer et al., 2017). These two accounts of the N400 have been extensively discussed in

the ERP literature on prediction (see e.g., Nieuwland et al., 2018; Nicenboim et al., 2020; Freunberger and Roehm, 2017; DeLong et al., 2005; Mantegna et al., 2019; Nieuwland et al., 2019). As a likely marker of memory retrieval during sentence comprehension, it is fair to assume that the N400 would show modulations due to retrieval interference. The few ERP studies that have investigated retrieval interference so far have indeed found negativities (Lee and Garnsey, 2015; Martin et al., 2014; Schoknecht et al., 2022; Vasishth and Drenhaus, 2011), although the topography and latency of these negativities was mostly not typical for the N400. It is important to highlight here that the previous ERP studies on interference either increased the number of distractors in a sentence or used grammatical gender or number as the critical retrieval cue. So, interference due to non-semantic features elicited negativities which were more or less similiar to the N400. Therefore, we expected the interference manipulation regardless of the type of retrieval cue (syntactic or semantic) to elicit N400 or similar effects. For a discussion of the relation between the N400 and other negativities, see Bornkessel-Schlesewsky and Schlesewsky (2019).

Alternately, retrieval interference – especially syntactic retrieval interference – might modulate the P600. The P600 is a positive deflection in the ERP wave form that is typically observed at least 600 ms after stimulus onset. It is associated with syntactic processing, e.g., morpho-syntactic errors and reanalysis (see, e.g., Osterhout and Holcomb, 1992; Kaan et al., 2000). High interference could increase difficulty during syntactic integration of subject and verb in our materials. Potential misretrievals of the distractor might lead to reanalysis which could manifest as an increased P600 amplitude in

high compared to low interference conditions.

Plausibility norming

Van Dyke conducted a pre-test to ensure that the distractor in high semantic interference conditions was a sufficiently probable subject of the critical verb, i.e., that it could potentially induce semantic interference. Following Van Dyke, we also conducted a plausibility norming experiment.

Participants in the plausibility norming study. Forty participants (18 female, 22 male, mean age: 27.3 years old, age range: 21 - 39 years old) from the SPR experiments were re-invited for the plausibility norming one month after they completed the SPR experiment. We used a subset of the participants who participated in the SPR study in order to ensure that the ratings for the norming came from the same pool of participants that did the reading study.

Materials in the plausibility study. For each of the 120 item quadruplets, four simple sentences with subject-verb-object word order were constructed, so that each of the nouns in the item once functioned as the subject of the critical verb. The norming materials corresponding to the item depicted in (2) are shown in (3).

(3) a. Subject:

Der Witwer trank.
the widower drank
'The widower drank.'

b. Animate distractor:

Der Einbrecher trank.
the burglar drank
'The burglar drank.'

c. Inanimate distractor:

Der Verlust trank.
the loss drank
'The loss drank.'

d. Introduction noun:

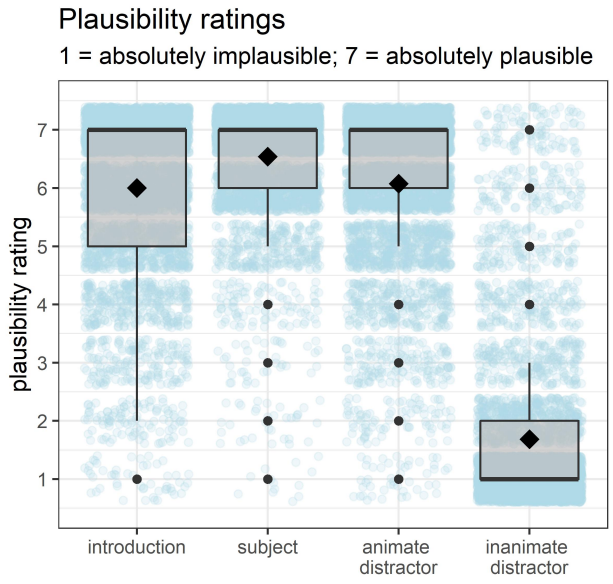
Die Nachbarin trank.
the_{FEM} $\text{neighbor}_{\text{FEM}}$ drank
'The neighbor drank.'

Thus, each of the forty participants were asked to rate 120×4 sentences, leading to 19,200 data points.

Procedure for the plausibility study. The plausibility norming experiment was implemented in PCIbex (Zehr and Schwarz, 2018) and was conducted online. Participants were asked to judge the plausibility of the sentences on a 7-point scale ranging from absolutely implausible (1) to absolutely plausible (7). All sentences belonging to one item were shown on one screen, so that they could be judged relative to each other. The plausibility norming experiment included 120 sentences from a different experiment which functioned as fillers. The participants received £3.40 as compensation for the norming experiment. A demonstration of the plausibility norming experiment can be accessed via https://farm.pcibex.net/r/WIeKYM/.

Results of the plausibility study. The plausibility ratings showed that the semantic manipulation worked as intended. The subject received the highest rating (mean: 6.5, sd: 0.9) and the animate distractor (high semantic interference conditions) was almost as plausible (mean: 6.1, sd: 1.3). In contrast, the inanimate distractor (low semantic interference conditions) was very implausible (mean: 1.7, sd: 1.4). The first noun in the target sentences, which was not part of the experimental manipulation, was tested as well and it received a high plausibility rating (mean: 6.0, sd: 1.4). These results are shown in Figure 1. The plausibility ratings are similar to those reported for Van Dyke's (2007) Experiment 3.

Figure 1: Plausibility ratings for the animate (high semantic interference conditions) and inanimate distractor (low semantic interference conditions), the introduction noun (*X believed that ...*) and the subject. The diamonds represent the mean rating. Blue dots show the individual data points.



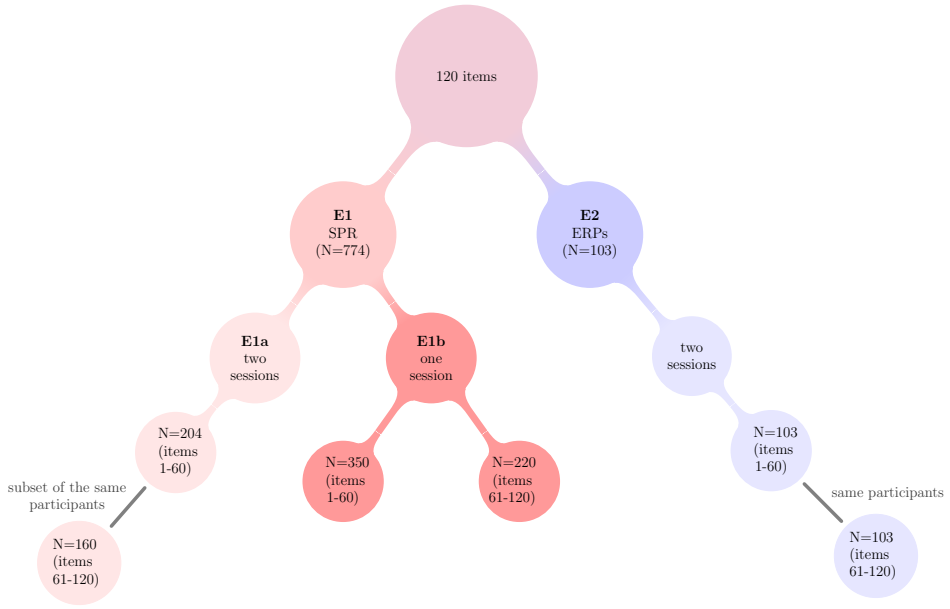
The structure and sample sizes of the SPR and EEG experiments

The SPR experiments were conducted web-based via Prolific and the EEG experiment was of course run in the lab. Because the signal-to-noise ratio of EEG data requires a large number of trials per condition and participant, we constructed 120 items. In order to keep the SPR and EEG studies comparable, we also used 120 items in the SPR experiments. In both methods, as it would have been very taxing for the participants to read all 120 items in one experimental session, we split the experiments into two sessions.

The SPR study, hereafter Experiment 1a, was initially designed so that participants read half of the items (1-60) in the first session and the other half of the items (61-120) in the second session on another day. The second session showed an adaptation effect (Prasad and Linzen, 2021): Reading times decreased overall. The adaptation effect is theoretically interesting per se (Fine et al., 2013) and we discuss it below; but it was problematic for our design because the average effect would then be confounded by adaptation. For this reason, we ran a second version of the SPR study, hereafter Experiment 1b, in which participants (N=570) were either shown items 1-60 or items 61-120 (but not both). In the Results section, we present the pooled data (N=774) from E1a and E1b. As mentioned above, because the ERP method requires a large number of items, all the ERP data come from participants who completed both sessions.

Figure 2 shows the structure of the present study, i.e., how many participants saw which subset of the items and whether these participants completed one or two experimental sessions.

Figure 2: The structure of the present study. In total, 120 items were used for the SPR (red branches) and the ERP experiments (blue branches). The color brightness reflects how many sessions were completed by each participant (lighter color: two sessions, darker color: one session).



Data Availability

All materials, data and analysis scripts can be accessed via https://osf.io/4fpru/

Experiments 1a and 1b: SPR

Methods

Participants

Native German-speaking participants were recruited via Prolific. In total, 908 participants took part in the self-paced reading experiments E1a and E1b. 117 participants were excluded due to low comprehension question

accuracy ($< 70\,\%$). Additionally, 16 participants were excluded because the demographic information in their Prolific profile and the information which they provided during the experiment did not match (regarding, e.g., their native language, language impairment, or age). One participant was excluded because of a technical error during their session. The data of 774 participants (mean age: 27 years old, age range: 18 - 40 years, 395 female, 369 male, 10 preferred to not provide sex information) were used for analyses. When asked about their highest level of education, 365 participants reported a bachelor's degree or higher university education. 334 participants reported to have a high-school diploma and 70 another secondary school certificate. Five participants replied with "other."

In Experiment 1a, drop-outs between sessions led to a lower sample size in the second session (204 vs. 160). In Experiment 1a, compensation for the first session was £3.5 and for the second session, it was £5. In Experiment 1b, compensation was £4.

Procedure

Participants completed a moving window self-paced reading task (Just et al., 1982) on the PCIbex Farm (Zehr and Schwarz, 2018). They were instructed to read for comprehension at a comfortable pace. All sentences were displayed word by word. Masked words were presented as a line indicating its length. Unmasked words were presented in 18 pt Courier font. The space bar on the keyboard was used to unmask words. The length of the sentences required line breaks. These were hard-coded so that in each sentence a line break appeared i) between the subject and the relative clause and ii) two words after the critical verb (see 4, critical word in bold, | indicate line

breaks in the moving window display). In a third of the trials, the sentence was followed by a yes/no comprehension question. The participants received no feedback on their performance.

In addition to the reading-for-comprehension task, the participants completed two types of attention and compliance checks. The attention checks required them to press one of two designated keyboard keys ten times during each session. The compliance checks required them to type in two types of fruit and two hobbies (session 1) and two German cities and two breakfast food items (session 2). No participant failed any of these checks. The reader can try the experiment via this link https://farm.pcibex.net/r/CBkSKl/.

(4) Die Nachbarin glaubte, dass der Witwer, | der erzählt hatte, dass
The neighbor believed that the widower | who told had that
der Verlust schrecklich war, regelmäßig abends **trank**,
the loss awful was regularly in.the.evening drank
um zu | vergessen.
in.order to | forget
'The neighbor believed that the widower, who had told her that the loss was awful, regularly drank in the evenings to forget.'

In Experiment 1a, participants were re-invited for session 2 after they completed session 1. The experimental sessions were separated by 1 - 20 days. The procedure of the sessions was identical. In session 1, the experimental items 1-60 were presented interspersed with 40 filler sentences. In session 2, the experimental items 61-120 were presented with another set of 40 filler sentences. The fillers were included to mask the critical manipulation. They were less syntactically complex than the experimental items

but had at least one embedded clause and generally provided more variety, e.g., by including modifiers and different types of subordinate clauses (e.g., "Der Teppichmacher, der früh in die Werkstatt gekommen war, reparierte den besonders schönen, alten Teppich während er die Nachrichten hörte", 'The carpet maker who came to the workshop early repaired the especially beautiful old carpet while listening to the news.'; "Der Jugendliche war genervt, weil seine Freundin, die nicht studieren wollte, oft die Schule schwänzte, um Computer zu spielen.", 'The teenager was annoyed because his girlfriend who did not want to go to University skipped school frequently to play video games.'). Each session lasted approximately 25 minutes.

Statistical analyses

Bayesian linear mixed models. Comprehension question accuracy was analyzed with Bayesian generalized linear mixed models, i.e., logistic regression, in R (R Core Team, 2024), using the brms package (Bürkner, 2021). The fixed effects were syntactic interference, semantic interference and their interaction. These were sum-contrast coded (high +0.5, low -0.5). Varying intercepts for participants and items were included. For the intercept, we used a Normal(0, 1.5) prior and for all fixed-effect slope parameters, a Normal(0, 0.1) prior. The priors for the variance components were the defaults specified in brms. Models were run with 4 chains and 8,000 iterations in each chain. 2,000 iterations in each chain consisted of a warm-up phase.

Self-paced reading times below 150 ms and above 3000 ms were excluded before analysis because these overly short or long reaction times were likely caused by inattentively performing the task. The procedure excluded 4.9 % of the data. Self-paced reading times were analyzed with Bayesian linear

mixed models in R (R Core Team, 2024) with log-normal likelihood, using the brms package (Bürkner, 2021). The critical word for analyses was the verb *trank* 'drank' in (2), constituting the retrieval site. Because previous work found effects in the pre-critical and post-critical region, reading times of the pre-critical word (*regelmäßig* 'regularly' in 2) and the spill-over region (*um* 'in.order.to' in 2) were analyzed. The models included fixed effects for syntactic interference (high +0.5, low -0.5), semantic interference (high +0.5, low -0.5) and their interaction. In addition to the predictors of interest, we included trial id to account for potential effects of fatigue or adaptation to the task. The trial id of the 100 trials per session was rescaled to span from 0 to 1. This was done to bring all fixed effects to a comparable scale which decreases the run time of the models. The models were run with full random effects, i.e., varying intercepts and slopes of all fixed effects and their interactions by participants and items.

Bayes factors for model comparison. In order to quantify the uncertainty on the parameters of interest, we report 95% credible intervals. These intervals represent the range over which we can be 95% certain that the values of the parameter lies, given the statistical model and the data. However, formal hypothesis testing cannot be carried out without a likelihood ratio test that compares two alternative models (Schad et al., 2022; Royall, 1997). For this reason, formal hypothesis tests for the presence or absence of the effects of interest were carried out using Bayes factors. Because Bayes factors can be very sensitive to prior specifications on the target parameter being tested (Schad et al., 2022), we report a sensitivity analysis using a range of prior specifications for the relevant parameters (see Table 1). The priors assume a

priori effect sizes for the main effects of syntactic and semantic interference, and their interaction, ranging from -8 to 8 ms, -40 to 40 ms, or -81 to 81 ms. These priors are based on the observed sizes of effects and uncertainties in previous reading studies that use the present design (Van Dyke, 2007; Mertzen et al., 2023), and on meta-analyses relating to interference effects (Jäger et al., 2017).

Table 1: Priors used for the analysis of the self-paced reading time data. The standard deviation of the fixed-effects slope priors was varied in order to conduct a Bayes factor sensitivity analysis following Schad et al. (2022). See the corresponding assumed a priori range of the difference between reading times under high vs. low interference on the millisecond scale.

Parameter	Prior	Assumed Range in ms
Intercept	Normal(6, 0.6)	[125, 1308]
slope	Normal(0, 0.01)	[-8, 8]
	Normal(0, 0.05)	[-40, 40]
	Normal(0, 0.1)	[-81, 81]
sigma	Normal(0, 0.5)	
SD	Normal(0, 0.1)	

Bayes factors were computed using the Savage-Dickey density ratio method. A Bayesian hypothesis test for which the posterior density is divided by the prior density at a specific parameter value of interest, e.g., zero (Wagenmakers et al., 2010; Vuorre, 2017; Dickey and Lientz, 1970; Dickey, 1971; Verdinelli and Wasserman, 1995). The Bayes factor is often written as BF_{10}. To obtain BF_{10} from the Savage-Dickey density ratio, we take its inverse. One

major advantage of the Bayes factor over frequentist ANOVA or likelihood ratio tests is that it takes the uncertainty of the parameters into account. This leads to more conservative inferences compared to frequentist ANOVA, which only takes the maximum likelihood estimate of the parameter into account (see Schad et al., 2022, for detailed discussion). Another important advantage of the Bayes factor – one that is very relevant for the present work – is that it is possible to find evidence for or against an effect. This stands in contrast to the frequentist ANOVA which, in its standard usage, is designed to only furnish evidence against the null.

Bayes factors can be interpreted as follows (e.g., Lee and Wagenmakers, 2014): if $BF_{10} > 1$, it provides evidence in favor of the effect of interest. If $BF_{10} < 1$, it provides evidence against the effect of interest. The larger the value of BF_{10}, the stronger the evidence for the effect and the smaller the value of BF_{10} is, the stronger the evidence against the effect of interest (see Table 2). In general, a large number of iterations is needed in the brms package in order to obtain stable estimates of the Bayes factor (Schad et al., 2022). For this reason, models were run with four chains and 20,000 iterations in each chain, with the first 2,000 iterations in each chain being discarded as the warm-up phase.

Results

Comprehension question accuracy (Experiments 1a and 1b combined)

After exclusion of participants with accuracy below 70 %, overall accuracy (including fillers) was good; on average 85.9 % (range: 71.9 - 100 %). Accuracy for the critical items was 81.1 % (range: 20 - 100 %). Condition-wise accuracy is shown in Table 3.

Table 2: Interpretation of Bayes factors (Lee and Wagenmakers, 2014).

BF_{10}	Interpretation
> 100	extreme evidence for the effect
30 - 100	very strong evidence for the effect
10 - 30	strong evidence for the effect
3 - 10	moderate evidence for the effect
1 - 3	anecdotal evidence for the effect
1	no evidence
1 - 0.3	anecdotal evidence against the effect
0.3 - 0.1	moderate evidence against the effect
0.1 - 0.03	strong evidence against the effect
0.03 - 0.001	very strong evidence against the effect
< 0.001	extreme evidence against the effect

Table 4 shows the results of the generalized mixed model (log-odds scale) analyzing the comprehension accuracy for the critical items (Experiments 1a and 1b combined). The estimates and 95 % credible intervals show primarily a reduction in comprehension accuracy for high compared to low semantic interference conditions.

Since the accuracies were not of primary interest, we did not carry out Bayes factors analyses for these.

Self-paced reading times (Experiments 1a and 1b combined)

Estimated effect sizes and their uncertainty. Reading times across the whole sentence are shown in Figure 3, separately for high (panel A) and low (panel

Table 3: By-condition accuracy in critical trials in the SPR experiment (E1a and E1b combined).

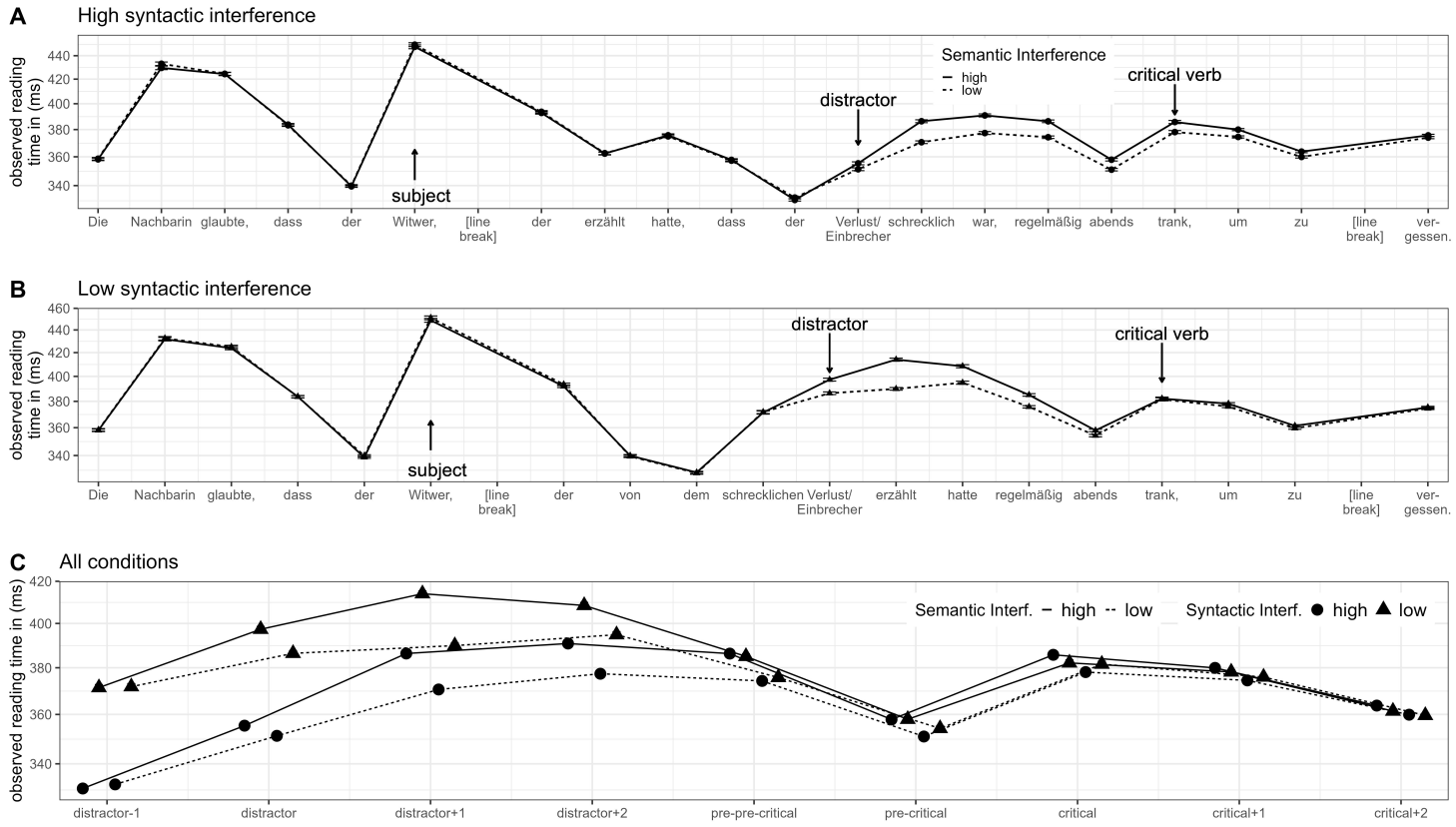
syntactic	semantic	accuracy %
low	low	86.0
low	high	77.7
high	low	85.8
high	high	74.9

Table 4: Results in log-odds from the Bayesian generalized model analyzing the comprehension accuracy in the SPR Experiments 1a and 1b.

	Estimate	95% CrI
Intercept	1.66	[1.38, 1.93]
syntactic	-0.08	[-0.16, -0.01]
semantic	-0.61	[-0.69, -0.54]
interaction	-0.10	[-0.22, 0.02]

B) syntactic interference because these conditions had partially different sentence structures.

Figure 3: Self-paced reading times with 95% confidence intervals. Panel A and B show the pooled reading times across the whole sentence; separately for high (A) and low (B) syntactic interference due to differing sentence structure. Panel C shows the reading times of the sentence focusing on the regions between the distractor and critical verb for all conditions.



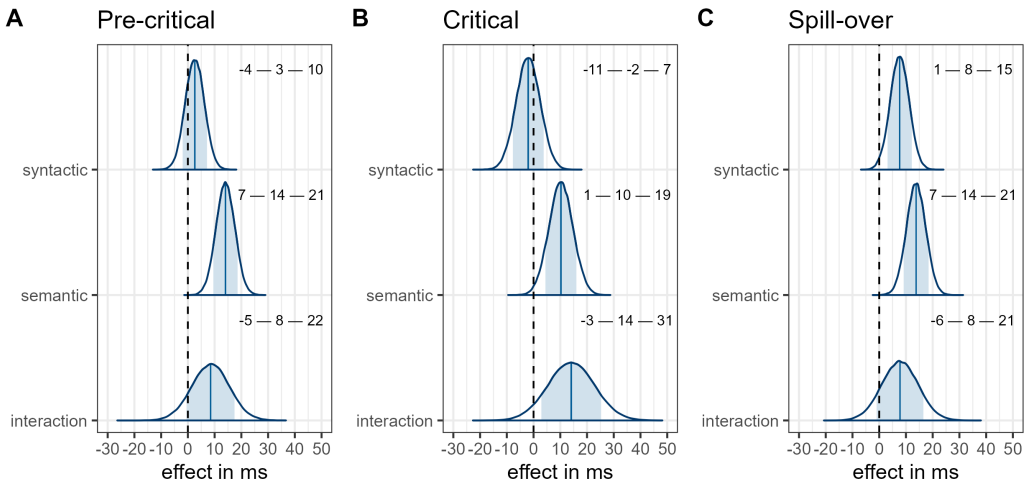
It is apparent from Figure 3 that, regardless of the syntactic manipulation, at the distractor, reading times between high and low semantic interference started to differ: distractors in the high semantic interference conditions induced longer reading times than in the low semantic interference conditions. To anticipate our discussion of the SPR results here, we believe that the difference in reading times starting at the distractor was caused by encoding interference (Oberauer and Kliegl, 2006). Additionally, the second session in Experiment 1a showed shorter reading times compared to the first session. As mentioned earlier, this attenuation in reading times was likely due to adaptation to the SPR task in the second session.

Crucially, the reading times difference starting at the distractor persisted in the following regions. The distractor and the immediately following regions differed between conditions, therefore we did not analyze their reading times. The reading times difference starting at the distractor and persisting almost until the end of the sentence is problematic for the analysis and interpretation of reading times in the later regions, i.e., the critical region. Any effects that might be present there cannot be attributed clearly to the processing of the respective region. However, for comparability with previous work, we briefly report analyses of the reading times of the later regions, i.e., the pre-critical, critical and spill-over region.

Figure 4 shows that the posterior distributions for the parameters are very similar in the pre-critical, critical and spill-over regions. The estimates of semantic interference were positive in all regions and the estimates of the interaction were mostly positive in all regions. The estimates of syntactic interference were less consistent across regions. They were almost centered

around zero in the critical region, showed mostly positive values in the pre-critical region and fully positive values in the spill-over region.

Figure 4: Posteriors for the syntactic and semantic interference effects and their interaction at the critical verb and surrounding regions in the pooled self-paced reading data (Experiments 1a and 1b combined). The numerical values are the means and 95 % credible intervals. The blue vertical lines represent the median and the blue shaded areas are 80 % credible intervals.



The fixed effect of trial id which was included in the models as a covariate revealed a huge adaptation effect: From the first to the last trial (100th), reading times in the regions of interest decreased by on average around 250 ms (pre-critical region: -219 ms [-233, -204], critical region: -296 ms [-317, -276], spill-over region: -246 ms [-260, -233]). This adaptation effect was probably caused by increasing familiarity with the task and employed sentence structures over the time course of the experiment. While this adaption is interesting in itself, it was not the focus of the present study, therefore we will not discuss it further.

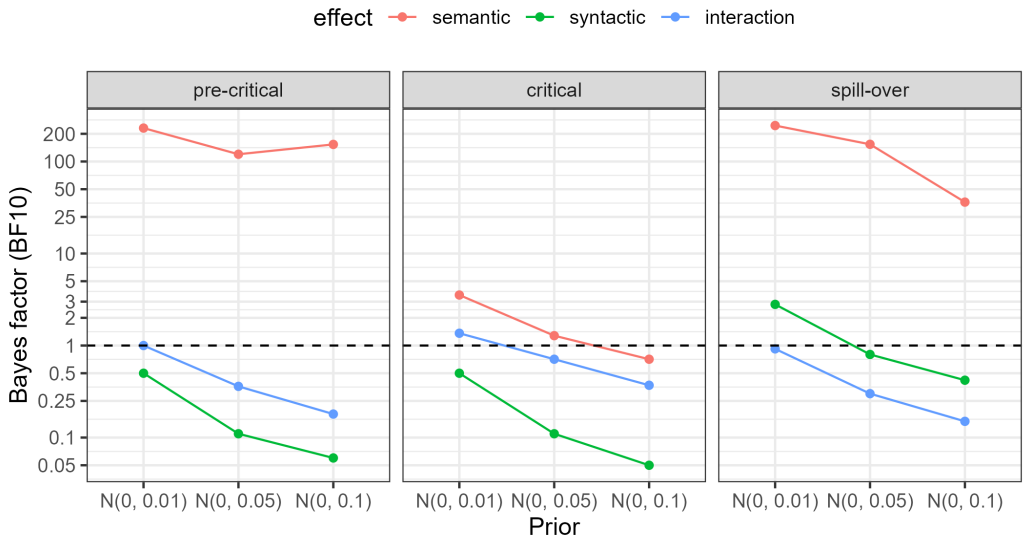
Hypothesis testing using Bayes factors. To complete the analysis of the effects in the critical and surrounding regions – although they are confounded by the long-lasting reading time differences starting at the distractor, we here present a Bayes factor analysis using all the data from Experiments 1a and 1b (see Figure 5). There was very strong to extreme evidence for the semantic interference effect in the pre-critical and spill-over region (BF_{10} between 36 and 231). In the critical region, the evidence for semantic interference was weak at best (Normal(0, 0.01): $BF_{10} = 3.5$, Normal(0, 0.05): $BF_{10} = 1.3$, Normal(0, 0.1): $BF_{10} = 0.7$). There was evidence against an effect of syntactic interference under almost all priors in all regions ($BF_{10} < 0.5$). Only in the spill-over region when small effects between -8 and 8 ms were assumed a priori, there was anecdotal evidence for syntactic interference ($BF_{10} = 2.8$). There was overall evidence against the interaction of syntactic and semantic interference ($BF_{10} < 1$). Under the narrowest prior in the critical region, the Bayes factor provided no evidence either way (neither for nor against the interaction ($BF_{10} = 1.4$)).

All in all, the Bayes factors provided very strong evidence for the semantic interference effect in the pre-critical and spill-over region, but at best weak evidence in the critical region. In contrast, the Bayes factors across all regions provided at best weak evidence for syntactic interference or even evidence against it and evidence against the interaction.

Discussion

We presented the to-date largest-sample self-paced reading study that aimed to investigate the use of syntactic and semantic features during subject-verb dependency formation. However, reading times started to differ before

Figure 5: Bayes factors for the effects of semantic interference, syntactic interference and their interaction in the reading times (combined Experiments 1a and 1b) at the critical verb and surrounding regions, under a range of priors on the target parameters.



this critical dependency could be formed. Starting at the distractor which intervened between subject and verb, high semantic interference conditions (animate distractors) were read slower than low semantic interference conditions (inanimate distractors). Because the distractor differed between conditions, we did not analyze distractor reading times. Instead, we treat the pre-critical region which was identical across conditions and preceded the critical verb as a proxy to investigate the difference starting at the distractor. Bayes factors provided extremely strong evidence for a semantic interference effect in the pre-critical region. In contrast, there was evidence against syntactic interference and an interaction in that region. We refrain from interpreting effects which were observed in later regions, i.e., the critical verb and spill-

over region, because it is unclear whether they were caused by the difference starting at the distractor or other processes. The reading times differences at the distractor and the pre-critical region are discussed below.

Reading times differences prior to retrieval

Animate distractors led to longer reading times than inanimate distractors (see Figure 3). One might argue that this could be due to differences in word length. The animate distractors in our materials were on average 1.1 letters longer than the inanimate distractors (mean length of animate distractors: 9.3, sd: 2.6, mean length of inanimate distractors: 8.2, sd: 2.9). It seems unlikely that this small difference of one letter in word length caused such large and long-lasting effects, but of course one cannot rule out this possibility with complete certainty.

Another possible explanation for the reading time differences starting at the distractor could be that the distractors induced a difference in plausibility between the high and low semantic interference conditions. To explore this possibility, we conducted a web-based plausibility judgement experiment on PCIbex (Zehr and Schwarz, 2018). Since we were only interested in the potential plausibility difference between the semantic interference conditions, we used a one-factorial design for the plausibility rating experiment investigating semantic interference within the high syntactic interference conditions (see 2a, repeated here as 5a for convenience).

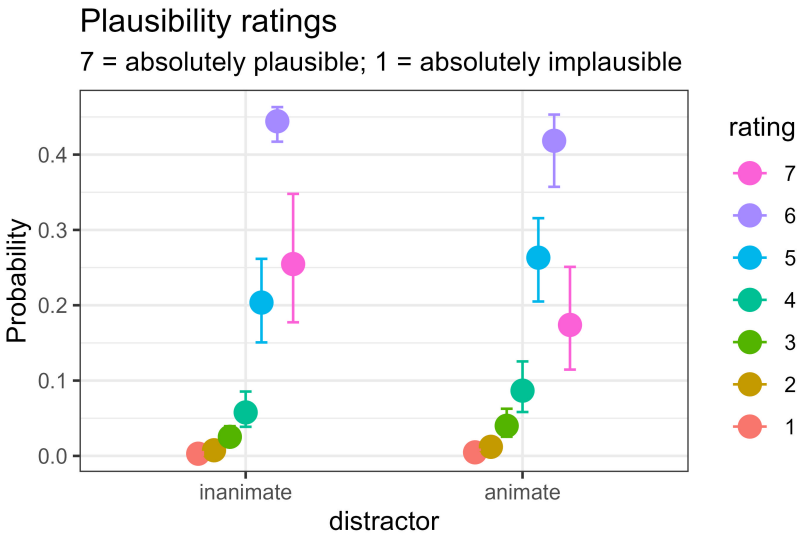
(5) Example item (critical word in bold, distractor in italics) of the present study:

a. High syntactic interference with high / low semantic interference:

Die Nachbarin glaubte, dass der Witwer, der erzählt
The$_{\text{FEM}}$ neighbor$_{\text{FEM}}$ believed that the widower who told
hatte, dass der *Einbrecher* / *Verlust* schrecklich war,
had that the burglar / loss awful was
abends regelmäßig **trank**, um zu vergessen.
in.the.evening regularly drank in.order to forget
'The neighbor believed that the widower, who had told her that the burglar / loss was awful, drank regularly in the evenings to forget.'

Forty-four native speakers of German (mean age: 26 years old, age range: 18 - 35 years, 21 female, 23 male), who were recruited over Prolific and had not participated in the SPR experiment or norming study, provided ratings on a scale from 1 (absolutely implausible) to 7 (absolutely plausible). The mean ratings per item ranged from 3.6 to 6.6. The experiment included 40 implausible filler sentences which were created by exchanging nouns between the fillers used in the SPR experiment (e.g., Der Sohn des Kaisers, der den Feldzug gewonnen hatte, plante äußerst strategisch, weil er die Gießkanne an sich reißen wollte., 'The son of the emperor who had won the campaign planned extremely strategically because he wanted to take the watering can by force.'). The mean rating of the implausible fillers ranged from 1.1 to 3.2. Participants spent approximately 37 minutes to complete the task and received £6.50 as compensation. The results of an ordinal brms model analyzing the ratings of the critical items are shown in Figure 6. Sentences with an animate distractor had a slightly lower probability to gain a high rating on the plausibility scale (β = -0.48, CrI [-0.67, -0.3]), i.e., the high semantic interference conditions were rated to be less plausible.

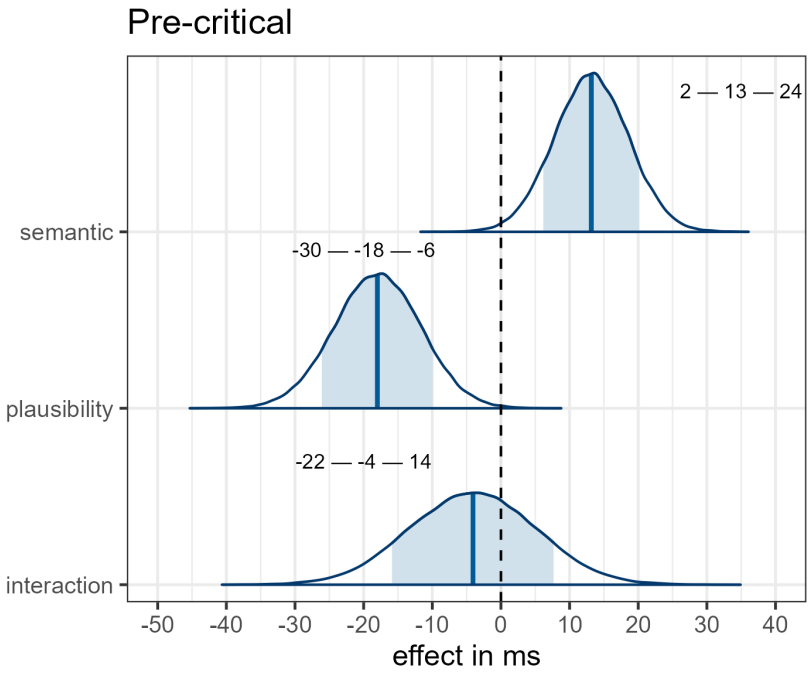
Figure 6: Posterior probability with 95% credible intervals of each rating for items with inanimate and animate distractors.



To investigate whether this plausibility difference was driving the reading time difference between the semantic interference conditions, we ran a Bayesian linear mixed model with log-normal likelihood on the reading times of the pre-critical region. This model included fixed effects for semantic interference (high +0.5, low -0.5), trial id (recoded to span from 0 to 1) and the centered plausibility ratings. We used the same priors as presented in Table 1. The model was run with full random effects, i.e., varying intercepts and slopes of all fixed effects and their interactions by participants and items. Figure 7 presents the posteriors from the model with the Normal(0, 0.05) prior on the slopes. Although plausibility influenced reading times (higher plausibility led to faster reading times, CrI [-30, -6] ms), there was an independent effect of semantic interference (CrI [2, 24] ms). The Bayes factors provided evidence for semantic interference in this analysis under all

priors except the most diffuse one (Normal(0, 0.01): $BF_{10} = 4.2$, Normal(0, 0.05): $BF_{10} = 2.4$, Normal(0, 0.1): $BF_{10} = 1.3$). The interaction of semantic interference and plausibility was centered around zero. In conclusion, the slowdown in the reading times results for high vs. low semantic interference starting at the distractor and lasting into later regions was partially caused by reduced plausibility of the high semantic interference conditions. However, the plausibility difference was not the only driver of the reading times slowdown, i.e., high semantic interference led to slower reading times independent of plausibility.

Figure 7: Posteriors of the semantic interference and plausibility effect and their interaction on the reading times in the pre-critical region. The numerical values are the means and 95 % credible intervals. The blue vertical lines represent the median and the blue shaded areas are 80 % credible intervals.



Since the reading times started to differ long before the critical verb which is assumed to be the locus of retrieval in the cue-based retrieval framework, retrieval interference cannot be the cause of this difference. The most likely explanation is encoding interference (Oberauer and Kliegl (2006), but see the potential confounds discussed above). The increased effort of encoding and subsequently maintaining representations of three different animate noun phrases (the introduction noun, the subject and the animate distractor) is the most likely reason for the slowdown in high vs. low semantic interference conditions (for similar findings, see e.g., Lago et al., 2021; Ness and Meltzer-Asscher, 2019, 2017; Kush et al., 2015a; Gordon et al., 2002).

Indeed, previous work using the same design as the present paper had also found semantic interference effects at the pre-critical region: both Van Dyke (2007) and Mertzen et al. (2023) found such effects. Mertzen et al.'s (2023) Figures 5 and 6 indicated reading time differences in earlier regions, especially for their German data (see their Figure 6). Van Dyke (2007) did not report reading times for the whole sentences / distractors; Van Dyke attributed the effects observed at the pre-critical region to plausibility differences between conditions, but as Mertzen et al. (2023) also pointed out, encoding interference could be an explanation even in that study. Given these earlier findings, our results are consistent with the encoding interference explanation.

Mertzen et al. (2023) observed both syntactic and semantic interference effects in the pre-critical region. In addition to encoding interference, Mertzen et al. (2023) proposed three other alternative explanations for their data: parafoveal-on-foveal effects, sentence structure confounds across conditions (two vs. one embedded clause between subject and verb in the high vs.

low syntactic interference conditions), and predictive processing effects. Our results cannot be explained by the parafoveal-on-foveal explanation, because there is no parafoveal preview in self-paced reading.

Regarding the possibility that sentence structure confounds caused the effects in the pre-critical region, this also seems implausible because the observed difference between the high and low semantic interference conditions is independent of the syntactic manipulation; it is the syntactic manipulation that has the confound. Therefore, the sentence structure confound is also not a good explanation for the semantic interference effect observed in our study.

Regarding the predictive-processing explanation, Mertzen et al. (2023) argued that the pre-critical adverb must attach to the upcoming verbal phrase; this leads to an anticipatory creation of a verb phrase chunk in memory, which triggers a retrieval of the subject already at the pre-critical region. However, in our study, the effect started at the distractor and became smaller as the two pre-critical adverbs were read (see Figure 3). The predictive-processing explanation would incorrectly predict that the effect begins at the first adverb, which is the first word of the verbal phrase. Consequently, the prediction explanation is also ruled out in our data, rendering encoding interference the most likely explanation of the effects found prior to the critical verb in our data. Crucially and as mentioned above, due to the long-lasting reading time difference starting at the distractor it is not possible to pinpoint the driver of effects observed in later regions. The effects in the critical region (and spill-over region) might be due to a retrieval initiated at the respective region, or still be due to encoding and maintaining the distractor in memory (Ness and Meltzer-Asscher, 2017), or a combination of encoding and retrieval

interference (Yadav et al., 2023).

We turn next to the event-related potentials experiment.

Experiment 2: EEG

Methods

Participants

146 participants from the University of Potsdam participant pool took part in the experiment. Three participants were excluded because they did not fulfill the demographic requirements (bilinguals or medical history). Four participants were excluded because they finished only one out of two experimental sessions. Seven participants were excluded due to EEG artifacts (below 20 artifact-free trials in at least one condition). Additionally, 29 participants were excluded because they showed poor comprehension question accuracy (below 70 %) in one of the experimental sessions. The data of 103 participants (mean age: 23.5 years old, age range: 18 – 38 years old, 81 female, 22 male) were used for the analyses presented here. These final participants were all right-handed, mono-lingual native speakers of German with normal or corrected-to-normal vision and no reported history of psychiatric or neurological disease. All participants gave written informed consent and were compensated with 40 Euros per experimental session (80 Euros in total) or course credit.

Procedure

The experiment was conducted in two experimental sessions for practical reasons (each of the sessions lasted approximately two hours). Sessions were

separated by one to eight weeks for each participant (the median gap was two weeks, the first and third quartiles being one and three weeks). The procedure of both sessions was identical and the same lists were used as in Experiment 1, with 60 critical sentences and 40 fillers per session.

During each experimental session, the EEG was recorded while participants were seated in a sound-proof booth. OpenSesame was used to present sentences word-by-word (Mathôt et al., 2012). Participants were familiarized with the procedure with two practice sentences. After that, the experiment was conducted in four blocks of 25 sentences each, presenting the items in pseudorandomized order, with breaks between the blocks. Each trial started with the presentation of a fixation cross in the center of the screen for 500 ms. Next, each word of the sentence was presented in the center of the screen. Word duration for words of interest (subject, distractor, pre-pre-critical word, pre-critical word, critical verb, post-critical word) was 500 ms. Word duration of all other words was 190 ms + 20 ms per character of the specific word. The inter-stimulus interval between all words was 400 ms. After a third of the trials, participants were asked to answer yes/no comprehension questions by pressing one of two keys on a standard keyboard. The j key was always mapped to "yes" answers and the f key was always mapped to "no" answers. The correct response was counterbalanced, so that half of the time a "no" response was correct and half of the time a "yes" response was correct.

EEG recording and processing

The EEG was recorded with 24 Ag/AgCl scalp electrodes, positioned according to the international 10-20 system. During recording, an electrode at the left mastoid was used as reference and AFz as ground. The sampling rate

was 500 Hz. Eye-movements were monitored with six electrodes which were positioned above, below and at the outer canthus of both eyes. Impedances of all electrodes were kept below 5 kΩ.

Processing of the EEG data was carried out with MNE python (Gramfort et al., 2013). The EEG was offline re-referenced to the average of the left and right mastoid electrodes. Eye-movements were corrected using independent component analysis (ICA) based on bipolar electro-oculogram channels. The data was band-pass filtered between 0.1 and 30 Hz. The data was segmented into epochs starting 200 ms preceding critical verb onset and lasting until 1000 ms following critical word onset. Epochs with artifacts were excluded automatically.

Statistical analyses

The statistical analyses of the comprehension accuracy in the EEG experiment were carried out in the same manner as the analyses of the comprehension accuracy in the SPR experiment.

We used Bayesian linear mixed models to analyze the single trial EEG data in response to the critical verb. We averaged the activity of 12 centro-parietal electrodes (Cz, C3/4, CPz, CP1/2, CP5/6, Pz, P3/4, POz) in the standard time windows of the N400 (300 to 500 ms post critical word onset) and P600 (600 to 900 ms post critical word onset) for all analyses. The models included sum-contrast coded fixed effects for syntactic interference (high +0.5, low -0.5), semantic interference (high +0.5, low -0.5) and their interaction. In addition to the fixed effects of interest, all models included the baseline EEG activity from 200 ms prior to critical word onset until critical word onset as a continuous predictor. This functioned as a regression-based

instead of traditional baseline correction (Alday, 2019). Varying intercepts for participants and items as well as by-participant and by-item random slopes for syntactic interference, semantic interference and their interaction were included. We used relatively informative priors for all parameters of the models (see Table 5).

Table 5: Relatively informative priors for the analysis of the event-related potentials (Nicenboim et al., 2023). The standard deviations of the slope priors were varied in order to conduct a Bayes factor sensitivity analysis. See the corresponding assumed a priori range of the difference between high vs. low interference.

Parameter	Prior	Assumed Range (μV)
Intercept	Normal(0, 5)	[-10, 10]
slope	Normal(0, 0.1)	[-0.2, 0.2]
	Normal(0, 0.5)	[-1, 1]
	Normal(0, 1)	[-2, 2]
	Normal(0, 2)	[-4, 4]
sigma	Normal(10, 5)	
SD	Normal(0, 2)	

Bayesian models were run with four chains and 20,000 iterations of which 2,000 were used as warm-up phase (Schad et al., 2022). For the calculation of Bayes factors and the corresponding sensitivity analysis, we defined a range of priors on the parameters of interest, assuming a range of effect sizes (Nicenboim et al., 2023; Schad et al., 2022). We chose slope priors which assume a wide range of effect sizes from $[-0.2, 0.2]$ to $[-4, 4]\,\mu V$.

Results

Comprehension question accuracy

After exclusion of participants with accuracy below 70 %, the overall accuracy (including fillers) was 82.8 % (range: [71.9, 100] %). Accuracy in critical trials was 75.6 % ([55, 95] %). By-condition accuracy is presented in Table 6.

Table 6: By-condition accuracy in critical trials in the EEG experiment.

syntactic	semantic	accuracy %
low	low	81.6
low	high	74.0
high	low	85.1
high	high	68.1

Table 7 presents the results of the generalized mixed model analyzing the comprehension accuracy in the critical trials. The log-odds estimates and 95 % credible intervals show reduced comprehension accuracy in high compared to low semantic interference conditions. Additionally, the results suggest an interaction, i.e., the difference between high and low semantic interference conditions was larger when syntactic interference was high vs. when it was low. Because the accuracies were not of primary interest, we did not carry out Bayes factors analyses for these.

Event-related potentials

Estimated effect sizes and their uncertainty. Figure 8 shows the grand average ERPs elicited by the critical verb for all conditions. Figure 9 shows the estimates for the brain activity in the spatio-temporal windows of the N400

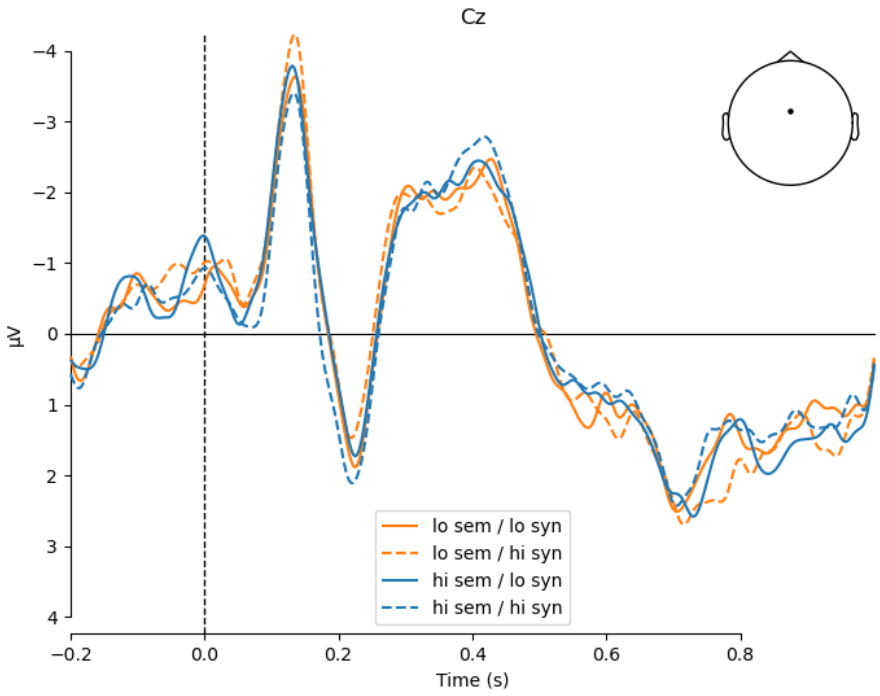
Table 7: Results in log-odds from a Bayesian generalized model analyzing the comprehension accuracy in critical trials of the EEG experiment.

	Estimate	95% CrI
Intercept	1.39	[1.10, 1.68]
syntactic	-0.04	[-0.16, 0.08]
semantic	-0.48	[-0.61, -0.36]
interaction	-0.17	[-0.34, -0.01]

and P600 from the linear mixed models with prior Normal(0, 0.5) which assumed interference effects with a difference possibly as large as $\pm 1\,\mu V$. The 95% credible intervals of the semantic interference effect showed more negative brain responses for high vs. low interference for both the N400 and the P600. The 95% credible intervals of the syntactic interference effect also showed more negative brain responses for high vs. low interference, but the mean estimate was about half the size of the semantic interference estimate, for both the N400 and the P600, respectively. The 95% credible interval of the interaction was centered around zero for the N400. For the P600, it included mostly negative values.

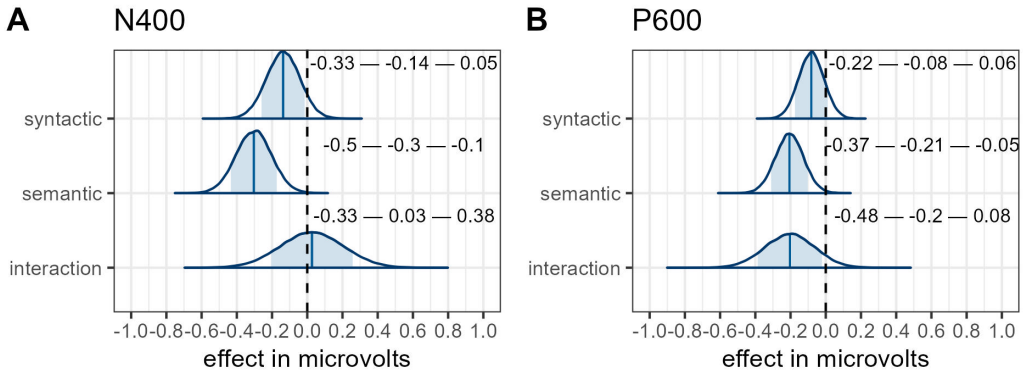
The semantic interference effect is further illustrated in Figure 10 (a) and (b). Figure 10 (a) shows that around 400 ms, the critical verb elicited a more negative ERP under high than under low semantic interference. Figure 10 (b) shows that the topography of the effect is rather broad but with a concentration of more negative values at centro-parietal electrodes. The timing and topography of this effect suggested that it was an N400 effect; thus, the N400 was modulated by semantic interference. Furthermore, Figure 10 (a)

Figure 8: ERPs elicited by the critical verb (word onset at 0 ms) at electrode Cz.



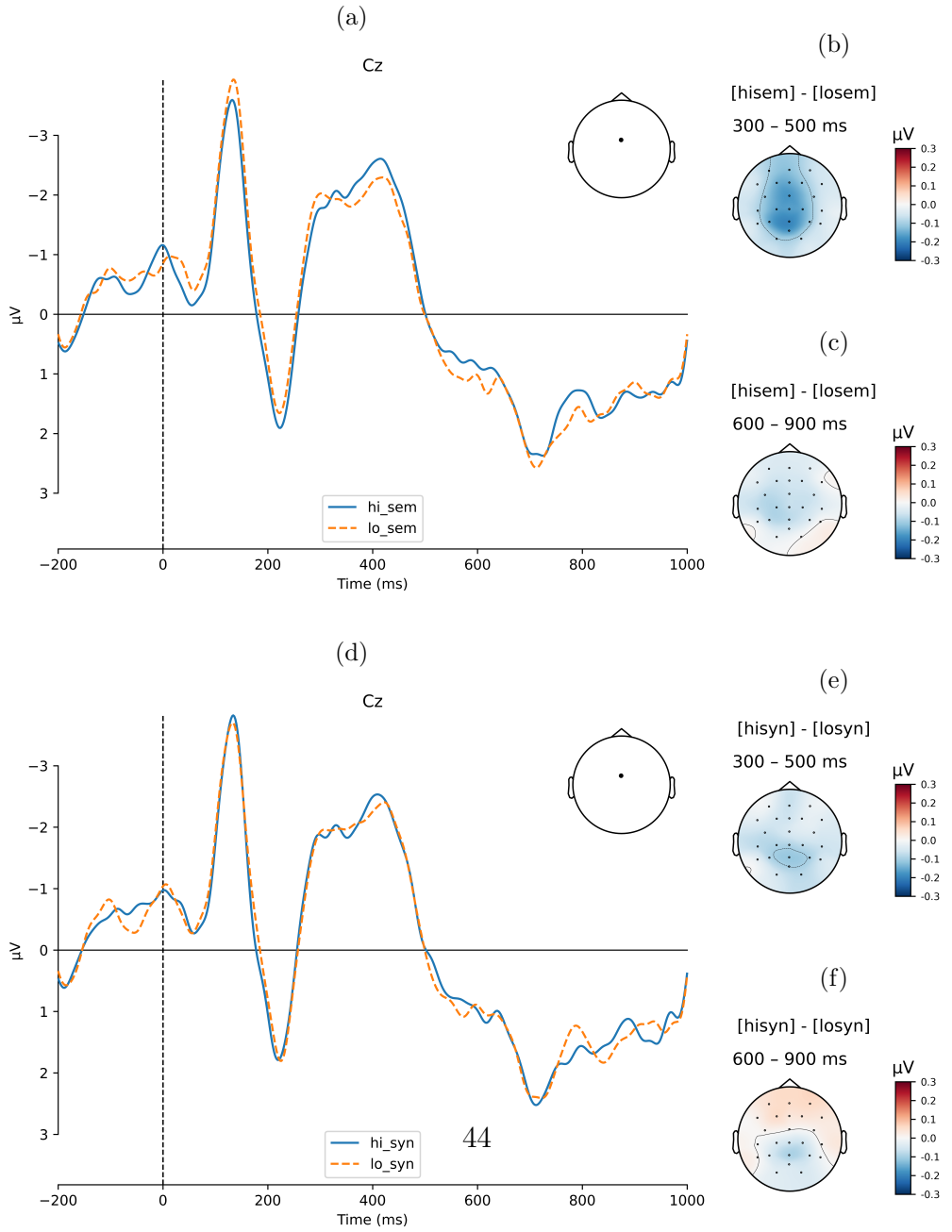
shows a slightly reduced amplitude for high vs. low semantic interference in the P600 time window. Figure 10 (c) shows that this effect had a broad distribution. Figure 10 (d) shows that syntactic interference affected the brain response to the critical verb in the N400 and P600 spatio-temporal windows to a smaller extend than semantic interference did. Figures 10 (d) and (e) show that high vs. low syntactic interference led to a slightly more negative ERP response with a broad distribution in the N400 time window. Figure 10 (f) shows a slightly reduced P600 amplitude for high vs. low syntactic interference with a centro-parietal distribution.

Figure 9: Posteriors of the syntactic and semantic interference effects and their interaction for the event-related potentials elicited by the critical verb in the spatio-temporal windows of the N400 (A) and P600 (B). Blue vertical lines represent the median and shaded areas are 80 % intervals.



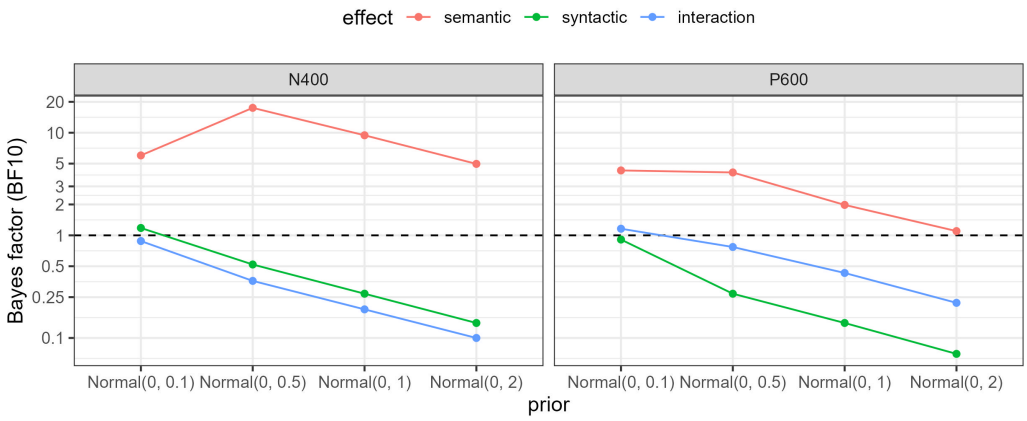
Hypothesis testing using Bayes factors. Bayes factors are shown in Figure 11. In the N400 spatio-temporal window, the Bayes factors provided moderate to strong evidence for the semantic interference effect under all priors (Normal(0, 0.1): $BF_{10} = 6$, Normal(0, 0.5): $BF_{10} = 17.5$, Normal(0, 1): $BF_{10} = 9.4$, Normal(0, 2): $BF_{10} = 5$). In the P600 spatio-temporal window, the Bayes factors provided anecdotal to moderate evidence (BF_{10} between 2 and 4.3) for the semantic interference effect under priors assuming effects smaller than $\pm 2\,\mu V$ and no evidence for it ($BF_{10} = 1.1$) under priors assuming effects in the range of $\pm 4\,\mu V$. By contrast, Bayes factors provided either no evidence for or even evidence against syntactic interference in both spatio-temporal windows ($BF_{10} < 1.2$). Similarly, Bayes factors provided either no evidence for or even evidence against the interaction in both spatio-temporal windows ($BF_{10} < 1.2$). In sum, in our ERP data, there was decisive evidence for the semantic interference effect and mainly evidence against syntactic

Figure 10: Brain responses elicited by the critical verb. (a) ERPs for semantic interference (word onset at 0 ms) at electrode Cz. Topographic maps of the semantic interference effect (high semantic interference - low semantic interference) in the N400 time window (b) and P600 time window (c). (d) ERPs for syntactic interference (word onset at 0 ms) at electrode Cz. Topographic maps of the syntactic interference effect (high syntactic interference - low syntactic interference) in the N400 time window (e) and P600 time window (f).



interference and the interaction.

Figure 11: Bayes factors for the effects of semantic interference, syntactic interference and their interaction in the N400 (300 - 500 ms post critical verb onset) and P600 (600 - 900 ms post critical verb onset) time windows at centro-parietal electrode sites.



Discussion

We presented a large-scale ERP experiment investigating syntactic and semantic interference during subject-verb dependency completion. To our knowledge, this is the first ERP experiment on this classic interference design, which has only been investigated using reading studies so far (Mertzen et al., 2023; Van Dyke, 2007).

Comprehension accuracy was reduced for high vs. low semantic interference conditions. There was no indication that syntactic interference affected accuracy. The sign of the effect was in the expected direction, but amounting to a 1% difference on the probability scale between the high and low syntactic interference conditions. There was a numerical suggestion of an interaction: the difference in accuracy between low and high semantic interference con-

ditions was larger when syntactic interference was high vs. when it was low (17 vs. 8 % difference).

The Bayes factors analyses found moderate to strong evidence for semantic interference in the N400 amplitude for the whole range of investigated effects sizes (± 0.2 to $\pm 4\,\mu V$). In the P600 spatio-temporal window, there was anecdotal to moderate evidence for semantic interference. Both the N400 and P600 were more negative under high vs. low semantic interference. In contrast, there was no evidence for effects of syntactic interference and the interaction or even evidence against them in both the N400 and P600 windows. In sum, the ERP results showed clear evidence for semantic interference and mostly evidence against syntactic interference and the interaction of syntactic and semantic interference. Consequently, we conclude that the ERP data showed no syntactic interference and no interaction.

Our finding that retrieval interference modulates the N400 amplitude supports the N400 as a marker of memory retrieval (Kutas and Federmeier, 2000, 2011; Brouwer et al., 2017; Lau et al., 2008). The more complex memory retrieval under high semantic interference resulted in a more negative N400. This finding is in line with previous ERP studies on interference (Lee and Garnsey, 2015; Vasishth and Drenhaus, 2011; Martin et al., 2014; Schoknecht et al., 2022). The semantic interference effect on the P600 is in the opposite direction than expected: High semantic interference led to a reduced P600 amplitude compared to low semantic interference. We provide a speculative interpretation of this unexpected result. Typically, the P600 amplitude is more positive when syntactic processing is complex, e.g., due to reanalysis (Osterhout and Holcomb, 1992). The reduced P600 amplitude under high

semantic interference in the present study might be explained by facilitatory interference (see e.g., Jäger et al., 2017). We are not aware of any other study that found inhibitory as well facilitatory interference in the same sentence configuration. Facilitatory interference is typically only found in ungrammatical sentences. Nevertheless, we think that the reduced P600 amplitude for high vs. low semantic interference in our data could be explained by facilitatory interference. In the present design, subject-verb integration might be facilitated, i.e., easier, under high semantic interference because there are two semantically suitable candidates to function as the subject. Even if the wrong noun, i.e., the distractor, was retrieved, it could be easily integrated with the verb. In contrast, misretrievals of the distractor in the low semantic interference conditions which might happen due to noisy processing would require reanalysis because the distractor did not match the semantic requirements of the verb. A similar effect was found by Tanner et al. (2017). In their study, the P600 elicited by ungrammatical verbs which were incongruent in number was reduced when the sentence included a matching distractor. However, the present study investigated only grammatical subject-verb integration and the size of the P600 effect was rather small. The small size of the effect on average could indicate that the facilitation occurred only occasionally when the distractor was misretrieved. Thus semantic cue overload would render memory retrieval more difficult causing a more negative N400 and if the distractor is falsely retrieved, integration with the verb would be easy resulting in a reduced, i.e., more negative, P600. This speculative claim would of course need to be tested in future work. Since the Bayes factor analysis provided stronger support for the semantic interference effect in the

N400 (largest $BF_{10} = 17.5$) than in the P600 (largest $BF_{10} = 4.3$), we focus on the N400 results in the remainder of this paper.

Similar to the self-paced reading results, there was clear evidence for semantic interference, and no evidence for syntactic interference in our ERP data. How do these observed patterns compare to the theoretical predictions of cue-based retrieval theory? We turn to this point next by discussing the quantitative predictions of the Lewis and Vasishth (2005) model.

Quantitative predictions of the Lewis and Vasishth (2005) model for ERPs elicited in the present design

The Lewis and Vasishth (2005) model is generally used to predict reading times. Recently, Mertzen et al. (2023) presented simulations for the present design that predicted approximately equally sized effects of syntactic and semantic interference in the reading times of the critical verb. We found that the reading times at the critical verb were confounded by reading times differences starting earlier in the sentence, which were best explained by *encoding* interference (see also Mertzen et al., 2023). Because the Lewis and Vasishth (2005) model is a model of retrieval interference and not encoding interference, and because we cannot disentangle encoding and retrieval interference in the present design, it would not make sense to compare the results from the present SPR experiment to the model's predictions. However, the ERP data can be compared to the model's predictions; all that has to be changed is that the measurement is now on the microvolt scale instead of the millisecond scale. In order to model the ERP data, we therefore adjust the scaling factor F, which rescales the activation of items in memory to the

relevant dependent measure. The modeling reported here is, to the best of our knowledge, the first time that the Lewis and Vasishth (2005) model has been fit quantitatively to ERP data.

We sampled the free scaling parameter F from a truncated normal prior distribution ($Normal_{lb=0,ub=0.05}(0.01, 0.01)$) with a lower bound of 0 and an upper bound of 0.05. This prior reflects the a priori belief that effects on the microvolt scale will range from 0 to 5 μV with higher likelihood for effect sizes below 3 μV (see Figure 12). This relatively uninformative prior was chosen because, although there exists no previously published ERP data on the Van Dyke (2007) design, it reflects the reasonable range of effect sizes in psycholinguistic ERP research. After sampling the scaling parameter, simulated data was generated from the model while holding all other parameters in the model at fixed values. We decided to model N400 amplitude differences because Bayes factors provided stronger evidence for the semantic interference effect on the N400 than on the P600 in the present ERP study. Since the scaling parameter F cannot be negative, an additional step was needed to reflect that larger predicted effect sizes correspond to more negative N400 amplitudes: The predicted effects from the model were multiplied with -1. In future work, the posterior distribution of the semantic interference effect in our ERP data, with range [-0.5, -0.1] μV, could be used as a basis for redefining a tighter prior on the scaling parameter F.

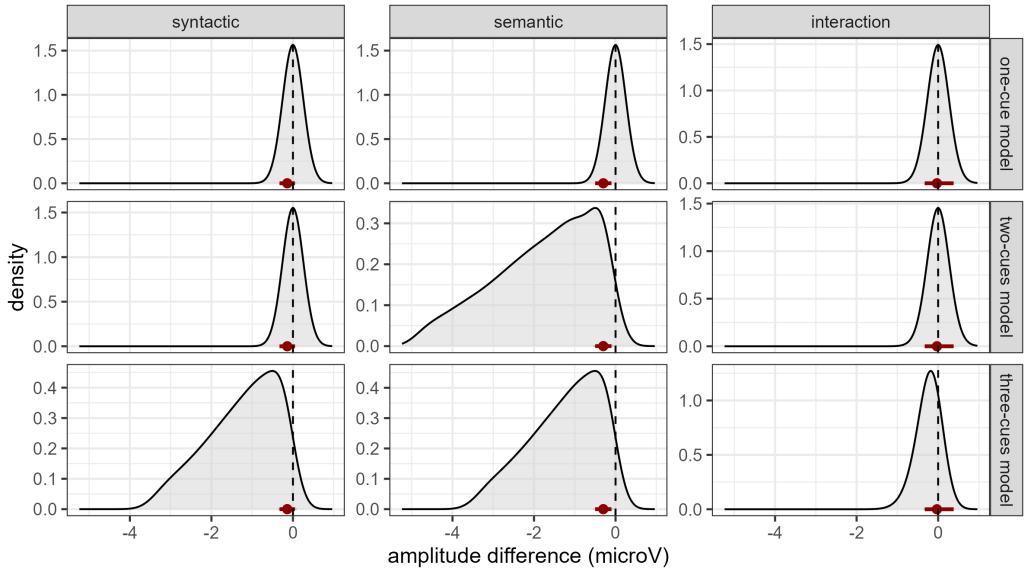
For their recent simulations, Mertzen et al. (2023) created a version of the Lewis and Vasishth (2005) model that assumes that in the Van Dyke (2007) design, three cues are used at the verb: {± animacy}, {± grammatical subject} and {± same clause}. The {± same clause} cue matches only

with nouns that occurred in the same clause as the verb. "The addition of the {± same clause} cue [was] necessary to identify the correct subject" and as a means for the model "to achieve correct retrieval" in the condition with high syntactic and high semantic interference (Mertzen et al., 2023, p. 14). This model predicted approximately equally sized effects of syntactic and semantic interference. Although it is clear that our ERP data does not support equally sized effects of syntactic and semantic interference, we decided to include their version of the model here because it has previously been claimed to be a good model for the present design. We have implemented two alternative versions of the model. Given that in our ERP data, there was no evidence for syntactic interference or depending on the a priori assumed effect size even evidence against it, it is plausible that, at least in this experimental design and language, syntactic interference plays little to no role. We speculate that this could be due to the parser tracking hierarchical structure and that this hierarchical information is used when searching for a noun in memory – even more so than suggested by Mertzen et al. (2023). So, we implemented a model which puts even more emphasis on hierarchical syntactic structure and thereby basically eliminates syntactic interference. Specifically, the relevant syntactic cues that might be used at the verb are not {± grammatical subject} and {± same clause} – as implemented by Mertzen et al. (2023) – but rather a composite cue {± subject-in-same-clause}. As already noted by Mertzen et al. (2023), structural information like [+ subject of clause$_i$] is not an inherent feature like for example [+ animate], but within in the cue-based retrieval framework, as a simplification, we assume that such a feature is encoded into memory and can consequently be used

to match the retrieval cue {± subject of clause$_i$}. This idea goes back to an assumption in the Lewis and Vasishth (2005) model by making the model "sensitive to whether or not it is parsing an embedded clause" (Lewis and Vasishth, 2005, p.388) and implies a simple clause tracking mechanism. The resulting model assumes two retrieval cues: {± subject-in-same-clause} and {± animate}. This proposal is furthermore in the spirit of the Dillon et al. (2013) and Sturt (2003) claim, that the parser intelligently targets only the syntactically relevant noun during antecedent retrieval. In the present case, it is possible that the syntactically relevant noun is only the one in the same clause as the verb. For more discussion of the role of structural retrieval cues, see Kush et al. (2015b), Franck and Wagers (2020) and Arnett and Wagers (2017). The proposed model here using two cues predicts no syntactic interference and no interaction, but does predict semantic interference effects. Lastly, we implemented a model with one cue which unambiguously matches the subject. This cue could be the composite cue {± animate-subject-in-same-clause} or it could be even more specific. This model naturally does not predict any interference effects. It can be regarded the null model.

Figure 12 shows the prior predictive distributions of the effects of interest for the three competing models discussed above in comparison to our ERP data. The null model using one cue predicts no interference effects. The new two-cues model, including a syntactic cue that utilizes syntactic hierarchy, predicts no effect of syntactic interference and no interaction, but a semantic interference effect. The prior predictions from the three-cues model show equally large effects of syntactic and semantic interference and an interaction. The predictions of all models overlap with the effects of interest in the data.

Figure 12: The prior predictions on the microvolt scale for syntactic and semantic interference and their interaction from three versions of the Lewis and Vasishth (2005) cue-based retrieval model. The data with 95 % credible intervals are shown in red.



To quantitatively evaluate the relative predictive performances of the models on the semantic interference data, we used a Bayes factor analysis. Since Bayes factors are sensitive to the choice of priors on the parameters of interest (Kass and Raftery, 1995), we computed Bayes factors under six different priors for the scaling parameter F, $Normal(0.01,0.01)$, $Normal(0.01,0.02)$, $Normal(0.01,0.005)$, $Normal(0.02,0.01)$, $Normal(0.02,0.02)$, and $Normal(0.02,0.005)$.

The procedure for computing Bayes factors was as follows. The Bayes factor computation required us to estimate the marginal likelihood of the competing models. We used the Monte Carlo integration method to estimate the marginal likelihood of each model. For Monte Carlo integration, we used

the importance sampling estimator where a large number of proposals were obtained from an importance density $q(F)$ and the average $likelihood \times prior$ was calculated across all these proposals to obtain the approximate marginal likelihood. The importance density should be chosen such that it covers the entire parameter space where $likelihood \times prior$ is non-zero and does not draw too many samples from the space where $likelihood \times prior$ is zero. We estimated the marginal likelihoods using Monte Carlo importance sampling as follows.

First, we chose an importance density $q(F)$ for the parameter F from which we draw 0.2 million samples of F; we used a truncated normal distribution $Normal_{lb=0,ub=0.05}(0, 0.025)$ for the importance density. This importance density allowed us to obtain more proposals for smaller values of F (for $F < 0.03$) while maintaining fatter tails for the larger values of F ($F > 0.03$); effectively, the whole parameter space of F between 0 and 0.05 was explored to draw 0.2 million proposals. Then, for each of these proposed parameter values, we computed the likelihood and prior density. The estimated marginal likelihood of a model will be given by

$$ML = \frac{1}{n}\sum_{i=1} n\frac{\mathcal{L}(F_i|y)p(F_i)}{q(F_i)} \text{ where } F_i \sim Normal_{lb=0,ub=0.05}(0, 0.025) \quad (1)$$

where n is the total number of proposals; y is the observed data, i.e., the observed semantic interference effect in our case; and F_i is the i^{th} proposal from the importance density $Normal_{lb=0,ub=0.05}(0, 0.025)$. The term $\mathcal{L}(F_i|y)$ gives the likelihood of proposal F_i given the data y, $p(F_i)$ gives the prior density of F_i, and $q(F_i)$ gives the importance density of F_i.

The Lewis and Vasishth (2005) model is a complex, non-deterministic model which can be simplified such that a likelihood function represents the model's generative process (e.g., in Nicenboim and Vasishth, 2018; Lissón et al., 2021). However, it is also possible to carry out model comparison without recasting the model as a likelihood function. In this case, model comparison is carried out using Approximate Bayesian Computation (ABC) (Sisson et al., 2018; Palestro et al., 2018). Using this approach, the model-generated data conditional on $F = F_i$ was compared with the actually observed data y. If the model-generated data $x_i \sim Model(F_i)$ for F_i is close to the actual data y, then F_i has higher likelihood and vice versa. The likelihood term can be rewritten as $\mathcal{L}(F_i|y) = \Psi(d(x_i, y), 0, \delta)$, where the function $\Psi(d, 0, \delta)$ represents a density kernel that weights a proposal F_i based on the distance d between model-generated data y and actual data x_i. We chose a Gaussian kernel, $Normal(0, \delta)$ where $\delta = 0.01$, such that the distance between model-generated data and observed data $d(x_i, y)$ is assumed to come from the distribution $Normal(0, 0.01)$. The δ parameter determines how good the approximation is: the lower the value of δ, the better the approximation of the likelihood. However, if δ is too low, the computational demand increases because too many proposals would end up with zero likelihood. We chose a reasonably small value of δ to obtain a good approximation of the likelihood without compromising on computational efficiency.

From the marginal likelihoods, we computed Bayes factors for the two-cues model compared to the three-cues model and the one-cue model. That is, the marginal likelihood of the two-cues model was divided by (i) the marginal likelihood of the three-cues model, and (ii) the marginal likelihood

of the one-cue model.

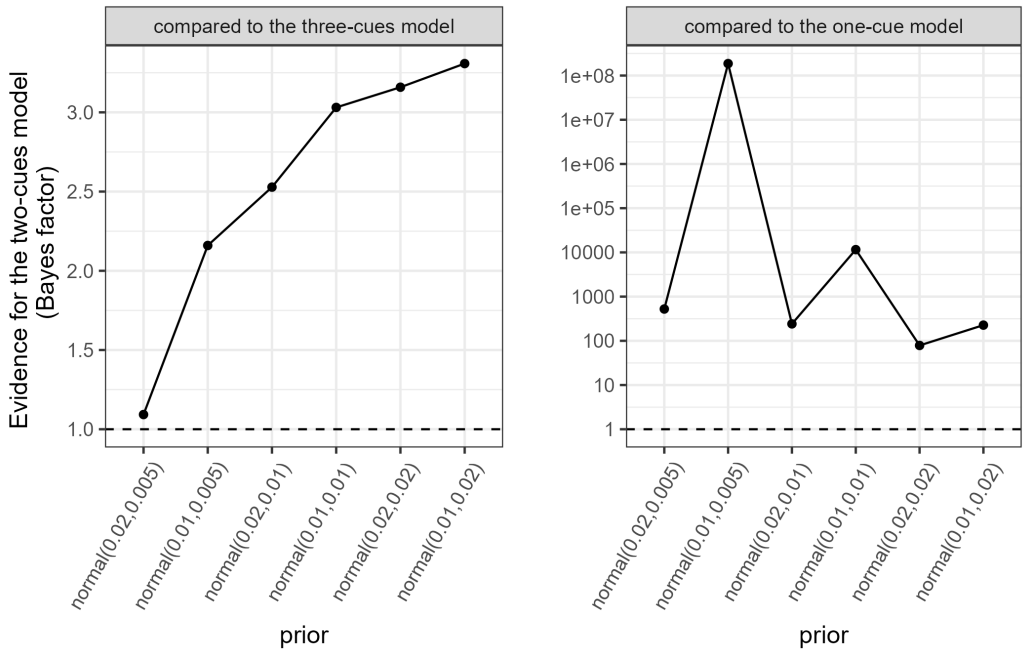
The Bayes factors under different priors on F are shown in Figure 13. The Bayes factors indicate strong evidence for the new two-cues model compared to the one-cue model (null model); this comparison shows that there is strong evidence for interference effects. Compared to the three-cues model proposed by Mertzen et al. (2023), the Bayes factors provide either no evidence for the two-cues model or anecdotal to moderate evidence for the new two-cues model depending on the prior (see Table 2 for the interpretation of Bayes factors). This model comparison shows that there is some support for the idea that the relevant syntactic cue is the {±subject-in-same-clause} cue – rather than the two separate, previously assumed cues {±grammatical subject} and {±same clause}. This result holds for prior assumptions that allow higher probability mass for small values of the scaling parameter $F.$, i.e., for $F < 0.01$.

Thus, the theoretical assumption that is consistent with the observed ERP data is that although the parser uses the animacy cue exactly as proposed in previous work on cue-based retrieval, the syntactic cue keeps track of the clausal location of the subject. This is of course a speculative claim that would need to be further investigated empirically.

General discussion

Taken together, the present SPR and ERP experiments using the 2 × 2 interference design developed by Van Dyke (2007) consistently showed effects of semantic interference and little or no effects of syntactic interference or an interaction. This pattern was not only found in the dependent measures of

Figure 13: Comparison of the predictive accuracies of the three versions of the Lewis and Vasishth (2005) cue-based retrieval model on our ERP data. The facets show how the new two-cues model performs compared the the three-cues model and the one-cue model, respectively. Larger Bayes factors correspond to better performance of the two-cues model.



the experiments that were of primary interest, but also in the comprehension accuracy. The modeling reported here suggests that one way to explain the observed pattern in the context of cue-based retrieval is to assume the following: the parser searches for a subject that is within the same clause, but uses the animacy cue without reference to the clause that a noun appears in. We revisit the connection between syntax and semantics below. A broader implication is that the human sentence comprehension system may be using the hierarchical syntactic structure to selectively target nouns for retrieval; this is an idea that has independent support in the literature (e.g., Sturt,

2003; Dillon et al., 2013; Yadav et al., 2022).

In the remainder of this section, we first provide an in-depth comparison of the present results to the previous studies' results which used the same design. Then we discuss two important issues that relate to the broader literature. The first is the role of encoding vs. retrieval interference, and the second is the use of syntactic and semantic information during sentence processing.

Comparison with previous findings

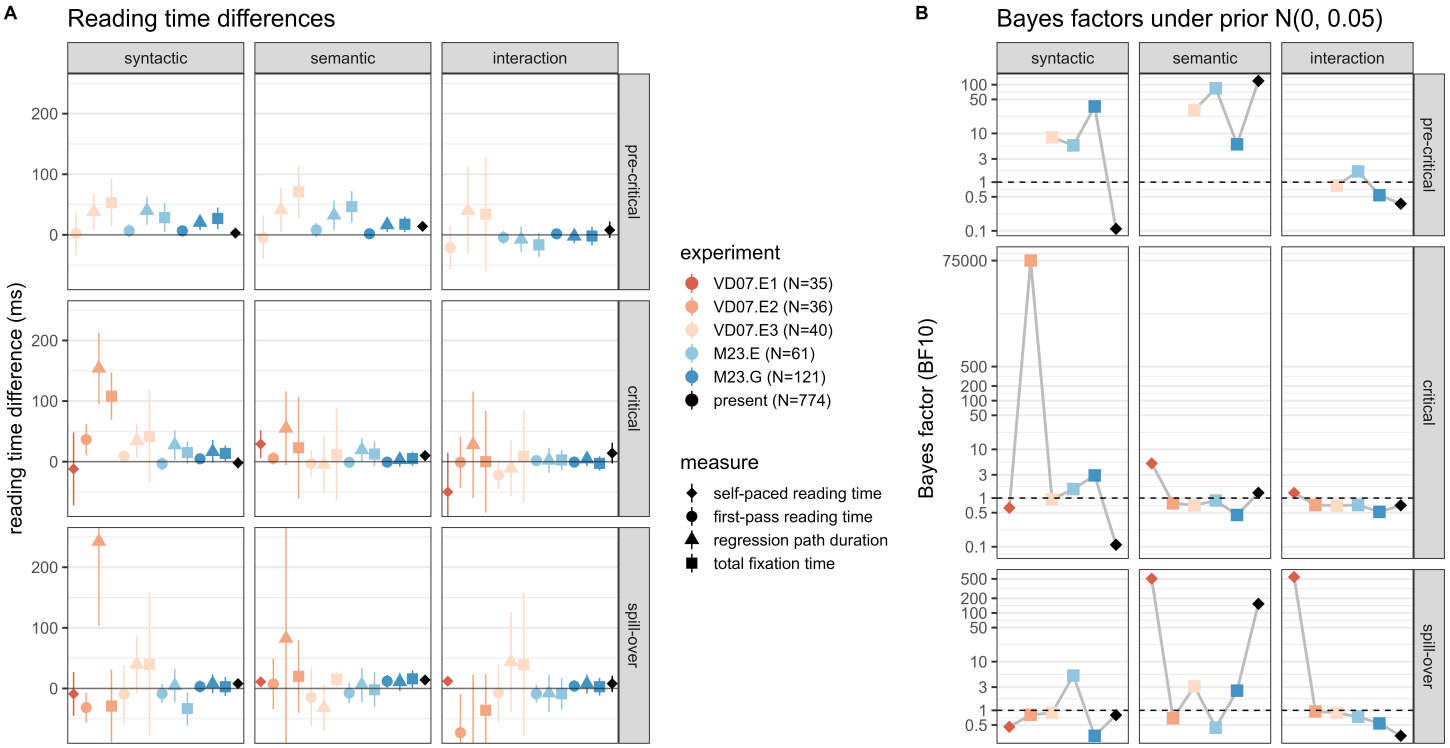
The present study used the same design as two previous studies (Van Dyke, 2007; Mertzen et al., 2023). Figure 14 A shows the estimated reading times differences from the previous studies and the present self-paced reading experiment. Because the Van Dyke (2007) data was not available to us, we derived estimates for their study from the summary statistics in the published paper. Mertzen et al. (2023) made their data available. In comparison to the previous estimates, the effects of the present study were relatively small and associated with less uncertainty in general. We used Bayes factors to compare whether there was evidence for the estimated effects in the previous studies vs. the present study. For maximal comparability, we used only the self-paced reading times and total fixation times from the previous studies for this comparison. Bayes factors for Van Dyke (2007) were computed using the Normal-Normal conjugate case (Lee, 2012) to derive posteriors from the summary statistics (mean and SE) in the published paper and a reasonable prior (Normal(0,0.5), which amounts to an a priori assumed effect size in the range of -40 ms to 40 ms). Bayes factors for Mertzen et al.'s (2023) data were computed from hierarchical mixed models which we ran on their data analo-

gously to our own analyses described in Statistical Analyses for Experiments 1a and b.

The Bayes factors for the effects of interest of the present and previous studies under prior Normal(0,0.5) are shown in Figure 14 B. The region-wise Bayes factors across studies were consistent with only a few exceptions (e.g., the majority of studies showed no evidence for the interaction), but in some cases there was no clear clustering in favor or against an effect (e.g., regarding semantic interference in the spill-over region). Only once did the present study deviate from a clear result if it was consistent across the previous studies: All previous studies showed evidence for syntactic interference in the pre-critical region, while the present study provided evidence against the effect. This difference in the results between the present and previous studies could be explained by differences in i) methodology, ii) items and/or iii) statistical power. We discuss each of these differences in turn.

An obvious difference between the studies is the experimental methodology. The majority of the previous experiments employed eye-tracking while reading, while we used self-paced reading and EEG / ERPs. However, word-by-word self-paced reading (SPR) and ERPs elicited by single words presented in the typical rapid serial presentation (RSVP) format, if anything might have a greater tendency to be affected by interference compared to natural reading (as in eye-tracking studies), due to the restricted reading format: Comprehenders have no opportunity to revisit previous material during conventional self-paced reading (but see Paape and Vasishth, 2022) and RSVP. Word-by-word presentation is likely more demanding on the comprehender's memory, which should make it easier to detect interference effects. So, the

Figure 14: A) Reading time differences in the regions of interest from the present and previous studies using the design by Van Dyke (2007). VD07.E1, VD07.E2 and VD07.E3 stand for Van Dyke's (2007) Experiments 1-3. The intervals for Van Dyke's (2007) estimates are 95% confidence intervals. M23.E and M23.G stand for Mertzen et al.'s (2023) English and German experiment, respectively. The intervals for their study and the present study are Bayesian 95% credible intervals. There are less estimates for the pre-critical region because Van Dyke's (2007) Experiment 1 and 2 did not have a pre-critical region. B) Bayes factors for the effects of interest in the present and previous studies under prior Normal(0,0.5).



differences in methodology do not offer a straightforward explanation for the lack of syntactic interference in our data compared to the previous studies. However, due to the demanding task comprehenders might trade off processing depth and adopt good-enough processing (Ferreira and Patson, 2007). Consequently, they would not fully parse embedded structures, hence there would be no syntactic interference from embedded subjects. Under this account, semantic interference might still arise because animacy information is readily available from the lexical entries with no need for detailed processing.

The difference in methodology led to another difference between our study and the previous studies. In their eye-tracking experiments, Van Dyke (2007) and Mertzen et al. (2023) presented the stimuli on a single line. Whether Van Dyke's (2007) SPR experiment used line breaks or not is not reported. In our SPR experiments, it was necessary to introduce line breaks. These were hard-coded to occur always after the subject of the critical verb and two words after the critical verb. The line break after the subject might have increased its prominence compared to the other nouns in the sentence, i.e., the distractor. It has been proposed by Engelmann et al. (2019) that prominence within the current context increases the activation of a memory representation which leads to faster retrieval and higher retrieval probability. Therefore, increased prominence of the subject could have decreased interference. Since we have found evidence for semantic, but not for syntactic interference, the increased prominence might have especially affected syntactic interference. For this point to hold, one would need to make the reasonable assumption that the line break provided additional structural information by emphasizing the syntactic structure of our sentences. However,

our EEG experiment used word-by-word presentation for all words, so there was no additional structural information which might have increased the prominence of the subject compared to the distractor. So, the line break might have decreased syntactic interference in our SPR experiments, but not in the EEG experiment. Nevertheless, in the present design, both the SPR and the EEG experiment showed predominantly evidence against syntactic interference.

We now turn to the differences regarding the experimental items. The three studies investigating syntactic and semantic interference with the same design used either English or German items. Cue-based retrieval as a theory on the memory processes during sentence comprehension implicitly states that these memory processes do not differ cross-linguistically (Lewis et al., 2006; Lewis and Vasishth, 2005). This is in line with no apparent clustering of the English vs. German results (see Figure 14 A). So, the use of materials in different languages cannot account for the differences in results between the studies.

The previous and present items differed in the pre-critical material. Van Dyke (2007) used one region consisting of two words (e.g., 'yesterday afternoon') which were analyzed together. Mertzen et al. (2023) used a one-word pre-critical region (e.g., 'tatsächlich', indeed). In the present study, we added a pre-pre-critical region to help us differentiate between the potential sources of the pre-critical effects in the literature as discussed in the Discussion of the SPR experiments. So, the present study was the only one which had a pre-critical region that did not directly follow a clause boundary and it was also the only one to not find syntactic interference in the pre-critical region.

These two observations combined suggest that the pre-critical syntactic interference effect in the previous studies might have been a spill-over effect from the syntactic manipulation. The syntactic interference manipulation in all the studies using this design was confounded with syntactic complexity of the material intervening between the subject and the (pre-)critical region. In the low syntactic interference conditions, it was a simple relative clause (who VP PP). In the high syntactic interference conditions, it was a relative clause with an embedded complement clause (who VP that NP VP). This difference in syntactic complexity could have "spilled over" into the pre-critical region of the previous studies causing the syntactic (interference) effect. While this seems like a plausible explanation for the previous effects, it would also apply to our pre-pre-critical region. However, Figure 3 C shows that the pre-pre-critical region in our items did not show a syntactic effect.

Another item-related difference is that all our experiments (SPR and EEG) displayed the subject with a comma. Commas are mandatory in German to separate subordinate clauses from the main clause. Similar to the line break discussed above, the comma could have emphasized the syntactic structure of the sentences and therefore decreased syntactic interference from the distractor which occurred in another clause than the critical verb. But if this comma provided information about the syntactic structure of the items then it should have done so in Mertzen et al.'s (2023) German study as well because they also presented the subject with a comma. But Mertzen et al. (2023) found syntactic interference and we did not. So, while it is plausible that commas provide structural information, it cannot explain the difference in findings between our study and the previous studies.

The contrast between our results and those from Mertzen et al.'s (2023) German experiment is surprising given that both used German items and partially even the same linguistic material. However, Mertzen et al. (2023) changed the structure of the relative clause modifying the subject by adding an additional animate distractor to all conditions (see an example of one of their English items in 6; their items had this structure cross-linguistically). This was done to "increase the strength of the manipulation." (Mertzen et al., 2023, p. 9).

(6) It turned out that the attorney whose secretary had forgotten that the visitor was important frequently complained about the salary at the firm. (Mertzen et al., 2023)

The additional distractor ('the secretary' in 6) was animate and the subject of the relative clause, therefore theoretically it should have affected both syntactic and semantic interference equally. But it is possible that it increased syntactic interference in Mertzen et al.'s (2023) experiment to a larger extent because the additional distractor ('the secretary') was a "stronger" subject than the manipulated distractor ('the visitor') because it was the subject of a full finite verb while the latter was always the subject of a verb phrase consisting of an auxiliar and an adjective. This additional distractor might have increased syntactic interference compared to our study. But it would have increased syntactic interference in comparison to Van Dyke's study as well and this was not the case (see Figure 14 A). All in all, differences in the items provide no clear explanation why the previous studies found syntactic interference and the present study did not.

A plausible explanation for the divergence between studies is a difference

in statistical power: The present study has higher statistical power than the previous studies. Van Dyke (2007) had 35 - 40 participants and 36 - 48 items in each of her three experiments, respectively. Mertzen et al. (2023) tested 61 English speakers and 121 German speakers with 40 items each. In contrast, in our SPR experiments, we tested 774 participants with at least 60 items each (a subset of 160 participants read all 120 items). Therefore, the present study is the study with the highest power to date using this 2×2 interference design. Consequently, the estimates from the present study are the most precise ones so far for this design. This is reflected by the tighter estimates in Figure 14 A. The smaller effect size in the present study is a consequence of higher power, and the magnitude of the effects observed is comparable to that of other large-scale reading studies (Nicenboim et al., 2018) and meta-analyses on interference (Jäger et al., 2017). In contrast, the large estimates of the previous studies are likely Type M errors, i.e., exaggerations, which are common in low-powered studies (Gelman and Carlin, 2014). That at least some of the previous studies had low power is further emphasized by the large uncertainty associated with some of the previous estimates (see Figure 14 A).

In sum, while there are differences in the employed methodology and items between the present and previous studies, the most likely cause of the different results is the difference in statistical power. However, an alternative explanation could be that syntactic interference might differ inter-individually (Yadav et al., 2022) and that the previous studies by chance sampled disproportionally many participants who showed syntactic interference, while we by chance sampled participants who did not.

While it seems unfortunate that the present and previous results together do not create a clear and unambiguous picture of the investigated phenomena, this divergence of results is not an anomaly within the scientific literature. In psycholinguistics alone, contradictory results have been reported for many phenomena: e.g., predictive parsing effects (Nieuwland et al., 2018; DeLong et al., 2005), reflexives processing (Jäger et al., 2020; Dillon et al., 2013), main verb/reduced relative garden paths (Trueswell et al., 1994; Ferreira and Clifton, 1986), local coherence (Tabor et al., 2004; Paape et al., 2025), locality effects in German (Levy and Keller, 2013; Vasishth et al., 2018), interference from extra-sentential distractors (Van Dyke and McElree, 2006; Mertzen et al., 2024), attenuation of agreement attraction due to case marking (Hartsuiker et al., 2001; Avetisyan et al., 2020), agreement attraction in grammatical sentences (Wagers et al., 2009a; Nicenboim et al., 2018) and relative clauses in Chinese (Hsiao and Gibson, 2003; Gibson and Wu, 2013; Vasishth et al., 2013; Jäger et al., 2015). The only way to resolve contradictions in the literature – like the contradictory results between the present and previous studies regarding syntactic interference – is by carrying out adequately powered studies in the future to reach consensus over time.

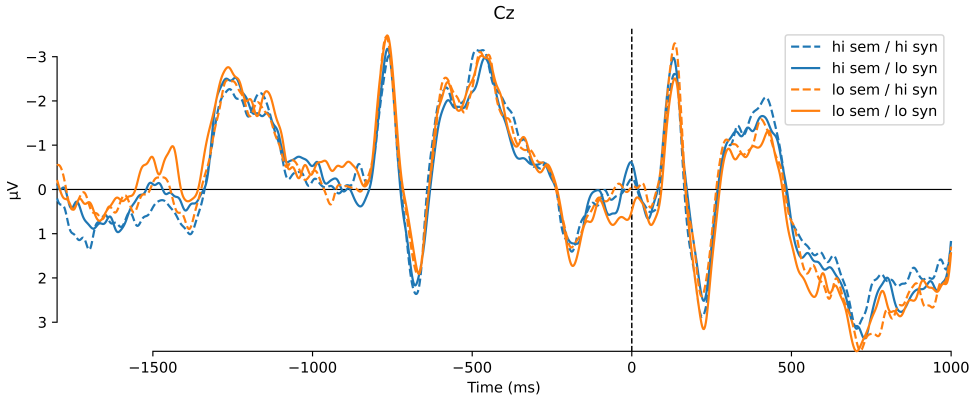
Encoding and retrieval interference

In the SPR data, we found that the semantic interference effect originated at the distractor and persisted throughout the following regions. Due to this time course, encoding interference (Yadav et al., 2023; Hammerly et al., 2019) is the best explanation for the interference effects in our SPR data. This raises the question whether the ERP effect was also caused by encoding interference, i.e., whether the ERP effect originated prior to the critical region. To answer

this question, it is necessary to look at the ERPs elicited by words earlier in the sentence. This is an unusual step in ERP research but makes sense as an exploratory step in the present context.

Figure 15 shows the ERPs for all conditions at the critical word and the two pre-critical adverbs which were identical across conditions. It is apparent from the ERPs of the pre-pre-critical and pre-critical word that the semantic interference effect found at the critical verb was not present at the words preceding it.

Figure 15: ERPs elicited by the pre-pre-critical adverb (word onset at -1800 ms), pre-critical adverb (word onset at -900 ms) and critical verb (word onset at 0 ms) at electrode Cz.

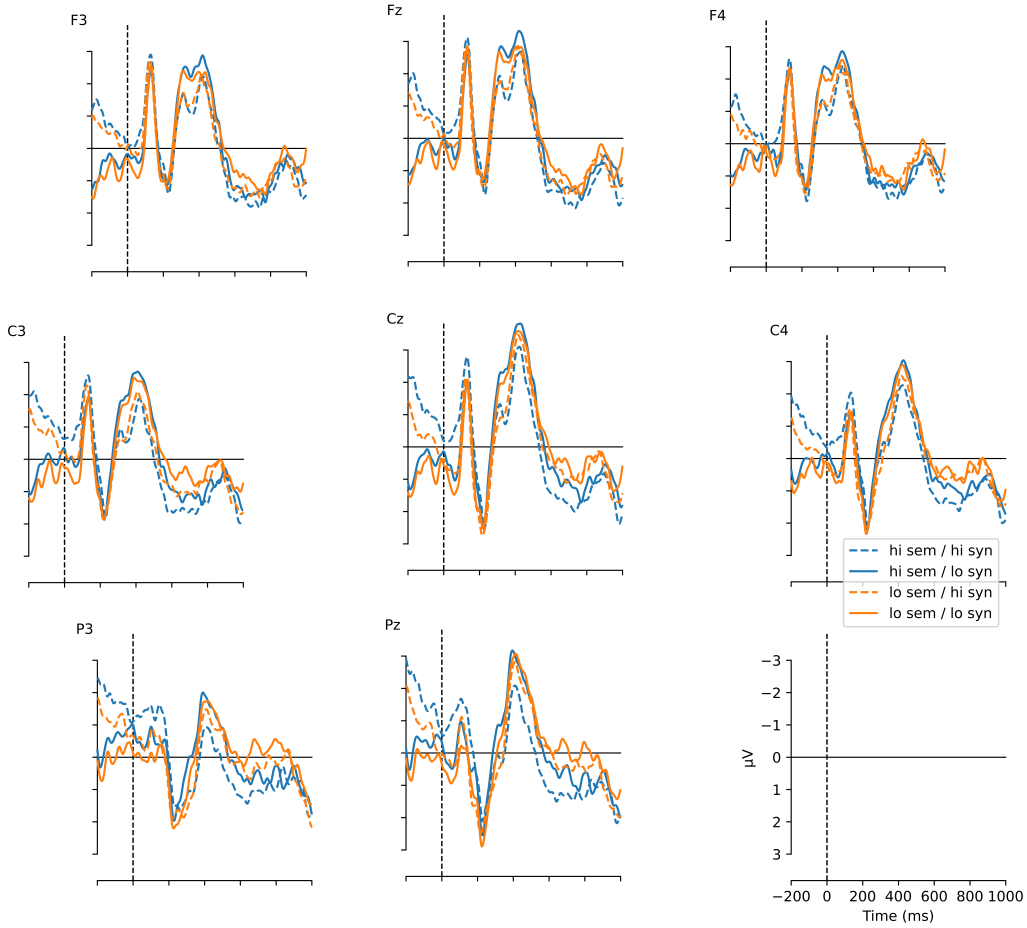


Thus, the words directly preceding the critical verb did not show a pattern consistent with semantic interference. A further question that arises is: which brain response was elicited by the distractor itself (see Figure 16)? An exploratory visual inspection of the ERPs elicited by the distractor shows several differences between conditions, but no semantic interference effect in the N400 spatio-temporal window.

We believe that the differences in the ERPs elicited by the distractor in the different conditions were caused by two factors. First, the high and low syntactic interference conditions had different word orders, i.e., the distractors were embedded in different phrase structures (simple noun phrase vs. noun phrase including an adjective within a prepositional phrase; see Example 2). This is also apparent from the large difference between high and low syntactic interference conditions before distractor onset. Second, the distractors were not identical across conditions; therefore, the ERPs in Figure 16 are elicited by different words in the high vs. low semantic interference conditions. In summary, this exploratory analysis suggests that the distractors did induce different brain responses across conditions, but these could be due to factors other than encoding interference. More importantly, these effects did not linger throughout the rest of the sentence (see Figure 15).

The comparison of the results of our SPR and ERP experiments suggests that the SPR results were strongly affected by encoding interference, while the ERPs elicited by the critical verb were not. This difference between methods might have been caused by the difference in presentation rate. In the self-paced reading experiment, as implied by the name, the participants had control over the presentation rate of the stimuli. The observed slowdown starting at the distractor might have resulted from the probably subconscious intent to slow down the sampling rate of the incoming stimuli to gain more time to process them. In contrast, in the ERP experiment, we used rapid serial visual presentation (RSVP), which presents stimuli at a fixed rate that is beyond the control of the participant. Therefore, they must adapt to the set rate and process the stimuli at the speed that the rate requires. Here, it

Figure 16: ERPs elicited by the distractor (word onset at 0 ms) at eight selected electrodes.



is noteworthy that the presentation duration of 500 ms per word and 400 ms between words which was used in the present ERP experiment is slower than usual natural reading speed. However, while in SPR the slowdown due to encoding and maintaining the distractor in memory overshadowed potential retrieval effects, the procedure of the ERP experiment might have prevented encoding interference from co-occurring with retrieval interference effects. In the ERP results, we observed an effect at the critical verb, which

could be attributed exclusively to the retrieval process hypothesized by cue-based retrieval accounts. Having said that, we cannot rule out the possibility that – due to the auto-paced nature of the ERP experiment – any encoding interference that started at or after the distractor is mixed in with the cost of retrieval interference. Indeed, recent modeling work on agreement attraction in reading (Yadav et al., 2023) has shown that some kind of feature distortion as well as cue-based retrieval are needed to explain the data-sets that were publicly available at the time that the modeling was done. If the Yadav et al. (2023) account extends to interference effects in general, it is reasonable to assume that the semantic interference effect observed at the verb in the present ERP study is driven by both encoding and retrieval interference. However, we cannot resolve the relative roles of these latent processes in the present study.

The use of syntactic and semantic information

In the sentence configurations that were investigated in the present study, we find a remarkable disconnect in the use of syntactic and semantic cues during subject-verb dependency resolution. The lack of decisive evidence for a main effect of syntactic interference for subject-verb dependency resolution in our study is not consistent with the assumption in previous work, e.g., Van Dyke (2007); Mertzen et al. (2023), that the parser searches for a subject simply by setting the retrieval cue {± grammatical subject}. A future avenue of research could be to find out what an appropriate syntactic manipulation would be to consistently trigger syntactic interference. This future work should utilize the findings from the literature on agreement attraction which has shown that the hierarchical position of the distractor can deter-

mine whether or not it affects processing (Franck et al., 2002, 2006; Franck and Wagers, 2020; Parker and An, 2018).

Although the syntactic cue included hierarchical information to match only with the correct subject within the same clause, this did not block the semantic cue from matching with the embedded distractor. This finding is not consistent with syntax-first models like garden-path theory (Frazier, 1987) which assume that syntactic processing precedes all other levels of linguistic evaluation (e.g., plausibility). Our findings rather suggest parallel, fully independent matching of different retrieval cues with the candidates in memory. In the modeling section above, we have already discussed how our results relate to predictions from the Lewis and Vasishth (2005) cue-based retrieval model assuming independent cues.

The semantic interference effect we found can be easily reconciled with other sentence processing theories that do not assign a special status to syntactic processing. Our results could be seen as consistent with the good-enough processing account (Ferreira and Patson, 2007), which assumes that comprehenders do not always aim to build a fully fleshed-out analysis of a sentence, but instead might accept incomplete, underspecified, or even incorrect analyses. The high number of participants that needed to be excluded in our experiments due to accuracy below 70 % (117 out of 908 participants in the SPR experiment, 29 out of 146 participants in the EEG experiment), suggests that our materials led to high processing demands. It is reasonable to assume that the participants might have – at least occasionally – adopted a good-enough processing mode when faced with high task demands (Swets et al., 2008; Logačev and Vasishth, 2015, 2016) or due to working memory

capacity limitations (von der Malsburg and Vasishth, 2013), or both. So, given a good-enough processing mode which leaves some syntactic relations within the sentence underspecified, it makes sense for the comprehender to use a simple heuristic for subject-verb dependency resolution. Animacy is an obvious choice for such a heuristic. Animate entities are proto-typical agents (Dowty, 1991) and therefore, the primary use of the {+ animate} cue to retrieve a subject leads to the correct analysis with high probability (not just within the experiment context, but also in everyday language use).

Finally, our findings are also consistent with the possibility that language processing relies predominantly on semantic associations to form (probabilistic) representation of event structures. This view has been put forward by Rabovsky et al. (2018), when they used the neural-network sentence gestalt model (McClelland et al., 1989), to model the N400 amplitude. This model assumes that language comprehension relies on associative form-to-meaning mapping instead of syntactic rules. This assumption is consistent with our results.

Without the modeling results reported above, it would be easy to conclude that our data show that syntactic cues play no role during subject-verb dependency resolution in the investigated sentence configurations. However, this conclusion would be problematic. First, as we showed above, the cue-based retrieval model suggests that – at least in our data – the {± grammatical subject} cue is used, but differently than previously assumed: the relevant cue may be {± subject-in-same-clause}. Second, there is plenty of independent evidence for the central role of syntactic information in comprehension; some examples are the role of case marking (e.g., Avetisyan et al.,

2020; Husain et al., 2014; Bhatia and Dillon, 2022; Bader et al., 2000; Bader and Bayer, 2006; Miyamoto, 2002), syntactic constraints like Principle A of the binding theory (e.g., Sturt, 2003; Dillon et al., 2013; Yadav et al., 2022), and the importance of word order (e.g., Meng and Bader, 2000).

There are of course important examples of the parser ignoring syntactic information. Three dramatic examples are the grammaticality illusion in multiple embeddings in English (Gibson and Thomas, 1999), syntactic local coherence effects (Tabor et al., 2004), and the agreement attraction phenomenon across different languages (e.g., Wagers et al., 2009b; Lago et al., 2021; Tucker et al., 2015). However, there exists evidence inconsistent with these findings. Regarding grammaticality illusions in double-center embeddings, the grammaticality illusion does not occur in German and Dutch (Vasishth et al., 2011; Frank et al., 2015b). Regarding syntactic local coherence effects, a large-sample study (Paape et al., 2025) presents evidence against syntactic local coherence. Regarding agreement attraction, there seems to be at least one language with rich case marking (Czech) that does not show convincing evidence of number agreement attraction (Chromý et al., 2023). Thus, it seems that, cross-linguistically, syntax does generally play a central role in building incremental structure and in completing dependencies. Moreover, it is entirely possible that a stronger syntactic manipulation than the one we used ends up showing a syntactic interference effect of the type that Van Dyke (2007) originally reported. Indeed, our own findings would need to be replicated, ideally by an independent research group, if we want to be sure that in the present design there is no evidence for syntactic interference. If such a replication attempt is carried out, it would also be useful to conduct

an lab-based self-paced reading study to investigate our speculation earlier that the absence of syntactic interference in our online SPR study may have been due to participants adopting a good-enough processing strategy, leading to only semantic interference being detectable.

Conclusion

This project investigated the use of syntactic and semantic cues during subject-verb dependency resolution with self-paced reading and event-related potentials. Both our experiments consistently showed semantic interference, i.e., more processing difficulty in the presence of a semantically matching distractor. This manifested in longer self-paced reading times starting at the distractor itself which persisted almost until the end of the sentence and in a more negative N400 amplitude for the high vs. low semantic interference conditions at the critical verb. The observed differences in the time course of the effect are likely due to the methods used in the two sets of experiments (self-paced vs. auto-paced reading), which might have lead to a different mixture of encoding and retrieval interference. Overall, in the present design, we found no decisive evidence for the use of a syntactic cue such as {± grammatical subject}, as was previously assumed in the literature; computational modeling shows that – at least given the present data – the parser uses the syntactic cue {± subject-in-same-clause} to identify the correct target for retrieval. A broader implication is that the parser may be able to use hierarchical structure in the input to target the correct syntactic dependent from memory, leading to no syntactic interference from a distractor that occurred within an embedded clause.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 317633480, SFB 1287 and the University of Potsdam, Germany. We thank Johanna Thieke, Romy Leue, Elise Oltrogge and Lisa Plagemann for assistance during data acquisition. We are grateful to Dario Paape for helpful discussions and three anonymous reviewers for thoughtful and constructive comments.

References

Alday, P.M., 2019. How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. Psychophysiology doi:10.1111/psyp.13451.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111, 1036 – 1060. doi:10.1037/0033-295X.111.4.1036.

Arnett, N., Wagers, M., 2017. Subject encodings and retrieval interference. Journal of Memory and Language 93, 22–54. doi:10.1016/j.jml.2016.07.005.

Avetisyan, S., Lago, S., Vasishth, S., 2020. Does case marking affect agreement attraction in comprehension? Journal of Memory and Language 112. doi:10.1016/j.jml.2020.104087.

Bach, E., Brown, C., Marslen-Wilson, W., 1986. Crossed and nested depen-

dencies in German and Dutch: A psycholinguistic study. Language and Cognitive Processes 1, 249–262. doi:10.1080/01690968608404677.

Bader, M., Bayer, J., 2006. Case and linking in language comprehension: Evidence from German. volume 34 of *Studies in theoretical psycholinguistics*. Springer. doi:10.1007/1-4020-4344-9.

Bader, M., Meng, M., Bayer, J., 2000. Case and reanalysis. Journal of Psycholinguistic Research 29, 37–52. doi:10.1023/A:1005120422899.

Bhatia, S., Dillon, B., 2022. Processing agreement in Hindi: When agreement feeds attraction. Journal of Memory and Language 125, 104322. doi:10.1016/j.jml.2022.104322.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2019. Toward a neurobiologically plausible model of language-related, negative event-related potentials. Frontiers in Psychology 10, 1 – 17. doi:10.3389/fpsyg.2019.00298.

Brouwer, H., Crocker, M.W., Venhuizen, Noortje J.and Hoeks, J.C., 2017. A neurocomputational model of the N400 and the P600 in language processing. Cognitive Science 41, 1318–1352. doi:10.1111/cogs.12461.

Bürkner, P.C., 2021. Bayesian item response modeling in R with brms and Stan. Journal of Statistical Software 100, 1–54. doi:10.18637/jss.v100.i05.

Chromý, J., Brand, J.L., Laurinavichyute, A., Lacina, R., 2023. Number agreement attraction in Czech and English comprehension: A direct experimental comparison. Glossa Psycholinguistics 2. doi:10.5070/G6011235.

Cunnings, I., Sturt, P., 2018. Retrieval interference and semantic interpretation. Journal of Memory and Language 102, 16–27. doi:10.1016/j.jml.2018.05.001.

DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word preactivation during language comprehension inferred from electrical brain activity. Nature Neuroscience 8, 1117 – 1121. doi:10.1038/nn1504.

Dickey, J.M., 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. The Annals of Mathematical Statistics 42, 204–223. doi:10.1214/aoms/1177693507.

Dickey, J.M., Lientz, B., 1970. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. The Annals of Mathematical Statistics 41, 214–226. doi:10.1214/aoms/1177697203.

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. doi:10.1016/j.jml.2013.04.003.

Dowty, D., 1991. Thematic proto-roles and argument selection. Language 67, 547–619. doi:10.2307/415037.

Engelmann, F., Jäger, L.A., Vasishth, S., 2019. The effect of prominence and cue association on retrieval processes: A computational account. Cognitive Science 43, e12800. doi:10.1111/cogs.12800.

Ferreira, F., Clifton, C., 1986. The independence of syntactic processing. Journal of Memory and Language 25, 348–368. doi:10.1016/0749-596X(86)90006-9.

Ferreira, F., Patson, N.D., 2007. The 'good enough' approach to language comprehension. Language and Linguistics Compass 1, 71–83. doi:10.1111/j.1749-818X.2007.00007.x.

Fine, A.B., Jaeger, T.F., Farmer, T.A., Qian, T., 2013. Rapid expectation adaptation during syntactic comprehension. PLoS ONE 8, e77661. doi:10.1371/journal.pone.0077661.

Franck, J., 2011. Reaching agreement as a core syntactic process: Commentary of Bock & Middleton *Reaching Agreement*. Natural Language & Linguistic Theory 29, 1071–1086. doi:10.1007/s11049-011-9153-1.

Franck, J., Lassi, G., Frauenfelder, U.H., Rizzi, L., 2006. Agreement and movement: A syntactic analysis of attraction. Cognition 101, 173–216. doi:10.1016/j.cognition.2005.10.003.

Franck, J., Vigliocco, G., Nicol, J., 2002. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. Language and Cognitive Processes 17, 371–404. doi:10.1080/01690960143000254.

Franck, J., Wagers, M., 2020. Hierarchical structure and memory mechanisms in agreement attraction. PLoS ONE 15, e0232163. doi:10.1371/journal.pone.0232163.

Frank, S.L., Otten, L.J., Galli, G., Vigliocco, G., 2015a. The ERP response

to the amount of information conveyed by words in sentences. Brain and Language 140, 1–11. doi:10.1016/j.bandl.2014.10.006.

Frank, S.L., Trompenaars, T., Vasishth, S., 2015b. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? Cognitive Science 40, 554–578. doi:10.1111/cogs.12247.

Frazier, L., 1987. Sentence processing: A tutorial review, in: Coltheart, M. (Ed.), Attention and Performance XII: The Psychology of Reading. Lawrence Erlbaum Associates, p. 559–586.

Freunberger, D., Roehm, D., 2017. The costs of being certain: Brain potential evidence for linguistic preactivation in sentence processing. Psychophysiology 54, 824–832. doi:10.1111/psyp.12848.

Gelman, A., Carlin, J., 2014. Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science 9, 641–651. doi:10.1177/1745691614551642.

Gibson, E., Thomas, J., 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. Language and Cognitive Processes 14(3), 225–248. doi:10.1080/016909699386293.

Gibson, E., Wu, H.H.I., 2013. Processing Chinese relative clauses in context. Language and Cognitive Processes 28, 125–155. doi:10.1080/01690965.2010.536656.

Gordon, P.C., Hendrick, R., Levine, W.H., 2002. Memory-load interference in syntactic processing. Psychological Science 13, 425 – 430. doi:10.1111/1467-9280.00475.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, C., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience 7. doi:10.3389/fnins.2013.00267.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M., 2004. Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441. doi:10.1126/science.1095455.

Hammerly, C., Staub, A., Dillon, B., 2019. The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. Cognitive Psychology 110, 70–104. doi:10.1016/j.cogpsych.2019.01.001.

Hartsuiker, R.J., Antón-Méndez, I., Van Zee, M., 2001. Object attraction in subject-verb agreement construction. Journal of Memory and Language 45, 546–572. doi:10.1006/jmla.2000.2787.

Hsiao, F.P.F., Gibson, E., 2003. Processing relative clauses in Chinese. Cognition 90, 3–27. doi:10.1016/S0010-0277(03)00124-0.

Husain, S., Vasishth, S., Srinivasan, N., 2014. Strong expectations cancel locality effects: Evidence from Hindi. PLoS ONE 9, 1–14. doi:10.1371/journal.pone.0100986.

Jäger, L.A., Chen, Z., Li, Q., Lin, C.J.C., Vasishth, S., 2015. The subject-relative advantage in Chinese: Evidence for expectation-based processing. Journal of Memory and Language 79, 97–120. doi:10.1016/j.jml.2014.10.005.

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:10.1016/j.jml.2017.01.004.

Jäger, L.A., Mertzen, D., Van Dyke, J.A., Vasishth, S., 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. Journal of Memory and Language 111, 104063. doi:10.1016/j.jml.2019.104063.

Just, M.A., Carpenter, P.A., Woolley, J.D., 1982. Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General 111, 228–238. doi:10.1037/0096-3445.111.2.228.

Kaan, E., Harris, A., Gibson, E., Holcomb, P., 2000. The P600 as an index of syntactic integration difficulty. Language and Cognitive Processes 15, 159–201. doi:10.1080/016909600386084.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795. doi:10.1080/01621459.1995.10476572.

Kush, D., Johns, C.L., Van Dyke, J.A., 2015a. Identifying the role of phonology in sentence-level reading. Journal of Memory and Language 79, 18–29. doi:10.1016/j.jml.2014.11.001.

Kush, D., Lidz, J., Phillips, C., 2015b. Relation-sensitive retrieval: Evidence from bound variable pronouns. Journal of Memory and Language 82, 18–40. doi:10.1016/j.jml.2015.02.003.

Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences 4, 463–470. doi:10.1016/s1364-6613(00)01560-6.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology 62, 621 – 647. doi:10.1146/annurev.psych.093008.131123.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. Science 207, 203 – 205. doi:10.1126/science.7350657.

Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161 – 163. doi:10.1038/307161a0.

Kutas, M., Iragui, V., 1998. The N400 in a semantic categorization task across 6 decades. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 108, 456 – 471. doi:10.1016/S0168-5597(98)00023-9.

Lago, S., Acuña Fariña, C., Meseguer, E., 2021. The Reading Signatures of Agreement Attraction. Open Mind 5, 132–153. doi:10.1162/opmi_a_00047.

Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de) constructing the N400. Nature Reviews Neuroscience 9, 920–933. doi:10.1038/nrn2532.

Lee, E.K., Garnsey, S.M., 2015. An ERP study of plural attraction in attachment ambiguity resolution: Evidence for retrieval interference. Journal of Neurolinguistics , 1–16doi:10.1016/j.jneuroling.2015.04.004.

Lee, M.D., Wagenmakers, E.J., 2014. Bayesian cognitive modeling: A practical course. Cambridge University Press. doi:10.1017/CBO9781139087759.

Lee, P.M., 2012. Bayesian statistics: An introduction. John Wiley & Sons.

Levy, R., Keller, F., 2013. Expectation and locality effects in German verb-final structures. Journal of Memory and Language 68, 199–222. doi:10.1016/j.jml.2012.02.005.

Lewis, R.L., Vasishth, S., 2005. An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science: A Multidisciplinary Journal 29, 375 – 419. doi:10.1207/s15516709cog0000_25.

Lewis, R.L., Vasishth, S., Van Dyke, J.A., 2006. Computational principles of working memory in sentence comprehension. Trends in Cognitive Science 10, 447 – 454. doi:10.1016/j.tics.2006.08.007.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., van het Nederend, M., Burchert, F., Stadie, N., Caplan, D., Vasishth, S., 2021. A computational evaluation of two models of retrieval processes in sentence processing in aphasia. Cognitive Science 45. doi:10.1111/cogs.12956.

Logačev, P., Vasishth, S., 2015. A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. Cognitive Science 40, 266–298. doi:10.1111/cogs.12228.

Logačev, P., Vasishth, S., 2016. Understanding underspecification: A comparison of two computational implementations. Quarterly Journal of Experimental Psychology 69, 996–1012. doi:10.1080/17470218.2015.1134602.

von der Malsburg, T., Vasishth, S., 2013. Scanpaths reveal syntactic underspecification and reanalysis strategies. Language and Cognitive Processes 28, 1545–1578. doi:10.1080/01690965.2012.728232.

Mantegna, F., Hintz, F., Ostarek, M., Alday, P.M., Huettig, F., 2019. Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. Neuropsychologia 134, 107199. doi:10.1016/j.neuropsychologia.2019.107199.

Martin, A.E., Nieuwland, M.S., Carreiras, M., 2014. Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. Brain and Language , 42 – 51doi:10.1016/j.bandl.2014.05.001.

Mathôt, S., Schreij, D., Theeuwes, J., 2012. Opensesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods 44, 314–324. doi:doi:10.3758/s13428-011-0168-7.

McClelland, J.L., St. John, M., Taraban, R., 1989. Sentence comprehension: A parallel distributed processing approach. Language and Cognitive Processes 4, 287– 335. doi:10.1080/01690968908406371.

McElree, B., 2000. Sentence comprehension is mediated by content-addressable memory structures. Journal of Psycholinguistic Research 29, 111 – 123. doi:10.1023/A:1005184709695.

Meng, M., Bader, M., 2000. Mode of disambiguation and garden-path strength: An investigation of subject-object ambiguities in German. Language and Speech 43, 43–74. doi:10.1177/00238309000430010201.

Mertzen, D., Laurinavichyute, A., Dillon, B.W., Engbert, R., Vasishth, S., 2024. Crosslinguistic evidence against interference from sentence-external distractors. Journal of Memory and Language 137. doi:10.1016/j.jml.2024.104514.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., Vasishth, S., 2023. Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data. Glossa Psycholinguistics 2. doi:10.5070/G60111266.

Miyamoto, E.T., 2002. Case markers as clause boundary inducers in Japanese. Journal of Psycholinguistic Research 31, 307–347. doi:10.1023/A:1019540324040.

Ness, T., Meltzer-Asscher, A., 2017. Working memory in the processing of long-distance dependencies: Interference and filler maintenance. Journal of Psycholinguistic Research 46, 1353–1365. doi:10.1007/s10936-017-9499-6.

Ness, T., Meltzer-Asscher, A., 2019. When is the verb a potential gap site? The influence of filler maintenance on the active search for a gap. Language,

Cognition and Neuroscience 34, 936–948. doi:10.1080/23273798.2019.1591471.

Nicenboim, B., Schad, D.J., Vasishth, S., 2023. Introduction to Bayesian data analysis for cognitive science. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series. URL: https://vasishth.github.io/bayescogsci/.

Nicenboim, B., Vasishth, S., 2018. Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. Journal of Memory and Language 99, 1–34. doi:10.1016/j.jml.2017.08.004.

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, 1075 – 1100. doi:10.1111/cogs.12589.

Nicenboim, B., Vasishth, S., Rösler, F., 2020. Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. Neuropsychologia , 107427doi:10.1016/j.neuropsychologia.2020.107427.

Nieuwland, M.S., Barr, D.J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D.I., Ferguson, H.J., Fu, X., Heyselaar, E., Huettig, F., et al., 2019. Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials.

Philosophical Transactions of the Royal Society B 375, 20180522. doi:10.1098/rstb.2018.0522.

Nieuwland, M.S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D.J., Rousselet, G.A., Ferguson, H.J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E.M., Donaldson, D.I., Kohút, Z., Rueschemeyer, S.A., Huettig, F., 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. eLife 7, e33468. doi:10.7554/eLife.33468.

Oberauer, K., Kliegl, R., 2006. A formal model of capacity limits in working memory. Journal of Memory and Language 55, 601 – 626. doi:10.1016/j.jml.2006.08.009. special Issue on Memory Models.

Osterhout, L., Holcomb, P.J., 1992. Event-related brain potentials elicited by syntactic anomaly. Journal of Memory and Language 31, 785–806. doi:10.1016/0749-596X(92)90039-Z.

Paape, D., Smith, G., Vasishth, S., 2025. Do local coherence effects exist in English reduced relative clauses? Journal of Memory and Language 140, 104578. doi:10.1016/j.jml.2024.104578.

Paape, D., Vasishth, S., 2022. Does conscious rereading lead to targeted regressions in garden-path sentences? Data from a novel stop-and-reread paradigm. PsyAxiv doi:10.31234/osf.io/d7pvz.

Palestro, J.J., Sederberg, P.B., Osth, A.F., Van Zandt, T., Turner, B.M.,

2018. Likelihood-free methods for cognitive science. Springer. doi:10.1007/978-3-319-72425-6.

Parker, D., An, A., 2018. Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. Frontiers in Psychology 9, 1566. doi:10.3389/fpsyg.2018.01566.

Prasad, G., Linzen, T., 2021. Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. Journal of Experimental Psychology: Learning, Memory, and Cognition 47, 1156. doi:10.1037/xlm0001046.

R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rabovsky, M., Hansen, S.S., McClelland, J.L., 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nature Human Behaviour 2, 693–705. doi:10.1038/s41562-018-0406-4.

Rayner, K., Kambe, G., Duffy, S.A., 2000. The effect of clause wrap-up on eye movements during reading. The Quarterly Journal of Experimental Psychology: Section A 53, 1061–1080. doi:10.1080/713755934.

Royall, R., 1997. Statistical Evidence: A likelihood paradigm. Chapman and Hall, CRC Press, New York. doi:10.1201/9780203738665.

Schad, D.J., Nicenboim, B., Bürkner, P.C., Betancourt, M., Vasishth, S., 2022. Workflow techniques for the robust use of Bayes factors. Psychological Methods doi:10.1037/met0000472.

Schoknecht, P., Roehm, D., Schlesewsky, M., Bornkessel-Schlesewsky, I., 2022. The interaction of predictive processing and similarity-based retrieval interference: an ERP study. Language, Cognition and Neuroscience , 1–19doi:10.1080/23273798.2022.2026421.

Sisson, S.A., Fan, Y., Beaumont, M., 2018. Handbook of approximate Bayesian computation. CRC Press.

Smith, G., Vasishth, S., 2020. A principled approach to feature selection in models of sentence processing. Cognitive Science 44. doi:10.1111/cogs.12918.

Sturt, P., 2003. The time-course of the application of binding constraints in reference resolution. Journal of Memory and Language 48, 542 – 562. doi:10.1016/S0749-596X(02)00536-3.

Swets, B., Desmet, T., Clifton, C., Ferreira, F., 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. Memory and Cognition 36, 201–216. doi:10.3758/MC.36.1.201.

Tabor, W., Galantucci, B., Richardson, D., 2004. Effects of merely local syntactic coherence on sentence processing. Journal of Memory and Language 50, 355–370. doi:10.1016/j.jml.2004.01.001.

Tanner, D., Grey, S., van Hell, J.G., 2017. Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. Psychophysiology 54, 248–259. doi:10.1111/psyp.12788.

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, 285–318. doi:10.1006/jmla.1994.1014.

Tucker, M.A., Idrissi, A., Almeida, D., 2015. Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. Frontiers in Psychology 6, 347. doi:10.3389/fpsyg.2015.00347.

Van Dyke, J.A., 2007. Interference effects from grammatically unavailable constituents during sentence processing. Journal of Experimental Psychology: Learning, Memory & Cognition 33, 407 – 430. doi:10.1037/0278-7393.33.2.407.

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. doi:10.1016/S0749-596X(03)00081-0.

Van Dyke, J.A., McElree, B., 2006. Retrieval interference in sentence comprehension. Journal of Memory and Language 55, 157 – 166. doi:10.1016/j.jml.2006.03.007.

Van Dyke, J.A., McElree, B., 2011. Cue-dependent interference in comprehension. Journal of Memory and Language 65, 247 – 263. doi:10.1016/j.jml.2011.05.002.

Vasishth, S., Chen, Z., Li, Q., Guo, G., 2013. Processing Chinese relative clauses: Evidence for the subject-relative advantage. PLoS ONE 8, 1–14. doi:10.1371/journal.pone.0077006.

Vasishth, S., Drenhaus, H., 2011. Locality in German. Dialogue & Discourse 2, 59–82. doi:10.5087/dad.2011.104.

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. doi:10.1016/j.jml.2018.07.004.

Vasishth, S., Suckow, K., Lewis, R.L., Kern, S., 2011. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures. Language and Cognitive Processes 25, 533–567. doi:10.1080/01690960903310587.

Verdinelli, I., Wasserman, L., 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. Journal of the American Statistical Association 90, 614–618. doi:10.2307/2291073.

Vuorre, M., 2017. Bayes Factors with brms. URL: https://vuorre.com/posts/2017-03-21-bayes-factors-with-brms.

Wagenmakers, E.J., Brown, S., 2007. On the linear relation between the mean and the standard deviation of a response time distribution. Psychological Review 114, 830. doi:10.1037/0033-295X.114.3.830.

Wagenmakers, E.J., Lodewyckx, T., Kuriyal, H., Grasman, R., 2010. Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. Cognitive Psychology 60, 158–189. doi:10.1016/j.cogpsych.2009.12.001.

Wagers, M.W., Lau, E.F., Phillips, C., 2009a. Agreement attraction in comprehension: Representations and processes. Journal of Memory and Language 61, 206–237. doi:10.1016/j.jml.2009.04.002.

Wagers, M.W., Lau, E.F., Phillips, C., 2009b. Agreement attraction in comprehension: Representations and processes. Journal of Memory and Language 61, 206–237. doi:10.1016/j.jml.2009.04.002.

Yadav, H., Paape, D., Smith, G., Dillon, B.W., Vasishth, S., 2022. Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation. Open Mind , 1–24doi:10.1162/opmi_a_00052.

Yadav, H., Smith, G., Reich, S., Vasishth, S., 2023. Number feature distortion modulates cue-based retrieval in reading. Journal of Memory and Language 129. doi:10.1016/j.jml.2022.104400.

Zehr, J., Schwarz, F., 2018. PennController for Internet Based Experiments (IBEX). doi:10.17605/OSF.IO/MD832.