November 26, 2024

Prof. Christopher Jarrold
Associate Editor
Journal of Memory and Language

Response letter for Submission JML-24-17, 2nd round of revisions

Dear Prof. Jarrold,

with my co-authors Shravan Vasishth and Himanshu Yadav, I would like to submit the second revision of our article titled "Do syntactic and semantic similarity lead to interference effects? Evidence from self-paced reading and event-related potentials using German".

We are grateful for the thoughtful and constructive comments and suggestions from yourself and the reviewers, and feel that the quality of the manuscript has been greatly improved by the revisions. To address the concern that we overstate the implications of our findings, we have clarified at various places in the abstract and discussion of our results that our findings only apply to the present design.

Clarifying text has been added in this second round of revisions (e.g., regarding the implications, the prolonged reading times starting at the distractor and the use of syntactic and semantic cues), and we have shortened the manuscript length by four pages. We have removed redundant text passages (e.g., in the description of the SPR results).

The OSF repository has been made public to ensure general accessibility. We have updated the link in the manuscript accordingly.

Below, we separately address each comment from the reviewers and yourself. All changes have been highlighted in the revised manuscript. The actionable parts of the comments are highlighted in bold.

Sincerely,

Pia Schoknecht
Postdoctoral researcher
Department of Linguistics
University of Potsdam, Germany

Editor's comments

Comment E.1

> "Dear Dr. Schoknecht,
>
> Thank you for submitting your revised manuscript to the Journal of Memory and Language. This has now been reviewed by the same three experts who reviewed the previous submission, and you can find their comments below.
>
> You will see from these that all three reviewers appreciate the thoroughness with which you responded to the issues raised on the initial submission. Reviewers 1 and 3 raise some remaining points but recommend acceptance, while Reviewer 2 is less persuaded that your revisions have adequately addressed their concerns. Given this set of reviews, I have come to a 'revise' decision because I would like you to have one more go at addressing the remaining points made by all three reviewers.
>
> I do not anticipate sending out the next version for review again, but rather anticipate being in a position to accept it at that point. However, it's important to note that that is not guaranteed, and one reason I have come to a 'revise' rather than 'accept subject to minor revisions' is that, like Reviewer 1, I was unable to access your data via your OSF link. **It is very important to us at JML that data are readily accessible, and that accessibility needs to be in place before any potential acceptance decision**."

Response to E.1

We have made the OSF repository public to grant general accessibility. We paste our data availability statement below.

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

Comment E.2

> "The other key points I would like to ask you to pay particular attention to in any response to the current set of reviews are:
>
> **Reviewer 2 and 3's clear view that you still overstate the implications of your findings in places**."

Response to E.2

We have revised the manuscript to be more cautious regarding the implications and generalizability of our findings. The most that we can claim in this paper is that *in the particular experiment design* that we use here, we have no decisive evidence for syntactic interference. But this does not mean that syntactic interference cannot occur in general. We make this point clear in our revision now.

See the revised abstract and related text edits below.

Revised 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.

In the general discussion (page 73), we already had an unambiguous statement that, given the broader evidence in sentence processing, syntax generally does play a role in dependency completion:

> Thus, it seems that, cross-linguistically, syntax does generally play a central role in building incremental structure and in completing dependencies.

We have now made this section even more explicit by stating that syntactic interference may well play a role in some other, stronger syntactic manipulation than the one that we used, and that our own results would need to be independently replicated in order to establish their robustness:

> 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.

Related text edits on pages 62, 70, 72 and 74:

> Nevertheless, in the present design, both the SPR and the EEG experiment showed predominantly evidence against syntactic interference.

> 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.

> 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
>
> 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.

Comment E.3

> "**Reviewer 2's additional point 1 that asks whether you have changed that particular analysis**."

Response to E.3

Yes, we changed all analyses in the first revision. As we had explained in the response to reviewers on August 29, 2024 (specifically in the response letter and our responses to the comments E.2 and R3.5), the priors were changed from directional priors in the original version to symmetrical priors in the revised version. This was necessary because we added trial id as a predictor as was suggested by a reviewer. There was no a priori hypothesis of the sign of the effect of trial id (do the participants get faster or slower over the course of the experiment?), therefore the prior for trial id needed to be symmetrical (positive and negative values). It is not possible to include symmetrical and directional priors within a single brms model (Bürkner, 2021), thus all priors were changed to symmetrical ones.

Additionally, we changed the method with which the Bayes factors were computed. Initially, we had used bridge sampling, but in the revision, we switched to the Savage-Dickey method because only using that method we were able to compare models with random slopes (this was also explained in the previous response letter, Response to R1.7). These changes did not change the results qualitatively, but led to some numerical changes which probably piqued the reviewer's interest.

Comment E.4

> "and **their final point about the length of the current version of the manuscript**."

Response to E.4

We have removed repetitions in the description of the statistical analyses and have shortened the description of the self-paced reading results (and the comparison to the previous studies). This reduced the length of the manuscript by four ages; after the first round of revisions the main text ended on page 77, now it ends on page 73.

Reviewer #1

Comment R1.1

> "I was the Reviewer #1 of the previous version of the manuscript.

All my concerns are addressed in this revision, so I am happy to say I don't see reasons to ask for resubmission or for another round of major, significant revisions. I found it a solid paper in the first round already, but with the revisions and especially given the modified and extended analysis, I find the whole story more convincing and also, fair.

I do have some suggestions. In my view these are only minor. They mainly concern just the way things are presented or discussed. I list my comments by page.

p. 16: osf data etc.

I wanted to check those, but **when I clicked the link, it said I don't have permissions and that I need to request access**."

Response to R1.1

We have made the OSF repository public to grant general accessibility (see also our Response to E.1).

Comment R1.2

"p. 17, SPR experiment

Did the analysis include any cut-off point for reading times to exclude data? I could not find this in the manuscript. If it did, could this information be included? If not, I would strongly suggest to use this. In my own experience with Prolific, I noticed the following very common pattern: people basically do not read around 10-20% of experimental items - that is, for quite a significant subset of participants I noticed that they have super fast reading times on some subset of items, most likely because they just rushed through those items by just pressing the space bar and keeping it pressed. This is sometimes followed by a long waiting time at some other point, so those participants would not be outliers when we check the total time in the experiment. These people and their data would not be excluded by checking their answers to comprehension questions, especially if only a small subset of items has comprehension questions, like one third, as in this experiment (basically, they could still easily pass with around 90% success rate). So if this was not checked, it would be good to check and **consider removing RTs based on small/large cut-off points and to see whether that affects the analysis. If this was done already, it would be good to report more details, in particular, the cut-off points and also the amount of data that was removed this way**."

Response to R1.2

We excluded reading times below 150 ms and above 3000 ms. This affected 4.9 % of the data. We have added this information to the manuscript on page 19:

> 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.

Comment R1.3

> "p. 36: "Regarding the possibility that sentence structure confounds caused the effects in the pre-critical region(.)"
>
> **I wonder whether it would make sense to briefly summarize what the sentence structure confound should be.** Otherwise the whole paragraph is hard to follow for readers who did not read or do not remember Mertzen et al. (2023)."

Response to R1.3

We have added the following text to the manuscript on page 35 (above the sentence which was quoted by the reviewer):

> 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.

Comment R1.4

> "p. 51: "For more discussion of the role of structural retrieval cues, see Franck and Wagers (2020) and Arnett and Wagers (2017)."
>
> **It would be good to cite Kush et al. (2015) here** (after all, a significant portion of Franck and Wagers, 2020, builds on Kush et al. 2015 to explore ways to deal with structural cues).
>
> Kush et al. (2015) Relation-sensitive retrieval: Evidence from bound variable pronouns. Journal of memory and language."

Response to R1.4

We have added the suggested reference on page 52:

> For more discussion of the role of structural retrieval cues, see Kush et al. (2015b), Franck and Wagers (2020) and Arnett and Wagers (2017).

Comment R1.5

> "p. 55: "$\Psi(d(x_i, y), 0, \delta)$, where $\Psi(.|\delta)$"
>
> I was a bit confused here, since the formula "$\Psi(d(x_i, y), 0, \delta)$ " did not have any "$\Psi(.|\delta)$" part."

Response to R1.5

We have changed the expression to make it clearer. On page 55, we now say:

> 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.

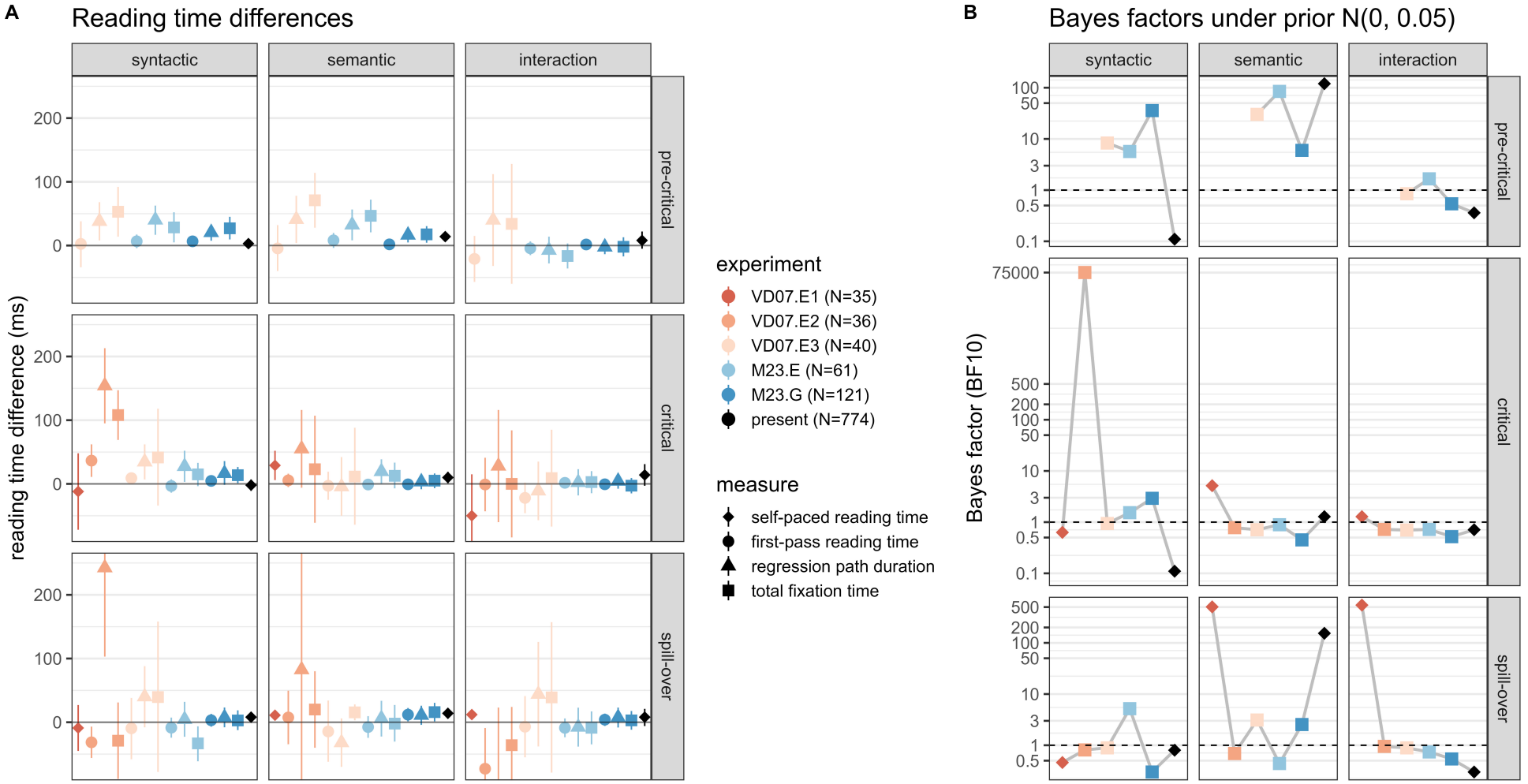
Comment R1.6

> "p. 61, **Figure 14: one value is outside the scale of the bottom left graph. If possible, it would be good to make it visible**."

Response to R1.6

We have changed the y-axis limits to include the mean estimate of the respective data point (VD07.E2, regression path duration). For convenience, we paste the revised Figure here.

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).



Comment R1.7

> "p. 64: "Word-by-word presentation is likely to be more demanding on the comprehender's memory, which should make it easier to detect interference effects. So, the differences in methodology do not offer a straightforward explanation for the lack of syntactic interference in our data compared to the previous studies."
>
> Actually, I thought here the authors basically provided an explanation of the differences between their findings and the previous findings because of differences in the methodology, even though their last sentence said otherwise. Let me spell out it: since the current methodology (SPR, EEG) is likely more demanding on memory, it is possible that people give up on detailed processing; rather, they do something akin to good-enough processing and do not fully process; consequently, they could still get interference from semantics (since the semantic interference is really simple, it is just lexical semantics and readers don't need to fully parse for that) but they don't correctly parse embedded subjects, hence the syntactic interference is gone. **The authors in fact come to good-enough processing in Section 7.3, but then it is not linked back to differences between their findings and the other findings (which I think should be, I do think this is a possible explanation).**"

Response to R1.7

We have added the reviewer's suggested explanation that the employed methods could have especially led to good-enough processing and that this might have contributed to the different findings between studies (see page 61):

> 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.

Comment R1.8

> "p. 76: **Chromy should appear like this: "Chromý" (note the correct diacritics above the "y")**"

Response to R1.8

We have corrected the spelling of Jan Chromý's last name.

Reviewer #2

Comment R2.1

> "I appreciate the authors' responses to my previous comments. However, despite the revisions, significant weaknesses remain, rendering in my opinion this study primarily a methodological contribution that adds little value to

our understanding of interference effects. Additionally, the authors' strong claims often come through as overstatements that could result in serious misconceptions. My overall assessment is as follows.

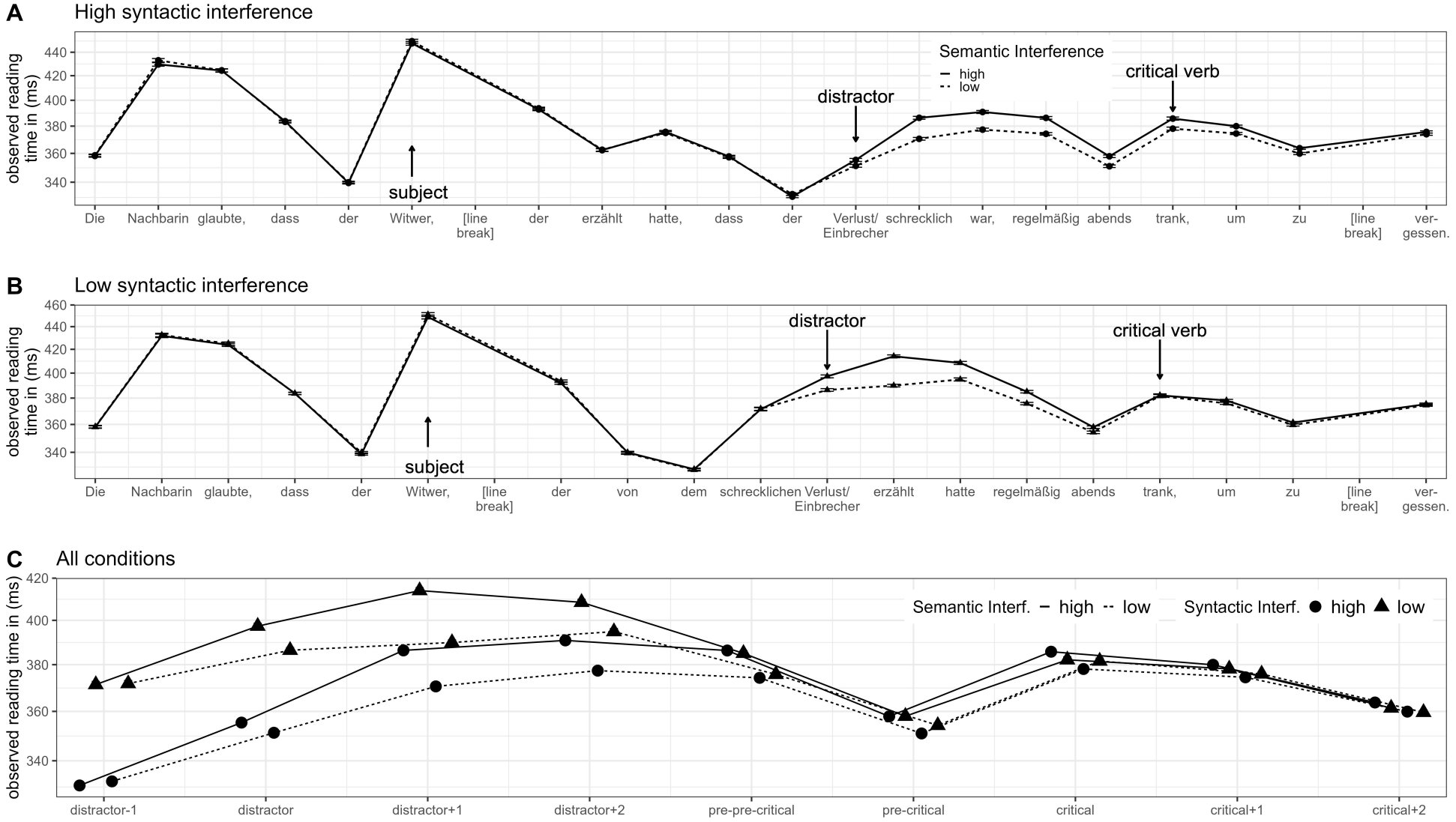
1. Interpretability of the SPR results: The results from the self-paced reading (SPR) task remain unclear. The only observable effect—the semantic interference—emerges well before the critical region of interest (specifically, at the distractor region) and persists without change beyond the critical region. As a result, any effect at the critical region is uninterpretable.

The authors acknowledge the interpretability issue with the SPR results in their response to my comment, and they now include the following statement: "Given the pre-critical reading time differences, effects in the later regions, i.e., the critical and spill-over regions, cannot be attributed clearly to the stimuli in these regions and sentence processing mechanisms associated with them. Therefore, we refrain from further discussing effects occurring in the later regions." **However, this comes after more than 10 full pages (pp. 25-37) of discussion on the reaction times (RTs) in the critical and pre-critical regions, which would lead any reader to believe these findings are meaningful. If, as the authors now realize, the effect emerges much earlier, this entire discussion is misleading: the SPR results after the distractor region are simply uninterpretable.**"

Response to R2.1

We agree with the reviewer that the presentation and discussion of the SPR results was misleading and have revised it accordingly. We have revised Figure 3 to remove the focus on the critical region (see below). At the beginning of the Results section, we now anticipate our interpretation of the results and state that we believe that the distractors led to encoding interference. We also state that the results at the critical (and spill-over) region are confounded because it is not possible to disentangle the long-lasting encoding effect from retrieval effects potentially arising in the critical region.

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.



The reviewer seems to believe that our results are not publication-worthy because encoding interference confounded the reading times of all regions after the distractor. However, as we had already discussed in the results of the SPR results, the results of the two published studies using the present design (Van Dyke, 2007; Mertzen et al., 2023) show indications that they were affected by encoding interference as well (see page 35):

> 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.

Therefore, our study replicates the encoding interference effect of the previous studies but it is the only study to prominently address that the reading times of the critical region are confounded by encoding interference. To ensure comparability to the previous studies, we believe that it is mandatory that we report the reading times effects in the same regions which have been reported by them: the pre-critical, critical and spill-over region – despite the confound discussed above. We have revised the Results section to be more concise, starting on page 26:

> 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. In our materials, there were two words between the distractor and the critical verb which were identical across conditions: the pre-pre-critical and pre-critical adverbs. The pre-pre-critical region was introduced in the present study to absorb potential effects caused by processing the clause boundary of the embedded clause, rendering it a problematic region for analyses. The pre-critical region was not adjacent to clause boundaries. The reading times of the pre-critical region were analyzed to investigate the reading times difference starting at the distractor. This difference starting at the distractor and persisting almost until the end of the sentence was problematic for the analysis and interpretation of reading times in the critical (and spill-over) 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 also report analyses of the reading times of the 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.

Because the finding that the self-paced reading times were primarily affected by semantic encoding interference is a central finding of our study, we think that it is appropriate that we investigated it. We did not analyze the distractor itself and the immediately following words because these were not identical across conditions. Instead, we used the reading times of the pre-critical region which was the same word across conditions (*abends*, 'in the evening', in the example item used for the x-axis labels in Figure 3 above) as a proxy to investigate the effect of encoding interference caused by the distractor. This pre-critical word preceded the critical verb, thus the reading times at the pre-critical word are unlikely to show retrieval interference. To emphasize this we have renamed the former section "Pre-critical effects" to "Reading times difference prior to retrieval". We have revised the beginning of the Discussion on page 30 accordingly:

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

Comment R2.2

"However, there are also problems with the interpretation of the effect at the distractor region. If the authors want to attribute this effect to encoding interference, then they must argue for the existence of long-lasting encoding effects, as this very same effect persists throughout the entire sentence. Yet, the mechanism behind such enduring encoding effects remains unclear, and the authors do not address this challenge. Feature overlap, a commonly proposed mechanism for encoding interference, seems unlikely. It would suggest that the two elements compete for the same feature throughout the entire

sentence, which is not plausible. Activation leveling, another potential mechanism, is also improbable, as it equalizes the activation of competing elements and should result in similar reaction times at some point, which we do not see here. **What is evident from these findings is that the inanimate distractors are read faster than the animate ones (but the opposite pattern is found in ERPs, see point 3), and this preference creates a reaction time difference that persists until and beyond the critical verb. If the authors wish to interpret this as evidence of encoding interference, they must provide compelling reasons to support the idea that encoding interference can have such a prolonged effect, along a clear mechanism for it**."

Response to R2.2

The focus of the present study was to investigate retrieval interference due to syntactic and semantic similarity. The finding that the reading times were primarily affected by encoding interference was unexpected (but see Mertzen et al.'s (2023) discussion). Given that the present manuscript contains two large-sample experiments and computational modeling (and its substantial length), we think that it is outside the scope of this manuscript to provide an in-depth discussion of a mechanism of encoding interference. Developing a complete treatment of encoding interference would require considerably more computational modeling, which would further increase the length and complexity of this paper. In our opinion, such a model should be developed, but that would need to be a stand-alone article, which would also require more experimental studies to validate and test competing models' predictions. For a recent attempt that we made to try to implement different versions of feature distortion, a phenomenon that is related to encoding interference, see:

Himanshu Yadav, Garrett Smith, Sebastian Reich, and Shravan Vasishth. Number feature distortion modulates cue-based retrieval in reading. Journal of Memory and Language, 129, 2023.

So, in summary, a proper treatment of encoding interference would require a modeling effort at the same scale as was carried out in the above-mentioned paper. Such a modeling effort, although very worthwhile, would take several years' of work and is therefore far beyond the scope of the present work.

Comment R2.3

"2. Lack of syntactic interference evidence: No evidence for syntactic interference was found in either task, contradicting two previous studies that tested similar structures (Van Dyke 2007; Mertzen et al. 2023). **In my previous review, I noted that the syntactic manipulation used was not appropriate and thus unlikely to induce interference effects. Although the authors concede that this might be true, they did not address the logical implications of this acknowledgment, as evidenced by their repeated strong assertions**, such as the one in the abstract: "Surprisingly, in both experiments, Bayes factor analyses showed evidence against interference due to syntactic cues". If the syntactic manipulation is irrelevant, then repeatedly concluding that syntactic interference is absent is not only an overstatement but also misleading. All this study allows us to conclude is that when cues that are irrelevant to syntax are manipulated (like [+being a subject of whichever clause at whichever level of embedding], underpowered studies (Van Dyke 2007; Mertzen et al. 2023) may falsely report interference

effects. While this is a valid conclusion and a sound methodological critique, it does not significantly advance our theoretical understanding of syntactic interference. Specifically, it fails to demonstrate that when relevant syntactic cues are manipulated, syntactic interference is not attested. Yet, this is the conclusion the authors consistently imply throughout their manuscript, leading to significant misconceptions. As I mentioned in my previous review, beyond the two studies the authors are building on, numerous studies in the literature show that hierarchically intervening elements do generate strong syntactic interference. It is a fine enterprise to correct inaccurate claims from previously underpowered studies that used irrelevant manipulations; however, this should not not suggest that the findings presented here have a broader significance than they do."

Response to R2.3

We agree with the reviewer that our present work does not rule out syntactic interference in general. We have tried to stress throughout the paper that the present study using the present design did not induce considerable syntactic interference. We have revised the abstract and manuscript main body accordingly (see response to E.2). In the general discussion, we have also added the following caveats:

> 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.

We feel that the above quote and the other re-statements in the paper (including the abstract) appropriately limit the claims we make in the paper. We never state anywhere in the paper that syntactic interference is generally absent in sentence processing.

Comment R2.4

> "3. Misalignment of SPR and ERP results: The results from the self-paced reading (SPR) and event-related potential (ERP) methods do not align, with the only observed effect—semantic interference—pointing in opposite directions: inhibitory in SPR but facilitatory in ERP. Specifically, the authors report a reversed effect in the P600 component, showing a reduced P600 in conditions of higher semantic interference. This contradicts typical predictions, where the P600 is usually greater under high interference. The authors interpret this as a facilitatory interference effect (i.e., easier processing with high semantic interference). However, if this were the case, **we would also expect faster reaction times (RTs) in high semantic interference conditions, while the opposite is observed in the SPR data**."

Response to R2.4

The SPR results and the ERP results in the N400 window both showed inhibitory interference, i.e., more processing difficulty under high vs. low semantic interference, manifesting in longer reading times and a more negative N400. We explain below in our response to R2.5 that the reduced P600 for high interference might be explained by facilitatory interference because under high semantic interference whatever noun was retrieved can be easily integrated.

In general, SPR results cannot show the same complex results pattern that the ERP results showed because they provide just one measure per region. Additionally, only at the critical (or spill-over) region facilitated integration might be observed and in line with the reviewer's other comments, we do not interpret the reading times of the these regions because they were overshadowed by the reading times differences starting at the distractor.

Comment R2.5

> "4. Conflicting directions in ERP components: Within the ERP results, the different components also point in opposite directions. The P600 component shows a reversed effect, with a reduced P600 in conditions of high semantic interference. However, the N400 component follows the expected pattern, with stronger activation under high semantic interference. **It remains unclear how the directionality of these two can be reconciled within the same theoretical framework.**"

Response to R2.5

We have revised our discussion of the ERP results on page 48 to clarify how the seemingly conflicting results can be reconciled (see below). Interference might enhance the N400 amplitude because it makes memory retrieval more difficult and it might reduce the P600 because no matter which noun was retrieved it can be easily integrated with the verb.

> 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.

Comment R2.6

"5. Unclear conflicting mechanisms: The authors attribute the SPR results to encoding interference and the ERP results to retrieval interference, leading to a contradictory conclusion. I won't delve further into this point, as the issues raised earlier already highlight more fundamental problems."

Response to R2.6

This comment does not call for further action. In the General Discussion (subsection "Encoding and retrieval interference"), we discuss that we believe that the difference in presentation rate and whether it was controlled by the participants or not led to different types of interference (encoding vs. retrieval).

Comment R2.7

"Additional points:

1. Original version: "By contrast, the syntactic interference effect had only anecdotal evidence under the two narrow priors (Normal-(0, 0.1): BF10 = 2.5, Normal-(0, 0.5): BF10 = 1.2). There was evidence against syntactic interference under wider priors (Normal-(0, 1): BF10 = 0.66, Normal-(0, 5): BF10 = 0.14). What this means is that, in our data, only if we assume a priori that the effect size is relatively small, there is very weak evidence in favor of syntactic interference."

Revised version: "By contrast, Bayes factors provided either no evidence for or even evidence against syntactic interference in both spatiotemporal windows (BF10 $<$ 1.2). Similarly, Bayes factors provided either no evidence for or even evidence against the interaction in both spatio-temporal windows (BF10 $<$ 1.2)."

Did the authors change the analyses? Why what was reported as anecdotal evidence under narrow priors is now reported as no evidence or evidence against?"

Response to R2.7

Yes, all analyses were changed in the first round of revision. Firstly, the priors were changed (directional priors in the original version and symmetrical priors in the revised version). Secondly, the method to compute Bayes factors was changed (from bridge sampling to Savage-Dickey, see also our Response to E.3). This resulted in small numerical

changes of the results, e.g., BF10 = 2.5 in the original analysis decreased to BF10 = 1.2 in the revised analysis. Furthermore in the original version of the manuscript, we were inconsistent in our description of small BFs, i.e., BF10 = 1.2. Since the first revision and in line with Lee and Wagenmakers (2014), we now consistently interpret such small BFs as providing no evidence either way.

Comment R2.8

> "2. As a side note, on p. 58, the authors conclude that "the parser searches for a subject that is within the same clause but uses the animacy cue without reference to the clause in which a noun appears." Does this mechanism truly seem plausible to the authors? What would this imply in practice? **Does it suggest that semantic and syntactic information are entirely encapsulated from each other?** If we were to set aside the issues I raised earlier, this would be one of the paper's central conclusions, yet the authors fail to provide a credible mechanism to support it."

Response to R2.8

To clarify the assumed independence of syntactic and semantic processing, we have revised the former section 'Relation to other sentence processing accounts', starting on page 70. We have also renamed the section to '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 determine 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.

Comment R2.9

"3. How do the fillers differ from the experimental items? In the example the authors reported there is retrieval at place again:

Experimental: The neighbor believed that the widower, who had told her that the loss was awful, regularly drank in the evenings to forget.
Filler: The carpet maker, who came to the workshop early, repaired the especially beautiful old carpet while listening to the news.

The authors state, "The fillers were less syntactically complex than the experimental items but had at least one embedded clause and generally provided more variety." **In what way did they provide more variety? In which respects? Both fillers and test items involve retrieval; what is the role of the fillers in this context?**"

Response to R2.9

The fillers included different types of subordinate clauses and (more) modifiers at different positions within the sentence. By deviating from the rigid structure of the critical items, they provided more variety in regard to the employed sentence structures. To avoid creating an obvious juxtaposition of very complex sentences (critical items) and very simple sentences (fillers), the fillers included long-distance dependencies and at least one subordinate clause. We have revised the description of the fillers in the manuscript on page 19 in the following way:

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.').

Comment R2.10

"4. There are numerous repetitions throughout the paper. A prime example is the paragraph on "hyp testing using BF" (p. 45), which is repeated almost verbatim in the "Discussion" section (p. 47). This redundancy occurs several times, making the paper excessively and unnecessarily long. **The authors should streamline the content to enhance clarity and conciseness.**"

Response to R2.10

We have removed the repetitions to improve clarity and conciseness (see also Response to E.4).

Reviewer #3

"This is a revision of a paper that I previously reviewed. (I was Reviewer #3 in the previous round.)

In the previous round of reviews, I was generally happy with the authors' findings. My queries were mostly related to the interpretation and generalizability of those findings. This is a routine issue in psycholinguistic research. A study looks at effects in one very specific (and often complicated) sentence type, and then draws conclusions about very broad notions such as 'syntax' and 'semantics'.

The authors have offered a thorough (daunting?) response to the reviews. I am more satisfied with some responses than others. But I do not think that this should stand in the way of publication in the JML special issue.

The authors acknowledge the concern about using relational notions such as 'subject of the same clause' as a memory retrieval cue. They provide some text that clarifies that they need to implement a clause tracking mechanism. I would prefer it if the text was clearer about the unfeasibility of a 'subject of the same clause' feature. But this does not impact their findings.

Another concern that I raised involves the scalability of the notion that semantic interference effects reflect the use of semantic retrieval cues such as [+animate]. At issue is how this extends to finer-grained properties that are routinely responsible for plausibility violations. In my previous review I used the example of "the teacher drove" vs. "the little boy drove". It is, of course, clear that teachers are more plausible drivers than young children. This is

part of our world knowledge. But it is less plausible that every time the noun 'teacher' is encountered the property [+can drive] is activated. The authors' response seems to be that animacy was sufficient for the materials in the current study, and that maybe other finer-grained features "become activated only when relevant". To my mind, this point removes much of the value of the cue-based retrieval theory. **Nevertheless, this concern is about the generalizability of the authors' claims, and not about the soundness of their results.**

Finally, and in a similar vein, I raised questions about whether there might be other reasons for the selective interference effects found in this study. Maybe a different mechanism could capture the special status of the subject-of-the-same-clause, or maybe some specific property of the current experimental materials could be responsible for the lack of interference from other subjects. The authors seem skeptical of the first of these, and they seem more sympathetic to the second. **This is all reasonable enough, and their text conveys some caution. The abstract, on the other hand, is rather more confident.** I am sympathetic to the authors' preferred conclusion. But I am less confident than they are about how well justified it is based on the current findings. This has nothing to do with the quantitative wizardry that the authors display in the paper. It's all about the issue of generalizing from a single sentence configuration (that participants saw in 60% of trials in the study)."

Response to Reviewer #3

We have revised the manuscript (including the abstract) to show more caution regarding the generalizability of our findings (see our response to E.2).