Department of Linguistics
University of Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam, Germany

Prof Genevieve McArthur
Associate Editor
PeerJ

Dear Prof McArthur,

Thank you for your quick review even in view of the Australian summer break and for the extremely helpful and insightful comments. Please accept our sincerest apologies for the long delay in returning our revisions.

We felt that all three reviewers' suggestions served to improve the clarity and preciseness of the manuscript and have therefore incorporated all of their suggestions. The major comments related mainly to the Introduction section, finding that this section lacked sufficient detail on the two accounts under investigation. The majority of our changes have therefore been to the Introduction, which we hope you find improved.

In the response letter below, we present your and each reviewer's comments in bolded text, and provide our response under each comment. Note that in some cases, we have split the comments into smaller chunks and/or paraphrased the comment for brevity. Where line numbers are given in the bolded reviewer comments, these refer to the original manuscript. New page numbers referring to the revised manuscript are provided in our responses.

Thank you again for your time and valuable feedback in reviewing the manuscript and we hope that we have sufficiently addressed the concerns raised.

Yours sincerely,

Kate Stone
Dr. Titus von der Malsburg
Prof. Shravan Vasishth

Responses to comments from the editor

1. **Predictions section (page 3). I found this section confusing. It seemed to "come out of the blue" - partly due to unclear wording, I believe, and partly because not much background had been provided about the two models that were being pitched against each other. I believe Reviewer 2 had a similar concern, and has offered some specific suggestions for how this might be addressed in the Introduction. In addition to those suggestions, please ensure that the Prediction section is precluded by a clear explanation of the two theories, and that the logic behind each prediction is described as clearly and simply as possible.**

   In order to address your concerns, we have used the suggestions of Reviewer 2 to restructure the Introduction. We hope you find that the revised Introduction now sets up the Predictions section more clearly. Specific examples of how the Introduction was revised are provided below under Reviewer 2's comments. Where possible, we provide specific page numbers where we have added requested information. However, the document containing tracked changes may, in some cases, be more useful where large sections of the Introduction have been reorganised.

2. **Participants section: Please clarify if "language" disorders include reading disorders.**

   We excluded participants with any kind of reading or production disorder and have rephrased this in the Participants sections on pages 5 and 12 to state "Participants were screened for acquired or developmental reading or language production disorders".

3. **Materials section: I was a bit confused by the presentation of the stimuli. Would it be possible to reformat the examples to improve clarity by adding a blank line between the two lines of the German/English stimuli, and also provide the meaning of the text prior to the stimuli?**

   Beginning on page 5, we have simplified this even further to condense each example into two sentences: one German and one translation. We have removed the English gloss altogether since we do not need

to annotate morpho-syntactic components. We provide one condition here as an example:

Small set/short distance:

Mit dem neu gekauften Lappen **schrubbte** sie die Teller in der Küche **ab**, um Platz zum Kochen zu schaffen.
*With the newly bought rag, she **scrubbed** the plates in the kitchen **off** to create space for cooking.*

4. **I understand why you might decide to outline the history of the development of the stimuli under Materials. However, the length of this history narrative the reader from the flow of information for Experiment.**

   We agree that this section disrupts the flow of the main text. We have revised this section to present only the cloze test results in the main text on page 6, since they are the directly relevant to the experiments presented. The frequency analysis has been moved to the appendices and the norming study removed entirely.

5. **At some point, there appeared to be an abrupt switch from the use of the term "predictability" to "entropy". If they are the same thing, it would help the reader to use the term "predictability" throughout the manuscript, since it is a less specialised word. However, if a switch to entropy is required, this needs to be explained clearly at the appropriate point in the narrative.**

   We appreciate that switching between terms without enough explanation has caused confusion. We do believe the switch is necessary, however, and have therefore revised the text to first state explicitly in the Introduction how predictability maps to set size (page 4), and that we also later be introducing the term entropy:

   > "Note that throughout the remainder of the article, we use *set size* as a proxy for predictability. Set size also relates to *entropy*, which we introduce in detail as it becomes relevant in the Cloze Test section."

   We then give a more detailed explanation of entropy where we first use it as a predictor in the cloze test analysis on page 7:

> "The *set size* manipulation was intended to induce uncertainty about the upcoming particle's lexical identity. One useful way of quantifying uncertainty is with *entropy*. Entropy provides a measure of how much information is carried by a new input in light of all possible outcomes. In our case, the new input is the particle. In a sentence context where many particles are plausible and cloze probability is uniformly low across all the plausible particles, we assume that uncertainty about the identity of the upcoming particle is high. Thus, each of the plausible particles carries a large amount of information about the meaning of the sentence and entropy is high. In a sentence where only few particles are plausible and one particle is much more probable than the others, we assume that uncertainty about that particle's identity and the meaning of the sentence is low, and so encountering the high-probability particle will be less informative; this is a low entropy situation. To calculate entropy for our experimental stimuli, we first calculated the cloze probability (P) of all verb particles given for each respective sentence in the cloze test. Entropy (H) of the target particle was then defined as:
>
> $$H = -\sum_i P_i log P_i$$

In the same section, we state explicitly how entropy maps to predictability:

> "We therefore present an analysis of entropy as a continuous predictor instead, since this maps better to our planned manipulation of predictability (high entropy = low predictability and vice versa)."

Throughout the remainder of the paper, we attempt to use both terms where possible, e.g. "high entropy (low predictability)".

Responses to comments from Reviewer 1

Major comments

1. **Does the verb in every stimuli sentence require a particle? If not, how was the "no particle" option incorporated in the**

cloze test and data analyses?

All items were confirmed to require a particle via cloze testing and native speaker judgements. Any items that did not require a particle (as evidenced by the Cloze test) were presented to participants as fillers, but not included in the final analysis. The text has been updated to state this more explicitly, both in the Introduction on page 4: "Using a cloze test, we confirmed that each sentence required a particle."; and under Materials on page 5: "Each experimental item was a quartet of four sentences in which the context required a particle for the sentence to be grammatical.".

2. **There is no mention of how spillover was taken into account, even though this phenomenon is prevalent in reading, in particular self-paced reading. Were reading times on words directly following the particle also taken considered? If not, could this be why the expected effects were not found?**

We did originally consider looking at the spillover region, but decided against is because the particle must be followed by a comma and it was not clear how the clause boundary and associated sentence wrap-up effects (Rayner et al., 2000) might interact with reading times in the spillover region. We therefore presented mean reading times across the sentence in Figure 1 (below). Figure 1 does not suggest that there was any difference in reading times in the spillover regions other than in the long-distance eye-tracking data where we already saw effects at the particle, and thus we did not analyse or discuss the spillover region further.

We have updated the text in both Data Analysis sections to include a statement about the spillover region:

> "We also considered analysing the spillover region, but decided against it as the particle had to be followed by a comma and it was not clear how the clause boundary and associated sentence wrap-up effects (Rayner et al., 2000) might interact with reading times in the spillover region. Instead, we present mean reading times across the sentence in Figure 1."

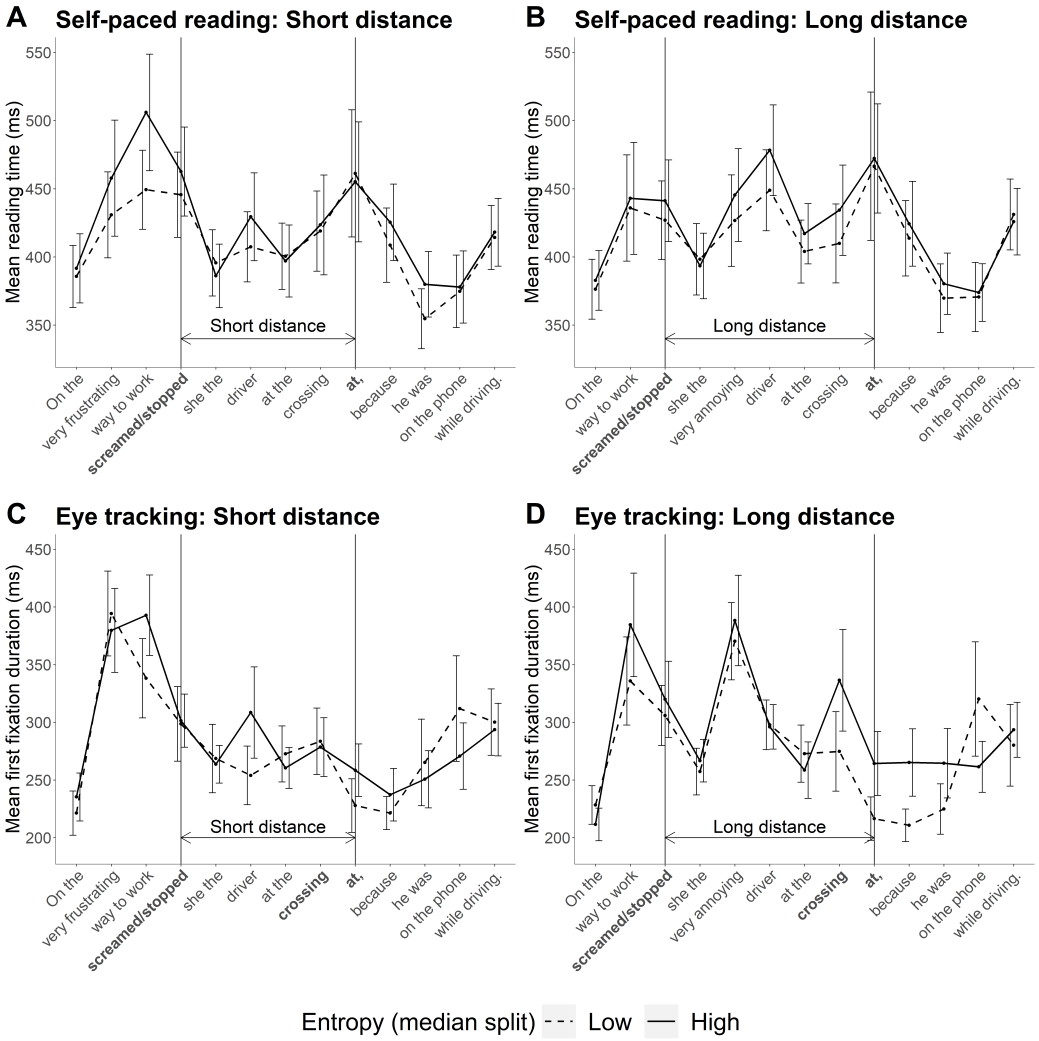

Figure 1: **Comparison of self-paced reading times and eye tracking total fixation times plotted across the sentence.** Error bars show 95% confidence intervals.

Minor comments

- **Line 124: what does it mean for something to be "anecdotally assumed"?**

  Here we were trying to express the fact that some papers assume acti-

vation decay has played a role in their findings on long-distance dependencies even though they don't specifically test it. We have updated the sentence on 3 to say this more explicitly:

> "There are few empirical experiments specifically testing decay in isolation, even though it is generally presumed to affect word processing times in long-distance dependencies (e.g. Xiang et al., 2014; Ness and Meltzer-Asscher, 2019; Chow and Zhou, 2019)."

- **When introducing German particle verbs, it would be good to mention that moving the particle to after the object NP is required in German.**

A sentence has been added to the Introduction to make this explicit on 3:

> "In German, however, the particle must appear after the direct object if the verb is transitive, usually at the right clause boundary (e.g. "Er raümte den Raum auf" *he tidied the room up*, but not "*Er raümte auf den Raum" *he tidied up the room*; Müller, 2002)."

- **Line 115: the Dutch prefix "ver" in "verdelen" is not a particle (i.e., it is not split: "hij deelt het ver" is not possible)**

This example has been replaced with an example from (Piai et al., 2013) on page 4:

> "In this study, it was hypothesised that Dutch verbs that can take a large number of possible particles (e.g. *spannen*, "to tense", which can take at least seven particles) would trigger preactivation of those particles, placing a larger demand on working memory than verbs with a small set size (e.g. *kleuren*, "to colour", which can take only two)."

- **Line 127-128: "self-paced reading and eye tracking modalities" and "reading modalities": shouldn't this be "paradigm" instead of "modality"? In both cases, the modality is written/visual.**

The word "modality" has been replaced with "experimental method" throughout the manuscript. We opted against "paradigm", only be-

cause we use this word elsewhere to describe the experimental presentation method.

- **Table 1 shows 95% CI instead of the standard error mentioned on line 221. Also, the caption is not quite accurate because the table presents cloze statistics but not the cloze test results.**

  The text on page 6 has been amended to state "Means and 95% confidence intervals of Beta distributions corresponding to the cloze probabilities for each factor level are presented in Table 1." and the caption of this table has been updated to state "cloze statistics" instead of "cloze test results".

- **It would be helpful if the goal of the cloze test data analysis were explained before the technical details.**

  We have revised the beginning of the cloze test section on page 6 to state the three purposes of the cloze test:

  > "In order to confirm that our sentence stimuli (i) elicited particles, (ii) that more particles were elicited by the large set condition than the small set condition, and to (iii) quantify the predictability of the target particle, a cloze test was conducted."

- **Cloze test analysis results: "the probability of the target particle was lower ... for the interaction": For which combination of factor levels was the probability lower?**

  The posterior for the interaction effect suggests there was no interaction of set size or distance that predicted cloze probability. Looking at the nested effects, there is a very small difference in cloze probability between large/small set verbs at long distance (lower probability for large set verbs), whereas there is no such difference at short distance. However, the difference in cloze probability between set sizes at long distance is about 0.025%, so this is not really meaningful. We have revised this paragraph on page 7 to remove any mention of the probability being lower and simply state that the posteriors are consistent with zero:

> "The model did not suggest that either set size, distance, or an interaction of the two influenced cloze probability. As can be seen in Figure 2, the posteriors for the probability of giving the target particle were more or less centred on zero, meaning that neither set size, distance, or their interaction made people any more or less likely to give the target particle."

- **The violin plots of Fig. 3 shows probability mass for negative values of entropy, even though entropy is by definition non-negative.**

  This was an error in the plot code and has been amended. The amended plot can be found on page 8.

- **Line 414: what did the preprocessing of eye-tracking data entail?**

  Apologies for having omitted this from the manuscript. The following detail has been added to the Analysis section of the eye-tracking section on page 15:

  > "Sampled data were exported from DataViewer (SR Research) and pre-processed in R using the *em2* package (Logačev and Vasishth, 2013). Trials containing blinks or track loss were excluded."

- **Line 417: the citation to R is "R Core Team", not just "Team".**

  This has been amended throughout the manuscript.

- **Line 448-449: the problem of evaluating multiple dependent measures is not a "limitation of the BF analysis" in particular, is it?**

  This is a good point, thank you. It is not a limitation of BFs, but rather of the way we have used them: applied to multiple dependent measures with no FWER correction. While the reference we cite relates to FWER in a frequentist framework, there may be an analogous issue in Bayesian analyses, but there is no current option for controlling or

correcting for this. We wanted to highlight this point and the fact that further confirmatory analysis is needed. We have rephrased this sentence on page 16 accordingly to state:

> "A possible limitation of our approach using Bayes factor analyses is that we are evaluating multiple measures, without any correction for family-wise error (**?**). While the family-wise error rate is a frequentist concept, it may be that an analogous issue exists in the Bayesian framework for which we have not controlled. Our analyses should therefore be considered exploratory and confirmed via future replication attempts.".

- **line 474: "The statistical analysis" should probably be "The outcome of the statistical analysis"**

This text no longer appears in the manuscript.

Responses to comments from Reviewer 2

Major comments

- **Clarity of opposing hypotheses, particularly the predictions of the LV05 model: If I understand correctly, the experiments set out to test the predictions of Surprisal vs. LV05. What is Surprisal theory, what are its main tenets? What is the LV05 model? What is it modelling, what are its assumptions?**

Our intent was actually to test surprisal versus decay rather than versus LV05 specifically, but we agree that the way the Introduction was formulated created confusion. We have therefore reframed the Introduction such that we compare the opposing predictions of predictability (as instantiated by surprisal) and temporal activation decay.

The Introduction first introduces predictability, beginning with this paragraph on page 2:

> "**Word predictability.** The surprisal theory of sentence processing provides an account of how words in a sentence become predictable and how predictability facilitates their processing (Levy, 2008; Hale, 2001). Surprisal is based on the assumption that the context of a sentence sets up expectations about what structural

information might appear next. Under surprisal, the difficulty of processing each new word in a sentence is equal to the negative log probability of that word appearing given the preceding context. The probability of a word given a context can be quantified using a probabilistic context-free grammar (PCFG; e.g. Levy, 2008). At each new word in a sentence, a set of plausible sentence continuations is generated based on the PCFG and held in parallel, ranked by their frequency. The degree of update that each new word induces in the distribution of probabilities over these structures is proportional to the difficulty of processing the new word; that is, the greater the update, the greater the processing difficulty or "surprisal". In broader terms, this means the more constraining a sentence is, the fewer likely possible continuations it will have and therefore the lower surprisal will be at an expected word. Conversely, at an unexpected word, surprisal will be higher. Lexical constraints are often not explicitly modelled in surprisal (Levy, 2008; Hale, 2001), but lexicalised PCFGs have demonstrated that their contribution to processing difficulty follows a similar pattern (Collins, 2003; Charniak, 2001)."

The Introduction then introduces decay, beginning with the following paragraphs on page 2:

"**Temporal activation decay.** A less well-studied factor in dependency processing is temporal activation decay. Decay is assumed to affect sentence processing in the following way: At any new word in a sentence, there may be a number of ways the sentence structure could plausibly continue. For example, the sentence *The secretary forgot...* could continue with a direct object NP (e.g. *the files*) or with a clause (e.g. *that the student...*); it has been proposed that both of these structures may be activated, but that only one will be pursued by the parser while the other is left to decay (Van Dyke and Lewis, 2003). Thus, if the parser pursues the sentence structure assuming an upcoming NP, but instead encounters the word *that...*, the decayed structure must be reactivated and reading time at the word *that* will be slower than if the expected NP had been encountered (Ferreira and Henderson, 1991; Gibson, 1998; Van Dyke and Lewis, 2003). Even if the NP parse proves to be correct, activation of the NP will decay over time such that, if it must be retrieved later (e.g. as the antecedent of a relative clause), retrieval time will become slower if

the retrieval is delayed (Lewis and Vasishth, 2005).

The above example concerns structural continuations of the sentence, but plausible continuations may also include the preactivation of specific lexical items, with the most probable item pre-integrated into the building sentence parse if its activation is strong enough (Kuperberg and Jaeger, 2016; Ness and Meltzer-Asscher, 2018). As for the structural example above, it can be assumed that lexical items preactivated but not pre-integrated are left to decay. Likewise, if future input indicates that the wrong lexical item was pre-integrated, then the decayed, correct item can be reactivated in order to repair the sentence, reflected by longer reading times. Reading times should therefore be faster if there is only one, highly probable lexical item, because the probability that the parser pursues a parse with the wrong lexical item will be low. With an increasing number of plausible lexical items, reading times should be slower, because the probability that the parser pursues a parse with the wrong lexical item increases and the reactivation of decayed items will occur more often. Even if the correct lexical item is pre-integrated, this item may too be subject to decay. However, due to stronger preactivation from the context, more predictable items are likely to have a higher starting activation and thus the effects of decay will not be as severe. Under these assumptions, less predictable lexical items are, on average, more sensitive to the effects of decay than more predictable items, leading to a more pronounced reading time slow-down (a locality effect) at less predictable dependency resolutions."

We no longer make reference to the LV05 model in the Introduction since the focus of that model is interference, which is not relevant to the current experiments. We had originally thought that LV05 might be a good framework for describing the predictive process for our particle verb stimuli and how this might interact with decay. LV05 does contain an element of predictive processing, in that it anticipates upcoming structure, and an explicit decay parameter. However, for the revision, we decided against this approach as it raised more questions than it answered. We do use the decay parameter of LV05 to simulate the effect of decay, but limit our discussion of this to the Predictions section.

- **Then, the authors should explain both frameworks' hypothe-**

ses about decay and its interaction with predictability. More explanation is needed, along with the relevant results (from German? Hindi? Persian?).

The Introduction has been revised to include a more specific account of how decay and surprisal might interact with lexical predictability; see the response to Reviewer 2's first question above. Here the changes are substantial and we refer reviewers to the sub-sections "Word predictability" and "Temporal activation decay" within the revised Introduction. In the second and third paragraphs under "Word predictability", we review evidence for the interaction of predictability with surprisal, specifically the finding that surprisal may only be a good predictor of reading times in high predictability sentences with low working memory load (Levy and Keller, 2013; Husain et al., 2014); although we note that this finding has been difficult to replicate (Vasishth et al., 2018). Then, under "Temporal activation decay", we describe a mechanism for how predictability might interact with decay by assuming that the same process of decay for structural material (Ferreira and Henderson, 1991; Gibson, 1998; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005) also applies to lexically preactivated material (Kuperberg and Jaeger, 2016).

- **Line 149 onwards "in the absence of interference, decay over distance ... will make the long condition more sensitive to predictability". Why? Do the authors claim that when a lexical item is highly predictable, it is integrated (prior to its occurrence in the input) and it is therefore amenable to decay? If so, it should be stated clearly.**

We have revised the section on page 2 to explain this more clearly. The revision of this section is large and so we do not quote it here, but to summarise: we assume that the effects of decay will show up more in *less* predictable items. The reason for this is that accounts of serial parsing propose that multiple plausible structural continuations of a sentences may be activated, but that only one parse structure is pursued while the others are left to decay (Ferreira and Henderson, 1991; Van Dyke and Lewis, 2003; Gibson, 1998). We then make the assumption that upcoming lexical items can also be pre-integrated into the pursued parse, especially if their identity is certain enough (Ku-

perberg and Jaeger, 2016; Ness and Meltzer-Asscher, 2018). However, when there is more uncertainty, the chance that the wrong word is pre-integrated would increase and the correct word will be left to decay. Thus when the real word is encountered, reactivation of the decayed, correct word will be necessary, increasing reading time. It may be that even when uncertainty is high, the correct word is still pre-integrated, but on average, this probability should be lower in the low predictability/high entropy condition. In contrast, a correct pre-integrated word (of which the probability is higher in the high predictability/low entropy condition) will not decay. Some accounts propose that decay only affects the structure pursued in working memory (Lewis and Vasishth, 2005), in which case the pre-integrated particle itself may also be subject to decay. Here we assume that more predictable particles will have a stronger starting activation, so the effects of decay at the particle site will not be as pronounced as for less predictable particles with a lower starting activation.

- **What's "highly predictable"? Consider for example a verb from the small set size group which takes five possible particles. If one of them appears in 80% of cases, and each of the other four – in 5% of cases, is the most probable one highly predictable, therefore integrated and amenable to decay? What about a "small set" verb with 60%-10%-10%-10%-10% distribution of particles and a "large set" verb with 60%-4%-4%... distribution? What happens when there's no one highly predictable completion?**

  This is certainly an important point and one that we feel is covered by the exploratory analysis using entropy instead of set size. Our quantification of entropy takes into account the distribution of possible particles activated and indeed indicated that there were experimental sentences in the small set condition (supposedly low entropy/high predictability) that actually had high entropy (low predictability) values, and vice versa for the large set condition. By collapsing the small/large set categories and using the entropy values as a predictor instead, this should mean that the distribution of particles at the low entropy (high predictability) end of the scale would have a distribution more like 60%-4%-4% and at the high entropy (low predictability) end of the scales, more like 30%-30%-30%.

It is true that we generally assume the highest probability word is integrated, even if the most probable word is only slightly more likely than other plausible words. For example, the distribution of probabilities may be something like 35%-30%-30%). In this case, the most probable word would still presumably be integrated. We would not consider this a "high predictability" situation however, because the small difference in probability might make it more likely that noise results in one of the other words being pre-integrated instead (e.g. the noise parameter in LV05 means that sometimes a word other than the highest activated word is (mis)retrieved). If there were no one highly predictable completion, then which particle gets pre-integrated could be random. It could also be that no particle is pre-integrated, depending on what the activation threshold for pre-integration is (based on your model in (Ness and Meltzer-Asscher, 2018)). The preactivated-but-not-pre-integrated particles could therefore still be affected by decay (the sentences required a particle so we assume that *something* was preactivated). In any case, all of these scenarios are captured by the continuous variable *entropy*.

- **The upshot from the last two questions is: shouldn't we look at constraint (cloze probabilities) *at the verb* in order to know what was preactivated/integrated there? Or perhaps at entropy, if it is assumed to modulate preactivation/integration (e.g. integration only happens when there are no strong competitors, i.e. low entropy), but again, *at the verb*?**

  The degree of constraint at the verb is definitely critical; however, to measure particle preactivation at the verb with a cloze test would be difficult. Because the context is so unconstrained at the verb region of the sentence, non-particle completions would represent a high number of cloze completions and we would need a large amount of data to get non-zero frequency counts for each particle, especially for verbs that take 10s of particles. Thus, while we assume particle preactivation occurs at the verb, it may only become strong enough to be detectable later in the sentence when the verb is combined with its arguments – at what exact point detectable preactivation occurs, it is difficult to know. For this particular paper, we therefore focused on whether preactivation could be sustained rather than on when it was triggered.

- **The manuscript does discuss entropy, but measured right before the particle. In the pre-test, it turns out that there's no difference between the two groups, but this is only discussed in the Results section, before carrying out the alternative analysis. I think it would be much better to acknowledge the potential problem when the pretest is presented.**

  Thank you for the recommendation. We have amended the pre-test section on page 7 by moving the discussion of the entropy issue here from the results section. We also state that although we will still present the planned analysis for transparency, the exploratory analysis with entropy is more relevant.

- **How were the verbs selected? Based on the cloze pretest, namely based on their preference after the object, before the particle? Or based on their particle selection options regardless of the specific object?**

  A section has been added to the materials section on 5 to explain how the stimuli were created:

  > "To develop the experimental stimuli, verbs were first selected
  >
  > using a corpus and dictionary search of verbs and all their possible particles. Verbs and their particle sets were grouped into small (fewer than 6 particles) and large (greater than 10 particles) categories and sentences constructed by German native speakers around small/large set pairings."

  and to the cloze text section on 6:

  > "An initial total of 48 items, each with 4 conditions (a-d), was truncated just before the particle such that the verb and the direct object of the sentence were known."

Other comments

- **I think it is natural to start the Introduction with the discussion of decay (which now appears in the second paragraph), as these are the more traditional approaches to distance effects. Then, Surprisal and anti-locality can be presented.**

This suggestion has been included in the restructure of the Introduction. Because the restructure of the Introduction is substantial, specific page references are difficult to present here. The tracked changes document may be more useful in reviewing the changes made.

- **The manipulation of decay was introduced by adding a very short constituent – a two-word phrase. Could that be the reason why no effect of decay was found? Does the LV05 predict an effect of decay with such a minimal manipulation? Related to this, line 526, "it would have been difficult to construct longer sentences without reintroducing these factors (interference), which supports the idea that they are the source of processing difficulty": why does it support this idea? I think it only means that it's very hard (perhaps impossible?) to test the influence of decay by itself.**

  Yes the constituent is very short – the example item contains a particularly short intervener, but others were longer (although not by much). The results of our simulations with the decay parameter of LV05 (see Figure 2, right panel, below) do predict a small amount of decay for a short constituent, but it is definitely possible that decay was undetectable in our stimuli.

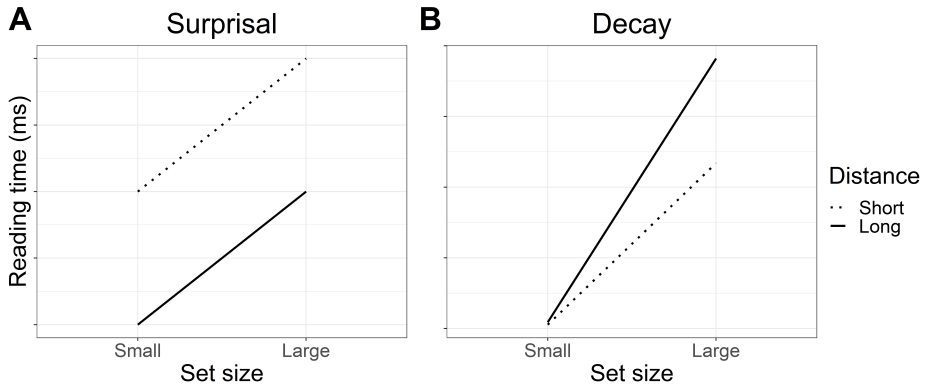

Figure 2: **Predicted interaction of lexical predictability and distance.** Informal predictions of the surprisal account and a simulation using the decay parameter of the LV05 model.

  However, the fact that decay is hard to test in isolation without introducing interference as a confound suggests that decay may just be

outweighed by interference in terms of its contribution to processing difficulty. This has certainly been the conclusion of a number of studies: we have expanded the discussion on this issue under Temporal Decay on page 21:

> "The evidence against an effect of temporal decay in both self-paced reading or eye tracking is entirely consistent with findings suggesting that decay is not an important factor influencing reading and memory recall times (Lewandowsky et al., 2009; Engelmann et al., 2019; Vasishth et al., 2019). In comparison to the sentences used in previous research, the sentences used in the current study were relatively simple, without interference or a particularly high working memory load added by the distance manipulation. However, the short adjectival modifiers used to introduce decay in our experimental stimuli may not have been long enough to introduce a detectable effect of decay. It would have been difficult to construct longer interveners without reintroducing interference or working memory load, which could support the idea that interference and working memory load are indeed the source of processing difficulty in longer sentences, rather than temporal decay. Alternatively, it could be argued that the difficulty in constructing longer sentences without introducing interference or working memory load means it is difficult or impossible to test decay in isolation and thus that we cannot know what the true effect of decay is. However, if the effect of decay is so small that it is undetectable in the face of interference and working memory load, and that these factors are almost unavoidable in constructing long dependencies, then decay is, as mentioned above, likely not a major influence on processing difficulty."

- **When entropy is first discussed, the concept should be explained – not only with a mathematical formula, but also with the intuition as to what it means.**

A more detailed explanation has been added to the Cloze Test section on page 7:

> "The *set size* manipulation was intended to induce uncertainty about the upcoming particle's lexical identity. One useful way of quantifying uncertainty is with *entropy*. Entropy provides a measure of how much information is carried by a new input in light

of all possible outcomes. In our case, the new input is the particle. In a sentence context where many particles are plausible and cloze probability is uniformly low across all the plausible particles, we assume that uncertainty about the identity of the upcoming particle is high. Thus, each of the plausible particles carries a large amount of information about the meaning of the sentence and entropy is high. In a sentence where only few particles are plausible and one particle is much more probable than the others, we assume that uncertainty about that particle's identity and the meaning of the sentence is low, and so encountering the high-probability particle will be less informative; this is a low entropy situation. To calculate entropy for our experimental stimuli, we first calculated the cloze probability (P) of all verb particles given for each respective sentence in the cloze test. Entropy (H) of the target particle was then defined as:

$$H = -\sum_{i} P_i log P_i$$

Minor comments

- **Line 47 "activation decay is anecdotally assumed. . . ": another relevant reference here is Chow & Zhou (2019), which is a replication of Wagers & Phillips (2014) (though the original authors do not frame their study as investigating decay).**

  Thanks! We have included it on line 130.

- **Line 52 "decay is not a useful predictor": perhaps also cite Van Dyke and Johns' (2012) review which argues against a role for decay in sentence processing.**

  This reference has also been included, on line 124, thank you for mentioning it.

- **Materials section: Do all the experimental verbs necessarily take particles at all? I assume this is the case, but I think this should be stated explicitly.**

All items were confirmed to require a particle via cloze testing and native speaker judgements. Any items that did not require a particle (as evidenced by the Cloze test) were presented to participants, but not included in the final analysis. The text has been updated to state this more explicitly, both in the Introduction on page 4: "Using a cloze test, we confirmed that each sentence required a particle."; and under Materials on page 5: "Each experimental item was a quartet of four sentences in which the context required a particle for the sentence to be grammatical.".

- **Line 217 "24 items that suited the experimental design" – meaning what? That they selected 6 or less, or 15 or more, particles?**

  Exactly: not only did the items elicit the required number of particles (less than 6 or more than 10), but they always elicited a particle. Detail has been added to the beginning of the Cloze test section on page 6 to state specifically how items were selected:

  > "In order to confirm that our sentence stimuli (i) elicited particles, (ii) that more particles were elicited by the large set condition than the small set condition, and to (iii) quantify the predictability of the target particle, a cloze test was conducted."

- **Online norming study (line 249 onwards): Why is this pretest necessary? In the experiment, the verb is several words upstream from the particle, so why are reading times of the verb+particle relevant?**

  It was felt that there might be some property of verb-particle constructions that leads to them being read faster or slower depending on the number of particles they take even when the verb and particle aren't separated. Perhaps, for example, base verbs that take more particles have more lexical and/or semantic associations with other lexical entries and thus net activation via passive spreading might be lower than for a verb that has fewer associates. Since we were interested in whether distance was the key factor in reading time changes, we needed to rule this possibility out. However, we have now removed this section since it is tangential to the main text.

- **Line 382, "a second possibility is that locality and antilocality**

effects simply cancelled each other out": how is this relevant to the effect of predictability, which is the topic of discussion? I would think that it is relevant to the (lack of) effect of decay, not predictability.

The suggestion that locality and antilocality may have cancelled each other out was a point about the mathematical consequences of averaging speed-ups and slow-downs, rather than a theoretical point related to predictability. However, this text no longer appears in the revised manuscript as we do not investigate it further due to space considerations.

- **Line 484 "speed up at the verb": this sounded to me like the authors were referring to a speed up at the verb relative to preceding material; it took me some time to understand that it means lower reading times in the large set verbs compared to the small set verbs.**

  This section on page 10 was indeed about the base verb region of the sentence – we do not conduct any analysis with the base verbs as they are not matched, but mention them here because it was odd that reading times were not faster for large set/high entropy verbs, despite having higher corpus frequency. We have rephrased this section to make this clearer:

  > "Mean reading times across the whole sentence for both experiments are plotted in Figure 1. One feature of these data that should be mentioned is that base verbs for sentences with higher entropy at the particle site had a higher corpus frequency than base verbs in sentences with lower entropy at the particle site (to compare verb frequency, we divided sentences into high and low entropy categories via a median split; see Appendix 2). Higher corpus frequency of the base verb should have resulted in faster reading times at the verb in high entropy sentences (Kliegl et al., 2004; Rayner and Duffy, 1986), but this was not the case in either experiment. The lack of a frequency effect at the base verb is discussed in the *General Discussion*."

- **Line 544, "a potential explanation for the lack of speed-up... more preactivated particles may have led to slower reading".**

I'm not sure I would predict this. I would think activations are not usually viewed as costly. Perhaps the source of increased reading times here is that the verbs are more ambiguous/vague, i.e. have more possible meanings?

This is a really good point and precisely what we were trying to get at – it is possible that the higher entropy base verbs themselves are more ambiguous, but a big driver of this ambiguity is the fact that they can be combined with a larger number of particles which change their meaning: so because readers can't see the particle immediately, the meaning of the verb is initially more ambiguous. This ambiguity, we assume, is associated with the increased range of preactivated particles, so I think it is conceivable that a larger amount of preactivated lexical material either results from or creates ambiguity and could be costly – this would be the idea behind the slower reading of low-constraint sentences, for example. We have updated the text to make our hypothesised link between preactivations and cost more explicit:

> "A potential explanation for the lack of a speed-up is that lexical entropy at the particle site reflected preactivation of particles at the verb. More preactivated particles would make the meaning of the verb more ambiguous, which in turn may have led to slower reading and cancelling out of the expected speed-up associated with higher frequency."

Typographical errors

The following typographical errors have been amended:

- **Line 34: length should be amount** (this sentence no longer appears in the revised Introduction)
- **Line 166: items should be item**
- **Line 128: delete second 'also'**
- **Line 319: delete second 'the'**

Other comments:

- **Line 341 and caption for Figure 5: I initially thought the RTs in the table are reading times for the particle (and wondered why they were so high). The text and caption should say that these are RTs for answering the comprehension questions. Same for line 440 and table 9.**

  The relevant text and figure captions on pages 9 and 15 have been updated to explicitly state "question response accuracy and reaction times".

- **Line 389: the number "1" is missing.**

  Here we have spelt out "one" as per APA guidelines.

- **Line 457, "the results of the statistical analysis": in all the reading time measures? If so, maybe "analyses"?**

  This has been updated to "analyses" on page 20.

Responses to comments from Reviewer 3

Major comments

- **Supplemental material is referenced... but I couldn't find supplementary material, either at the end of the PDF, in the PeerJ review materials, nor in the OSF repository.**

  This may be a lack of clarity on our part - the phrasing in the original text could be interpreted to mean that the code and the supplementary materials were separate entities. What Reviewer 3 found on OSF was indeed the entirety of the supplementary material. Throughout the manuscript, we have therefore updated any reference to code to simply state "code" rather than "code in the supplementary materials".

Minor comment (change encouraged but not mandatory):

- **Given that the stated surprisal predictions are not supported by simulations, I suggest the authors temper their claim about the predictions of surprisal [or]..., alternatively, the authors could also acknowledge that it is simply not clear**

whether or not surprisal predicts an antilocality effect for these data.

This is absolutely correct and we have tempered any statements about the predictions of the surprisal model to reinforce that they were informal; for example, in the Predictions section on page 4, where we have stated:

> "In the absence of formal quantifications for whether surprisal would predict an antilocality effect for our sentences, these predictions should be taken as an approximation of surprisal's general claim that long distance should always result in faster reading times and that higher lexical predictability should further sharpen expectations (Levy, 2008).";

as well as in the Conclusions section on page 22, where we compare our results to our predictions:

> "We compared two hypotheses of dependency processing in separable verb-particle constructions: informal predictions based on the surprisal account suggested that delaying the appearance of a verb particle could have elicited an antilocality effect, stronger in high vs. low predictable particles (Levy, 2008);"

As an aside: we would like to thank Reviewer 3 very much for the simulation results, these are very reassuring and even seem to resemble the pattern of reading times we found; at least in eye-tracking.

References

Charniak, E. (2001). Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 124–131, Toulouse, France. Association for Computational Linguistics.

Chow, W.-Y. and Zhou, Y. (2019). Eye-tracking evidence for active gap-filling regardless of dependency length. *Quarterly Journal of Experimental Psychology*, 72(6):1297–1307. Publisher: SAGE Publications.

Collins, M. (2003). Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics*, 29(4):589–637. Publisher: MIT Press.

Engelmann, F., Jäger, L. A., and Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, 43(12).

Ferreira, F. and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6):725–745.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76. ISBN: 0010-0277.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8.

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PloS one*, 9(7):e100986. Publisher: Public Library of Science.

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1/2):262–284. ISBN: 0954-1446.

Kuperberg, G. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, 31(1). ISBN: 2327-3798 2327-3801.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R. and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, 68(2):199–222. Publisher: Elsevier Inc.

Lewandowsky, S., Oberauer, K., and Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, 13(3):120–126.

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

Logačev, P. and Vasishth, S. (2013). em2: A package for computing reading time measures for psycholinguistics.

Müller, S. (2002). Particle Verbs. In Müller, S., editor, *Complex predicates: verbal complexes, resultative constructions and particle verbs in German.*, pages 253–390. CSLI: Leland Stanford Junior University.

Ness, T. and Meltzer-Asscher, A. (2018). Predictive Pre-updating and Working Memory Capacity: Evidence from Event-related Potentials. *Journal of Cognitive Neuroscience*, 30(12):1916–1938.

Ness, T. and 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(7):936–948.

Piai, V., Meyer, L., Schreuder, R., and Bastiaansen, M. C. M. (2013). Sit down and read on: Working memory and long-term memory in particle-verb processing. *Brain and Language*, 127(2):296–306.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Rayner, K., Kambe, G., and Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1061–1080. Publisher: Routledge _eprint: https://doi.org/10.1080/713755934.

Van Dyke, J. A. and 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(3):285–316. Publisher: Elsevier.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*.

Xiang, M., Dillon, B., Wagers, M., Liu, F., and Guo, T. (2014). Processing covert dependencies: an SAT study on Mandarin wh-in-situ questions. *Journal of East Asian Linguistics*, 23(2):207–232.