The effect of decay and lexical uncertainty on processing long-distance dependencies in reading

Kate Stone Corresp., [1] , **Titus von der Malsburg** [1, 2] , **Shravan Vasishth** [1]

[1] Department of Linguistics, Universität Potsdam, Potsdam, Germany

[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Corresponding Author: Kate Stone
Email address: stone@uni-potsdam.de

To make sense of a sentence, the human reader must keep track of dependent relationships between words, such as between a noun and a verb. Increasing the distance between such dependent elements may facilitate reading as expectation builds about the position and identity of the distant word; otherwise known as the antilocality effect. On the other hand, the intervening information may slow down reading via interference, working memory load, and temporal activation decay; the locality effect. While the cost of storage, integration, and similarity-based interference have well-established effects on dependency processing, the effect of temporal decay has been more difficult to test in isolation. In one self-paced reading and one eye tracking experiment, we investigated the effect of decay by delaying the appearance of a verb particle that was syntactically necessary but varied in lexical predictability. Importantly, the delay-inducing information carried no additional information about the lexical identity of the particle, or any interference-inducing components. The surprisal account predicts that expectation for the appearance of the syntactically required particle should result in an antilocality effect when its appearance is delayed, perhaps stronger with increased lexical predictability. Other accounts predict that the temporal decay may result in a locality effect when the particle is delayed, but that increased lexical predictability of the particle may make its activation more resistant to decay. The self-paced reading study provided no evidence that either temporal decay or predictability affected reading times. The eye tracking experiment provided evidence that higher predictability sped up early and total reading times, but no evidence that either decay or the interaction of predictability and decay played a role. The findings are consistent with previous research suggesting that predictability affects the early stages of word processing and that decay is not a strong influence on reading times.

The effect of decay and lexical uncertainty on processing long-distance dependencies in reading

Kate Stone[1], **Titus von der Malsburg**[1,2], **and Shravan Vasishth**[1]

[1]**Department of Linguistics, Universität Potsdam, Germany**
[2]**Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology**

Corresponding author:
Kate Stone[1]

Email address: stone@uni-potsdam.de; OrcID: 0000-0002-2180-9736

ABSTRACT

To make sense of a sentence, the human reader must keep track of dependent relationships between words, such as between a noun and a verb. Increasing the distance between such dependent elements may facilitate reading as expectation builds about the position and identity of the distant word; otherwise known as the antilocality effect. On the other hand, the intervening information may slow down reading via interference, working memory load, and temporal activation decay; the locality effect. While the cost of storage, integration, and similarity-based interference have well-established effects on dependency processing, the effect of temporal decay has been more difficult to test in isolation. In one self-paced reading and one eye tracking experiment, we investigated the effect of decay by delaying the appearance of a verb particle that was syntactically necessary but varied in lexical predictability. Importantly, the delay-inducing information carried no additional information about the lexical identity of the particle, or any interference-inducing components. The surprisal account predicts that expectation for the appearance of the syntactically required particle should result in an antilocality effect when its appearance is delayed, perhaps stronger with increased lexical predictability. Other accounts predict that the temporal decay may result in a locality effect when the particle is delayed, but that increased lexical predictability of the particle may make its activation more resistant to decay. The self-paced reading study provided no evidence that either temporal decay or predictability affected reading times. The eye tracking experiment provided evidence that higher predictability sped up early and total reading times, but no evidence that either decay or the interaction of predictability and decay played a role. The findings are consistent with previous research suggesting that predictability affects the early stages of word processing and that decay is not a strong influence on reading times.

INTRODUCTION

The speed with which an individual word in a sentence is read depends on factors such as its length, frequency, and predictability given the context (Kliegl et al., 2004). Processing a dependency *between* two words is subject to additional factors and depends on the type and length of information separating the two words. There are various accounts modelling the effect of intervening information on dependency processing. The surprisal account predicts that increasing distance between two dependent words should sharpen expectation for the distant word (Levy, 2008). However, some have suggested that distance may only sharpen expectation if working memory load is relatively low (Levy and Keller, 2013), or if the distant element is highly predictable (Husain et al., 2014; Konieczny, 2000). If the distant element is less predictable, interference and working memory constraints may negatively impact its processing (Gibson, 1998, 2000; Lewis and Vasishth, 2005; Husain et al., 2014). A further factor influencing long-distance dependencies is that of activation decay over time.

Temporal decay is presumed to play a role in sentence processing in a number of accounts and models, which predict that plausible sentence parses activated by the parser but not pursued will be left to decay (Van Dyke and Lewis, 2003; Ferreira and Henderson, 1991; Gibson, 1998; Lewis and Vasishth, 2005; Vasishth and Lewis, 2006). If a decayed parse then turns out to be the correct parse, it must be reactivated, prolonging retrieval and slowing reading time. Activation decay over time is anecdotally assumed to affect word processing times in long-distance dependencies (e.g. Xiang et al., 2014; Ness and Meltzer-Asscher, 2019) and an empirical study has demonstrated its effects over and above that of interference (Van Dyke and Lewis, 2003). However, computational models of empirical reading time data have demonstrated that the effects of temporal decay can be explained entirely by interference (Lewandowsky et al., 2009) or that decay is not a useful predictor (Engelmann et al., 2019; Vasishth et al., 2019). On the other hand, these modelling predictions are largely based on data from experiments testing interference rather than specifically testing decay. The current experiments therefore sought to test the role of temporal activation decay by manipulating distance between highly dependent sentence elements without adding similarity-based interference (Lewis and Vasishth, 2005) or new discourse referents (Gibson, 1998, 2000). In addition, we tested whether higher lexical predictability may make a word more resistant to decay.

The LV05 model (Lewis and Vasishth, 2005), while intended as a model of similarity-based interference, also makes predictions with regard to lexical predictability and decay. If an upcoming lexical item is highly predictable, it can be pre-integrated into the pursued parse, facilitating its retrieval once encountered. However, if there is uncertainty about the lexical identity of a word, this will increase the likelihood that the parser either pursues a parse with a different lexical item to the one yet to be encountered, or makes no lexical prediction at all. Both of these will increase retrieval time at the word in question, by requiring either reactivation of the parse with the correct lexical item that was left to decay, or initial activation of the unpredicted lexical item. LV05 therefore predicts that less predictable lexical items should be 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. This differs from the surprisal account, which predicts that delaying any expected syntactic or lexical element should result in faster reading times (an antilocality effect; Levy, 2008; Vasishth and Lewis, 2006).

Previous experiments directly and indirectly testing the interaction of distance and predictability have produced conflicting results. In German, it was found that reading times at the head-final verb of a relative clause were faster when a single dative argument preceded the verb than when an adjunct was added (Levy and Keller, 2013). This was taken as support for the surprisal account in low working memory load conditions, but also hinted at a potential role of verb predictability, since corpus-based conditional verb probability was higher in the dative-only than in the dative-plus-adjunct condition. Casting doubt on those results, however, is a replication attempt finding that only increased working memory load hindered reading time, regardless of what information preceded the verb (Vasishth et al., 2018).

A more direct test of the predictability/distance interaction was carried out in Hindi and Persian, with results again appearing to depend on the type of information separating the dependency. In Hindi, a highly predictable complex predicate verb appeared to outweigh the effects of long distance to be read faster than a low-predictable verb in a simple noun-verb complex (Husain et al., 2014). In comparable constructions in Persian, additional distance slowed reading of the distant verb, regardless of its predictability, even though higher predictability was associated with faster reading times overall (Safavi et al., 2016). The difference between the Hindi and Persian studies was the type of information added within the complex predicate dependencies. In Persian, a relative clause and a prepositional phrase were used as interveners (Safavi et al., 2016). Both of these introduce additional discourse referents and interference, both of which are predicted to burden working memory resources and slow reading (Gibson, 1998, 2000; Lewis and Vasishth, 2005), although discourse referents may not be the only source of slowing in longer dependencies (Gibson and Wu, 2013). In comparison, distance in the Hindi experiments was increased with adverbials, which are presumed not to add working memory load, but rather increase evidence for the position and lexical identity of the upcoming verb (Hale, 2001; Levy, 2008). Taken together, these results suggest that predictability may not be sufficient to outweigh working memory load unless the information in working memory confirms expectations.

In the current study, we sought to test the predictability/decay interaction using German particle verbs, which are complex predicates similar to the constructions used in the Hindi and Persian studies (Husain et al., 2014; Safavi et al., 2016). German particle verbs are comparable to English particle verbs in that they are composed of a base verb (e.g. “räumen”, to tidy) and a particle (e.g. “auf”, up) which can be separated (Müller, 2002). Particle verbs form a very strong dependency because the full meaning of the verb “aufräumen” (to tidy up) can only be interpreted once both the verb and particle are known. Delaying appearance of the particle therefore creates a very strong structural expectation if the context makes a particle necessary, but potentially also a strong lexical expectation for a specific particle. In

English particle verb constructions, the delay between a base verb and its particle is usually not very long; consider *to tidy up* versus *?/*to tidy the mess left after the party on Saturday up*. In German, however, long-distance separations are common. To manipulate lexical predictability of the distant particle, we compared base verbs that could take a large number of particles (10+) with verbs that can take only a small number of particles (6 or fewer). We hypothesised that the set of potential particles would be preactivated at the verb and that a larger set of particles would create more uncertainty (weaker predictability) about the eventual identity of the particle. Large set verbs therefore formed a low predictability condition and small set verbs a high predictability condition. To induce decay between the verb and its particle, we manipulated distance with a neutral intervener that added neither interference nor working memory load, nor semantic clues about the lexical identity of the dependency resolution. Any effects of the intervener on reading time should therefore be attributable to temporal decay.

The design was based on a study of Dutch particle verbs (Piai et al., 2013). In this study, it was hypothesised that Dutch verbs that can take a large number of possible particles (e.g. *delen*, which can take the particles *in, mee, op* and *ver*) should involve a larger demand on working memory than verbs with a small set size (e.g. *verdienen*, which can only take *bij*). Based on the finding that left anterior negativity (LAN) amplitude did not differ between large and small set verbs, the authors concluded that the particles themselves were *not* preactivated, but rather only the *possibility* of a downstream particle. The verb was then maintained in working memory to facilitate retrieval if and when the particle was encountered. We reasoned, however, that the distinction between small and large particle set sizes in the Dutch study was possibly too small; i.e. *small set* verbs took 2-3 particles and *large set* verbs, at least 5. We therefore categorised our German verbs into *small set* verbs that took up to 5 particles (in one case, 6), and *large set* verbs that took at least 10 particles. The current experiments therefore tested the hypotheses that 1) verbs that take particles trigger preactivation of those particles; 2) that delaying the appearance of the particle would slow reading times through temporal decay; but that 3) higher predictability would make reading times more resistant to the effects of decay.

We tested the hypotheses in self-paced reading and eye tracking modalities, both to confirm that any effects seen were not limited to a particular reading modality, but also because the two methods also provide complementary information. Self-paced reading has the advantage of forcing readers to view each word in the sentence, while eye tracking allows words to be skipped. In the current study, the target word, a particle, was very short and more likely to be skipped, making self-paced reading data valuable in examining reading time effects at the particle. On the other hand, eye tracking has the advantage of more closely resembling natural reading and is able to measure phenomena such as regressive eye movements to previous regions of the sentence and forward saccades to upcoming regions of the sentence. This allows us to generate hypotheses about the cognitive processes subserving slower or faster reading at a particular word and complements observations made in self-paced reading.

Predictions

Despite attempts to calculate surprisal using the Incremental Top-Down Parser (Roark and Bachrach, 2009) and two different types of annotated corpora (the Tiger newspaper corpus, Brants et al., 2004; and a larger corpus of novels annotated with the German version of the Stanford CoreNLP natural language software, Manning et al., 2014), the particular verb-particle combinations used in the experimental stimuli were likely too infrequent and were thus incorrectly categorised by the parser (e.g. as adverbs, verbs, and even nouns). The parser's surprisal estimates were therefore unreliable. Instead, we present informal predictions for the surprisal account, visualised in Figure 1. These should be taken as an approximation of the model's general claim that long distance should always result in faster reading times and that higher lexical predictability should further sharpen expectations (Levy, 2008; Konieczny and Döring, 2003). Note that, from here on, *set size* is used as a proxy for predictability, where a large set of particles is presumed to result in low predictability, while a small set would result in high predictability.

In contrast, a simulation using the decay parameter of the LV05 model predicts that, in the absence of interference, decay over distance will make the long distance condition more sensitive to the predictability of the particle than the short distance condition (Lewis and Vasishth, 2005). Code for the simulation is included in the supplementary materials. Figure 1 shows that the simulation predicts a larger magnitude slow-down between small and large set size in the long distance condition than in the short distance condition.

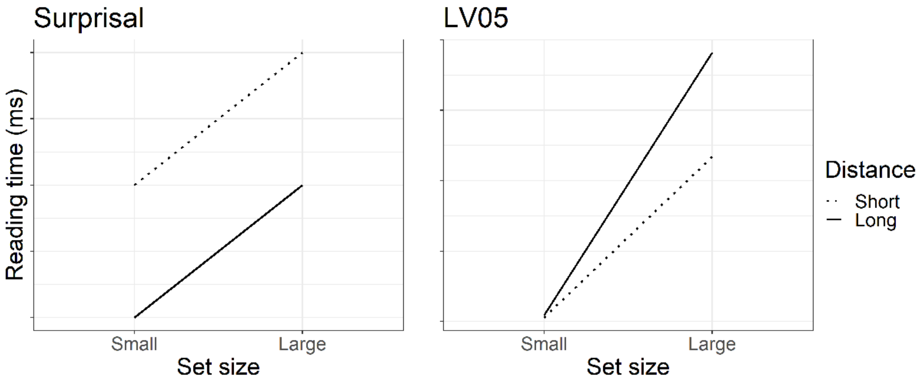

Figure 1. Predicted interaction of lexical predictability and distance. Informal predictions of the surprisal account and a simulation using the decay parameter of the LV05 model.

EXPERIMENT 1: SELF-PACED READING

Methods

Participants

Experiment 1 included a total of 60 participants (14 male, mean age = 24 years, SD = 6 years, range = 18-55 years) recruited via an in-house database. Participants were screened for acquired or developmental language disorders, neurological or psychological disorders, hearing disorders, and visual limitations that would prevent them from adequately reading sentences from the presentation computer. All participants provided written informed consent in accordance with the Declaration of Helsinki. In accordance with German law, IRB review was not required.

Materials

The study had a 2 × 2 design with *set size* (small vs. large) and *distance* (short vs. long) as factors. Each experimental items was a quartet of four sentences. In the example of an experimental item in 1 below, the verb *schrubben* (to scrub) in (a/b) can take only 2 different particles, while *spülen* (to rinse) in (c/d) can take 13. To increase distance between the verb and the particle, we added a long-distance condition where an adjectival phrase was introduced between the verb and its particle (underlined). Importantly, the adjectival phrase did not introduce any new discourse referents and did not possess any features that would interfere with the particle's retrieval. This meant that any slowing due to the additional distance could only be attributed to decay. To balance the number of words between conditions, in the short-distance condition, the intervener was inserted before the verb:

(1) a. Small set/short distance:

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

With the newly bought rag, she scrubbed the plates in the kitchen to create space for cooking.

b. Small set/long distance:

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

With the newly bought rag, she scrubbed the plates in the kitchen to create space for cooking.

c. Large set/short distance:

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

With the newly bought rag, she rinsed the plates in the kitchen to make space for cooking.

d. Large set/long distance:

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

With the rag, she rinsed the newly bought plates in the kitchen to make space for cooking.

In each experimental item, contexts were matched word-for-word, with the exception of the verb. The purpose of this was to ensure that the properties of the verb were the only factors contributing to reading times. Ideally, these properties included the number of particles each verb could take. Naturally, it cannot be ruled out that some factor resulting from the internal properties of each verb or its combination with the context contributed to differences in reading times (for example, *scrubbing* may not generate as strong an expectation for an object as *rinsing*, or vice versa). Furthermore, due to the difficulty of creating sentences with different verbs in matched contexts, it was also not possible to match the frequency of the base verb between conditions. Both of these factors are taken into consideration in interpretation of the results.

The materials used for the self-paced reading study were 24 items selected from a cloze test, separated into four lists and presented in random order. The lists were compiled using a Latin square design, such that each participant only saw one condition from each item. Each participant therefore saw 24 target sentences, interspersed with 72 filler items. The filler items were either sentences that used particle verbs in other tenses and other syntactic arrangements, or short declarative statements.

Cloze test

An initial total of 48 items, each with 4 conditions (a-d) were developed by German native speakers. A paper-and-pencil cloze test was conducted with 126 native German speakers (25 male, mean age 25 years, standard deviation 7 years, range 17-53 years). The 48 sentences were split into 4 lists such that each participant saw only one condition from every item. The 48 target sentences were randomly interspersed with 63 filler sentences, giving a total of 111 sentences per cloze test. Each sentence was cut off before either the particle (target sentences) or a clause final word (filler sentences). Participants were instructed to fill the gap with the word or words that first came to mind. The results of the cloze test yielded 24 items that suited the experimental design. It should be noted that in 8% of the stimuli, the highest cloze particle was not used as the target particle. This was because the target particle had to be matched across conditions and the highest cloze particle in one condition was therefore not always the highest cloze particle in another condition. Wherever possible, however, the highest cloze particle was used. Means and standard errors of Beta distributions corresponding to the cloze probabilities for each factor level are presented in Table 1. Since the distributions of cloze probabilities were non-normal, the means are actually not particularly informative. Entropy is therefore also presented as a measure of the uncertainty induced by each factor level. Entropy (H) was calculated as the negative logarithm of cloze probabilities (P):

$$H = -\sum_i P_i log P_i$$

A logistic mixed model was fit in *brms* (Buerkner, 2017) to the cloze probabilities of the target particles, with factor levels contrast coded as follows: small set -0.5 / large set 0.5, short distance -0.5 / long distance 0.5. The *brms* zero/one inflated Beta family was used for the likelihood to account for the presence of 0s and 1s in the data. Uninformative priors were selected for each of the predictors set size, distance, and their interaction: $\beta \sim Normal(0, 0.25)$. The full prior and model specification can be

Condition	Cloze probability Mean	Cloze probability 95% CI	Entropy Mean	Entropy 95% CI
Small set	0.51	0.28, 0.73	1.10	1.09, 1.12
Large set	0.55	0.35, 0.75	1.20	1.19, 1.22
Short distance	0.52	0.31, 0.73	1.15	1.14, 1.16
Long distance	0.53	0.32, 0.75	1.15	1.13, 1.16

Table 1. Results of the cloze test for the final set of 24 items.

found in the code provided in the supplementary materials. 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 probability of giving the target particle was lower for large set and long distance conditions than for small set and short distance conditions, as well as for the interaction. However, each of the posteriors was more or less centred on zero.

A lognormal regression model was fitted to the entropy data with the same contrast coding. The likelihood was assumed to have a lognormal distribution and the *brms* hurdle lognormal family was used to account for zeros in the data. Uninformative priors were used for the predictors set size, distance, and their interaction: $\beta \sim Normal(0, 0.01)$. This model did not suggest that entropy varied with set size, distance, or their interaction, as can be seen in Figure 2, although the mean entropy was a little higher in the large than the small set condition.

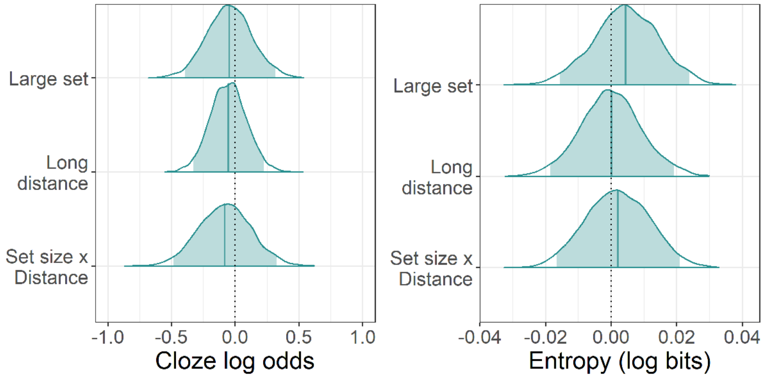

Figure 2. Change in cloze log odds and entropy of the target particle associated with each predictor. The posterior distributions are those for large set size and long distance relative to the grand mean of each condition (the dotted line). The posteriors for the small set size and short distance conditions can therefore be assumed to be the mirror image on the opposite side of the dotted line. The shaded areas are the 95% credible intervals.

Particle verb frequencies

Frequencies were computed for both the base verb and the verb-particle structure using the Tübingen aNotated Data Retrieval Application, TüNDRA, (Martens, 2013). The treebank used was the automatic dependency parse of the German Wikipedia with over 48.26 million sentences. Frequencies are presented as the incidence of the verb or particle verb per 1000 words. As can be seen in Table 2, while the frequencies of the verb+particle constructions were comparable, frequency of the base verb was notably higher in the large set condition.

Online norming study

The stimuli for the main experiment used particle verbs in sentences where the base verb appeared in second position, from which the particle was separated by verbal arguments and an intervener. The goal of the experiment was to assess whether the number of potential particles pre-activated at the verb would affect reading times at the particle itself. It was therefore important to rule out whether the verb-particle combinations themselves were associated with different reading times, even if they were not separated.

	Verb only		Verb+particle	
Condition	**Mean**	**95% CI**	**Mean**	**95% CI**
Small set	0.12	0.06, 0.25	0.04	0.02, 0.09
Large set	0.72	0.44, 1.17	0.04	0.02, 0.08

Table 2. **Mean verb and particle verb frequency per 1000 words.**

For this reason, we conducted a small online norming study to assess reading times of verb-particle constructions where the verb and particle were adjacent. The stimuli for the main experiment were therefore rearranged such that the target sentence became a subordinate clause, meaning that the base verb then appeared in final position with its particle affixed, as in the following example:

(2) a. Small set:

Die Hausfrau sagte, dass sie mit dem neu gekauften Lappen die Teller in der
The housewife said, that she with the newly bought rag the plates in the
Küche **abschrubbte**, um Platz zum Kochen zu schaffen.
kitchen **scrubbed off**, in order space for cooking to create.

The housewife said that she scrubbed/rinsed the plates in the kitchen with the newly bought rag to make space for cooking.

b. Large set:

Die Hausfrau sagte, dass sie mit dem neu gekauften Lappen die Teller in der
The housewife said, that she with the newly bought rag the plates in the
Küche **abspülte**, um Platz zum Kochen zu schaffen.
kitchen **rinsed off**, in order space for cooking to create.

The housewife said that she scrubbed/rinsed the plates in the kitchen with the newly bought rag to make space for cooking.

Participants were 20 German native speakers (6 female; mean age = 32.65, range = 21-55, sd = 10.33) recruited via the platform Prolific (www.prolific.ac). Participants received a financial reimbursement for their participation in the 30 min experiment. The only requirements for participation were German as a native language, no history of neurological or psychological illness, and access to a computer for completion of the study. One participant was excluded as their accuracy suggested inattention (M = 63%, 95% CI = 45-73%), leaving a final sample size of 19.

The items were divided into two lists and presented in random order, interspersed with 70 fillers. As for the main experiments, each participant only saw one condition from each item. Button-press time data were recorded using Ibex (Drummond, 2016). Due to the online nature of the experiment, we could not ensure that participants were attending to the task as we could in a lab setting. We therefore excluded reading times below 150 ms and above 2000 ms as indicating that participants were either speeding through the sentence without reading or reading strategically (2.57% of the data). Mean reading times by condition are shown in Table 3. Linear mixed models were fitted to the exported Ibex data using *brms* in R with full variance-covariance matrices estimated for the random effects of participant and item. Table 4 shows the reciprocal transformed estimates of the effect of set size on reading times. Large set verb-particle constructions were read faster than their small set counterparts; however, as can be seen in the model posterior in Figure 3, zero is still well within the 95% credible interval and the speed-up therefore unlikely to be meaningful.

Procedure

Participants sat in a quiet cabin in the laboratory and read the sentences in 20 point Helvetica font from a 22-inch monitor with 1680 × 1050 screen resolution. Participants saw 7 practice items before the experiment proper. The sentences were presented word-by-word in random order using the masked

Condition	Mean reading time (ms)	95% CrI
Small set	381	$358,405$
Large set	367	$346,390$

Table 3. Mean reading times for the norming study of non-separated verb-particle constructions.

Predictor	$\hat{\beta}$ (words/sec)	95% CrI
Intercept	3.02	$2.64,3.42$
Set size	0.08	$-0.05,0.22$

Table 4. Model estimates for the norming study of non-separated verb-particle constructions. The reciprocal transform means that $\hat{\beta}$ represents the model's estimated effect for each of the predictors in words per second. A positive sign therefore indicates faster reading (more words per second) and a negative sign, a slow-down. The 95% credible interval gives the range in which 95% of the model's samples fell.

self-paced reading design of Linger (Rohde, 2003). The masked words were presented as underscores separated by spaces. This meant that the participant had some clue as to the length of each word and of the sentence. Participants pressed on the space bar to reveal the next word. The previous word disappeared when the next word appeared, meaning that only one word was visible at any time. Linger recorded the time between word onset and spacebar press, and this data was exported for analysis. After each sentence, a yes/no question appeared which participants answered with the *u* (No) and *r* (Yes) keyboard keys. Feedback was not given. The questions concerned the content of the sentences; for example, in the example Item 1 above, the question was "Were the plates in the kitchen?". We ensured that the questions targeted a balanced range of sentence regions. A break was offered after every 50 sentences. All other settings were left at their defaults.

Analysis

Linear mixed models with full variance-covariance matrices estimated for the random effects of participant and item were fitted to the exported Linger data using *brms* (Buerkner, 2017) in R. The dependent variable was reading time at the particle with a reciprocal transform as suggested by the Box Cox procedure (Box and Cox, 1964). The predictors *set size* and *distance* were effect contrast coded: -0.5 (small set/short distance), 0.5 (large set/long distance). The model priors were as follows:

$$\begin{gathered}
\beta_0 \sim Normal(3, 0.5) \\
\beta_{1,2,3} \sim Normal(0, 0.5) \\
\upsilon \sim Normal(0, \sigma_\upsilon) \\
\gamma \sim Normal(0, \sigma_\gamma) \\
\sigma_\upsilon, \sigma_\gamma \sim Normal_+(0, 0.25) \\
\rho_\upsilon, \rho_\gamma \sim LKJ(2) \\
\sigma \sim Normal_+(0, 0.25)
\end{gathered}$$

The prior distribution of the intercept was determined using domain knowledge that mean reading time is approximately 3 words per second under a 1000/y reciprocal transform and that 95% of reading speeds should fall within a range of 2 and 4 words per second. The slope adjustments, for example β_1 (*set size*), were centred on zero and assumed that the expected the effect of set size would be to either increase or decrease reading speed by 1 word per second. By-subject and by-trial adjustments to the slope and intercept (υ, γ) were also centred on zero with respective priors reflecting their plausible standard deviations. The prior for the correlation parameters ρ of these random effects is a so-called LKJ prior in Stan, which takes a hyperparameter η with value 2; this LKJ(2) prior represents a distribution ranging from -1 to $+1$, but favouring correlations closer to 0. Finally, the prior for the standard deviation parameter σ for the residual is a $Normal(0, 0.25)$ truncated at 0. The full model specification can be found in the supplementary materials.

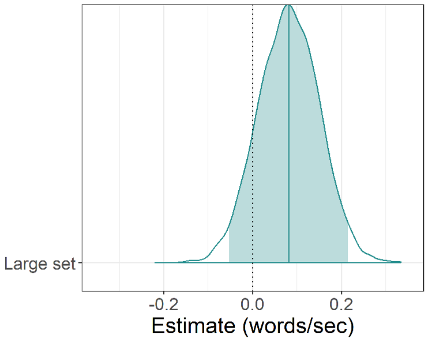

Figure 3. Change in self-paced reading speed in the online norming study. The curve is the posterior distribution associated with the large set condition relative to the grand mean of large and small set conditions (dotted line). Due to the reciprocal transform, a shift in the posterior to the right of zero indicates faster reading speed in the large than in the small set condition. The shaded area is the 95% credible interval.

To decide whether the effects of *distance* and *set size* were consistent with the null hypothesis that there was no effect, Bayes factors (BF) were computed. The BF gives the ratio of marginal likelihoods for one model against another (Jeffreys, 1939). We therefore compared the planned analysis model including all predictors (described above) against reduced models without the predictor of interest. For example, when we wanted to decide whether the effect of *set size* was not zero, we computed a BF for the model with set size (referred to as model 1) versus a reduced model without set size (referred to as model 0), i.e. BF_{10}. A BF of around 1 indicates no evidence in favour of either model. A BF of greater than 3 (when the comparison is BF_{10}) will be taken as evidence in favour of the model with the effect, and a BF of less than $\frac{1}{3}$ as evidence in favour of the null hypothesis. We assessed the strength of the evidence with reference to the conventional BF classification scheme (Jeffreys, 1939). We computed BFs not only for the planned models, but also for models with more and less informative priors. Computing BFs with a variety of priors is recommended, since the BF is sensitive to the prior used (Lee and Wagenmakers, 2013).

RESULTS

Accuracy and reaction times

Mean comprehension accuracy and reaction times in all four conditions are set out in Table 5.

	Accuracy (%)		Reaction time (ms)	
Condition	**Mean**	**95% CI**	**Mean**	**95% CI**
(a) Small set, short distance	92	89, 95	1944	1862, 2031
(b) Small set, long distance	93	90, 95	2020	1918, 2128
(c) Large set, short distance	94	91, 96	1996	1897, 2100
(d) Large set, long distance	93	91, 96	1963	1872, 2058

Table 5. Summary of accuracy and reaction times for the self-paced reading experiment.

Planned analysis

Mean self-paced reading speed by condition are shown in Table 6 and the model estimates in Table 7. The 95% credible intervals of each of the posteriors contain zero, suggesting that there was uncertainty about how these factors influenced reading speed, if at all. The Bayes factors for all effects were between weakly and strongly in favour of the null hypothesis.

The categorical predictor *set size* used in the planned analysis was intended as a proxy for entropy, where a large set size was supposed to reflect high entropy and thus lower predictability. However, although these categories may have reflected the number of particles associated with each base verb, the results of the cloze test suggested they did not represent the range of particle completions provided at

Condition	Mean reading time (ms)	95% CrI
(a) Small set, short distance	442	$421, 464$
(b) Small set, long distance	451	$429, 474$
(c) Large set, short distance	428	$408, 448$
(d) Large set, long distance	429	$409, 449$

Table 6. Mean self-paced reading speed by condition.

			BF_{10}:		
Predictor	$\hat{\beta}$ **(words/sec)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
Intercept	2.50	$2.33, 2.67$	-	-	-
Set size	0.07	$-0.02, 0.16$	1.32	0.28	0.20
Distance	-0.02	$-0.09, 0.06$	0.31	0.07	0.05
Set size x Distance	0.02	$-0.15, 0.18$	0.88	0.23	0.07

Table 7. Self-paced reading speed model estimates with *set size* as a categorical predictor. The reciprocal transform means that $\hat{\beta}$ represents the model's estimated effect for each of the predictors in words per second. A positive sign therefore indicates faster reading (more words per second) and a negative sign, slower reading. The 95% credible interval gives the range in which 95% of the model's samples fell.

the particle site. This can be seen in Figure 4: sentences in the large set condition elicited, on average, a broader variety of particle completions (higher entropy), but there were items in both conditions that elicited both a large and a small set of particle completions. We therefore decided to analyse entropy as a continuous predictor instead, since this would map much better to our planned manipulation of predictability (high entropy = low predictability and vice versa).

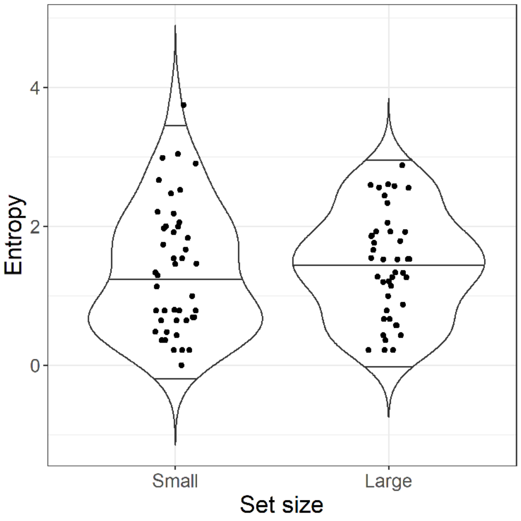

Figure 4. By-item entropy within small and large set categories. Violin plots show the median and 95% quantiles.

Exploratory analysis

Entropy as a continuous predictor

In an exploratory analysis, entropy at the particle was refitted as a continuous predictor and its effect on reading speed examined. The priors and model specification remained the same as for the planned analysis. Reading speed predicted by the model is plotted in Figure 5. The numerical pattern suggests an interesting mix of the two models; that is, when predictability was high (low entropy), reading speed was faster at long distance in line with the surprisal accounts. In contrast, when predictability was low (high entropy), the pattern more closely resembles that predicted by the LV05 model. However, these patterns are not further interpreted as the statistical analysis did not support an interaction effect.

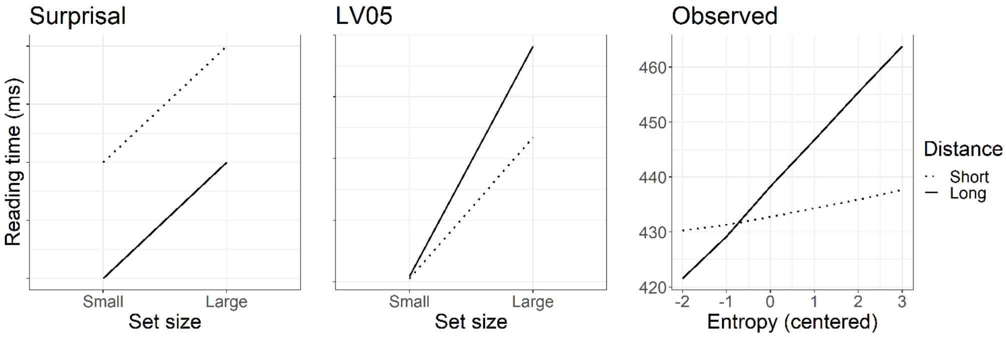

Figure 5. **Predicted versus modelled self-paced reading times.** Note that this figure represents only a predicted reading speed pattern based on the model output and that there was no statistical support for the interaction of distance and entropy.

The model coefficients are summarised in Table 8. As can also be seen in Figure 6, zero is well within the 95% credible interval for the posterior of the all predictors. The Bayes factor analysis found no evidence for any of the predictors over the null hypothesis. In other words, there was no evidence that either entropy, distance, or their interaction affected reading speed.

				BF_{10}:	
Predictor	$\hat{\beta}$ **(words/sec)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
Intercept	2.51	$2.32, 2.69$	-	-	-
Entropy	-0.04	$-0.13, 0.05$	0.51	0.14	0.07
Distance	-0.02	$-0.11, 0.07$	0.42	0.10	0.05
Entropy x Distance	-0.02	$-0.15, 0.10$	0.52	0.05	0.01

Table 8. **Self-paced reading speed estimates with entropy as a continuous predictor.** As for the planned analysis, the reciprocal transform means that $\hat{\beta}$ represents the model's estimated effect for each of the predictors in words per second. A positive sign therefore indicates faster reading (more words per second) and a negative sign, slower reading. The 95% credible interval gives the range in which 95% of the model's samples fell. Bayes factors are presented for a range of β priors including, from left to right: more informative than the prior used in the planned analysis, $N(0, 0.1)$; the prior used in the planned analysis, $N(0, 0.5)$; and more diffuse than the prior used in the planned analysis, $N(0, 1)$. BF_{10} indicates the Bayes factor for the full model (1) against a reduced model (0). BFs of less than $\frac{1}{3}$ indicate evidence for the reduced model, while BFs greater than 3 suggest evidence for the full model.

Discussion of self-paced reading results

We hypothesised that temporal activation decay would lead to slower reading of verb particles at long distance versus short, but that higher lexical uncertainty about the identity of the particle (lower predictability) would be more sensitive to the effects of long distance than when the particle was predictable. Neither the planned nor the exploratory analyses supported these hypotheses, contrasting with both the surprisal and the LV05 model predictions. One potential explanation may lie in the very small differences in cloze probably and entropy at the particle site, meaning that entropy between set size conditions was effectively matched at that point in the sentence. Examples of entropy differences between condition means discussed elsewhere in the literature include 0.38 or 0.50 bits (Levy, 2008), 0.57 bits (Linzen and Jaeger, 2016), and reductions of up to 53 bits (Hale, 2006). In comparison, our between-category difference was only 0.10 bits. However, the examples given from the literature are derived from syntactic entropy *of the*

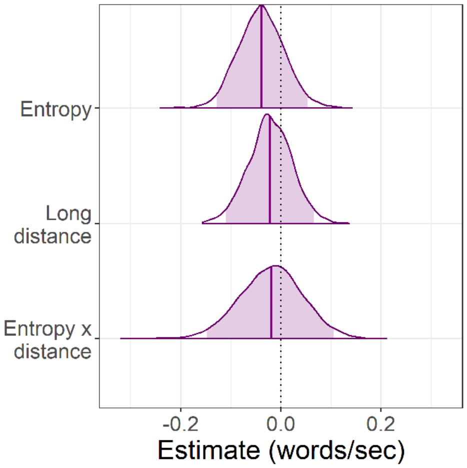

Figure 6. Change in self-paced reading speed at the particle with entropy as a continuous predictor. Now that entropy is a continuous predictor, the posterior represents the change in reading time elicited by a 1-unit increase in entropy. Due to the reciprocal transform, a shift in the posterior to the left of zero indicates slower reading speeds. The dotted line represents the grand mean of the two factor levels of each predictor and the shaded areas, the 95% credible intervals.

rest of the sentence, while ours were based on lexical entropy *at the particle*. Nonetheless, the small between-category difference should have been ameliorated by the reanalysis of entropy as a continuous predictor and yet this was not the case. A second possibility is that locality and antilocality effects simply cancelled each other out. We therefore turn to the eye tracking results for further information.

EXPERIMENT 2: EYE TRACKING

The eye-tracking experiment was conducted using the same materials as the self-paced reading study and maintained the original hypotheses visualised in Figure 1.

METHODS

Participants

Sixty German native speakers were recruited, of which one was excluded due to the presence of a neurological disorder. The remaining 59 (13 male) were free of current or developmental disorders, speech or hearing disorders, or vision impairments that could not be corrected without impeding the eye-tracker (e.g. glasses and contacts occasionally caused reflection preventing accurate calibration of the eye-tracker, meaning that these participants had to be excluded if they were unable to read without visual correction). The mean age of the participants was 26 (SD = 6, range = 18-47) and all were university educated. All participants provided written informed consent in accordance with the Declaration of Helsinki. In accordance with German law, IRB review was not required.

Materials

The experimental materials and presentation lists were identical to those used in the self-paced reading study.

Procedure

Right eye monocular tracking was conducted using an EyeLink 1000 eye-tracker (SR Research) with a desktop-mounted camera and a sampling rate of 1000 Hz. The head was stabilised using a chin and forehead rest which set the eyes at a distance of approximately 66cm from the presentation monitor. The experimental paradigm was built and presented using Experiment Builder (SR Research). The 22-inch presentation monitor had a screen resolution of 1680 x 1050. Sentences were presented in size 16-point Courier New font on a pale grey background (hex code #cccccc). Each experimental session began with calibration of the eye-tracker, which was repeated if necessary during the experiment. The experimental sentences were preceded by six practice sentences. Participants fixated on a dot at the centre left of the screen before each sentence was presented. Once they had finished reading, they fixated on a dot at the bottom right of the screen. Each of the experimental sentences was followed by the same yes/no question used in the self-paced reading study, which the participant answered using a gamepad. Each session lasted approximately 30 minutes.

Data analysis

Sampled data were exported from DataViewer (SR Research) and pre-processed in R using the *em2* package (Logačev and Vasishth, 2013). Linear mixed-effects models with full variance-covariance matrices estimated for the random effects of participant and item were fitted using *brms* (Buerkner, 2017) in R (Team, 2018) separately to data for each of four reading time measures, first fixation duration (FFD), first pass reading time (FPRT), total fixation time (TFT), and regression path duration (RPD). This range of measures was selected as both early and late measures have been found to be affected by predictability (Kliegl et al., 2004; Boston et al., 2008), although perhaps earlier measures are more sensitive (Staub, 2015). The target region of the sentence was the particle plus the immediately preceding word, since the particles were usually short (2-3 letters) and therefore not always fixated. The preceding rather than the following word was chosen because the target particle was at the right clause boundary. The dependent variable was reading time at the particle, log transformed as indicated by the Box Cox procedure. The predictors set size and distance were effect contrast coded: -0.5 (small set/short distance), 0.5 (large set/long distance). The model priors were as follows:

$$\begin{gathered}
\beta_0 \sim Normal(5.7, 0.5) \\
\beta_{1,2,3} \sim Normal(0, 0.5) \\
\upsilon \sim Normal(0, \sigma_\upsilon) \\
\gamma \sim Normal(0, \sigma_\gamma) \\
\sigma_\upsilon, \sigma_\gamma \sim Normal_+(0, 1) \\
\rho_\upsilon, \rho_\gamma \sim LKJ(2) \\
\sigma \sim Normal_+(0, 1)
\end{gathered}$$

The prior distribution of the intercept was determined using domain knowledge that mean reading time is approximately 300 ms (5.7 on the log scale) and that 95% of reading times should fall within a range of 110 and 812 ms. We expected the effect of the predictors would mostly lie somewhere between a speed-up of 190 ms and a slow-down of 513 ms. Priors for the random effects parameters were as shown above. The full model specification can be found in the code in the supplementary materials.

RESULTS

Accuracy and reaction times

Mean comprehension accuracy and reaction times in all four conditions are set out in Table 9.

	Accuracy (%)		Reaction time (ms)	
Condition	**Mean**	**95% CI**	**Mean**	**95% CI**
(a) Small set, short distance	91	88, 94	2052	1967, 2141
(b) Small set, long distance	92	89, 95	2090	2007, 2177
(c) Large set, short distance	96	94, 98	2007	1928, 2089
(d) Large set, long distance	97	94, 98	2051	1978, 2126

Table 9. Summary of accuracy and reaction times in the eye tracking experiment.

Planned analysis

Observed reading times per condition are summarised in Table 10. The model estimates for each reading time measure are shown in Table 11. The 95% credible interval for each of the posteriors contains zero, suggesting that it was uncertain whether the predictors' effect on any reading time was positive or negative, or zero. However, as for the self-paced reading experiment (Experiment 1), the categorical distinction of large and small set size was probably inappropriate, and thus an exploratory analysis using entropy as a continuous predictor is presented next. One limitation of the Bayes factors analyses is that we are evaluating multiple dependent measures which are correlated to each other (von der Malsburg and Angele, 2016). Our analyses should therefore be considered exploratory, and should be confirmed via future replication attempts.

Measure	Condition	Mean reading time (ms)	95% CrI
FFD	(a) Small set, short distance	284	269, 299
	(b) Small set, long distance	285	270, 301
	(c) Large set, short distance	292	277, 309
	(d) Large set, long distance	303	287, 319
FPRT	(a) Small set, short distance	316	297, 335
	(b) Small set, long distance	313	294, 333
	(c) Large set, short distance	324	304, 345
	(d) Large set, long distance	337	317, 357
TFT	(a) Small set, short distance	368	343, 395
	(b) Small set, long distance	364	338, 391
	(c) Large set, short distance	370	344, 397
	(d) Large set, long distance	381	355, 408
RPD	(a) Small set, short distance	354	330, 379
	(b) Small set, long distance	355	330, 382
	(c) Large set, short distance	359	334, 386
	(d) Large set, long distance	380	354, 408

Table 10. Mean eye-tracking reading times by condition.

Exploratory analyses

Entropy as a continuous predictor

As for the self-paced reading analysis, models were refit using entropy as a continuous predictor. The predicted versus observed interactions of distance and entropy are plotted in Figure 7. Numerically, the pattern of reading times again appeared to be a mixture of the surprisal and LV05 predictions. However, the results of the statistical analysis did not support an interaction of entropy and distance, and so this pattern is not further interpreted.

The model estimates can be seen in Table 12 and the model posteriors in Figure 8. The Bayes factor (BF) analysis found evidence for an effect of entropy on first fixation duration (FFD), first pass reading time (FPRT), and total fixation time (TFT), in that increasing entropy slowed reading times. With more informative priors, BFs suggested evidence for the effect of entropy in each of these three measures was strong. At the planned (non-informative, regularising) prior for regression path duration (RPD), BF evidence for an effect of entropy was inconclusive. However, when the more informative prior was used, evidence for an effect of entropy on RPD was strong. The BFs for the remaining predictors (distance, entropy x distance) were in favour of the null hypothesis.

DISCUSSION OF EYE-TRACKING RESULTS

The planned analysis with the categorical predictor *set size* again did not find any support for our hypotheses that temporal activation decay would be more prominent when lexical predictability was low. Reconfiguring set size as the continuous predictor *entropy*, however, found support for the hypothesis that increased uncertainty about the lexical identity of the particle would slow reading times. There was no evidence that temporal decay alone, or in interaction with entropy, influenced reading times.

SELF-PACED AND EYE-TRACKING READING TIMES COMPARED

The statistical analysis at the particle region differed quite considerably between self-paced reading (SPR) and eye tracking, finding no effect of any predictor in SPR but an effect of entropy in eye tracking.

					BF_{10}:	
Measure	**Predictor**	$\hat{\beta}$ **(log ms)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
FFD	Intercept	5.66	$5.55, 5.75$	-	-	-
	Set size	0.02	$-0.01, 0.05$	1.69	0.10	0.02
	Distance	0.01	$-0.02, 0.03$	0.27	0.06	0.04
	Set size x Distance	0.01	$-0.02, 0.03$	0.19	0.00	0.00
FPRT	Intercept	5.74	$5.58, 5.89$	-	-	-
	Set size	0.02	$-0.01, 0.05$	2.02	0.10	0.02
	Distance	0.00	$-0.02, 0.03$	0.27	0.05	0.03
	Set size x Distance	0.01	$-0.02, 0.03$	0.32	0.01	0.00
TFT	Intercept	5.89	$5.71, 6.06$	-	-	-
	Set size	0.00	$-0.04, 0.04$	1.16	0.09	0.02
	Distance	0.00	$-0.03, 0.03$	0.28	0.05	0.03
	Set size x Distance	0.01	$-0.04, 0.04$	0.59	0.02	0.00
RPD	Intercept	5.86	$5.69, 6.03$	-	-	-
	Set size	0.01	$-0.03, 0.05$	1.38	0.08	0.02
	Distance	0.01	$-0.02, 0.04$	0.41	0.07	0.04
	Set size x Distance	0.01	$-0.02, 0.04$	0.80	0.05	0.01

Table 11. Eye-tracking model estimates for the planned analysis with *set size* as a categorical predictor. $\hat{\beta}$ represents the model's estimated effect for each of the predictors on the log scale. The log transform means that estimates with a positive sign indicate slower reading times and that readers who are slower on average will be more affected by the manipulation than faster readers. The 95% credible interval gives the range in which 95% of the model's samples fell.

Despite the lack of statistical congruity between the two modalities, Figure 5 and Figure 7 suggested a similar numerical pattern of effects at the particle. The numerical pattern suggested that when lexical predictability was high (low entropy), a surprisal-like antilocality effect was seen at long distance. In contrast, when lexical predictability was low (high entropy), a locality effect was seen, congruent with the hypothesis that low predictability would be more sensitive to the effects of temporal decay. Across the rest of the sentence, reading times were also similar between modalities, as can be seen in Figure 9. However, the statistical analysis at the particle region and the 95% confidence intervals for the mean reading times over the rest of the sentences in Figure 9 warn against overinterpretation of these patterns.

One feature of Figure 9 that should be mentioned, however, is that there does not appear to be a speed up at the verb in either modality as would be expected with the higher frequency of *large set* verbs (Kliegl et al., 2004; Rayner and Duffy, 1986). However, in light of the fact that *set size* was not a good proxy for lexical entropy, we recalculated verb frequency for entropy divided into high and low categories via a median split. As can be seen in Table 13, frequency of the base verb was still higher in the high entropy category, meaning that a speed-up at high-entropy verbs should still have been expected. This is discussed below.

GENERAL DISCUSSION

In two reading time experiments, we tested whether delaying the appearance of a structurally necessary verb particle would increase reading speed in line with the surprisal account (Levy, 2008), or whether the particle's lexical predictability might interact with the effects of decay in line with the LV05 model (Lewis and Vasishth, 2005). The planned analyses of both a self-paced reading and an eye tracking experiment provided no evidence of an effect of either the predictability of the particle or of delaying its appearance. In a more appropriate exploratory analysis using entropy as a continuous predictor at the particle site, there was again no evidence of an effect of either predictor on self-paced reading times. However, there was evidence in eye-tracking that higher particle predictability led to faster reading times, although there was again no evidence of an effect of distance.

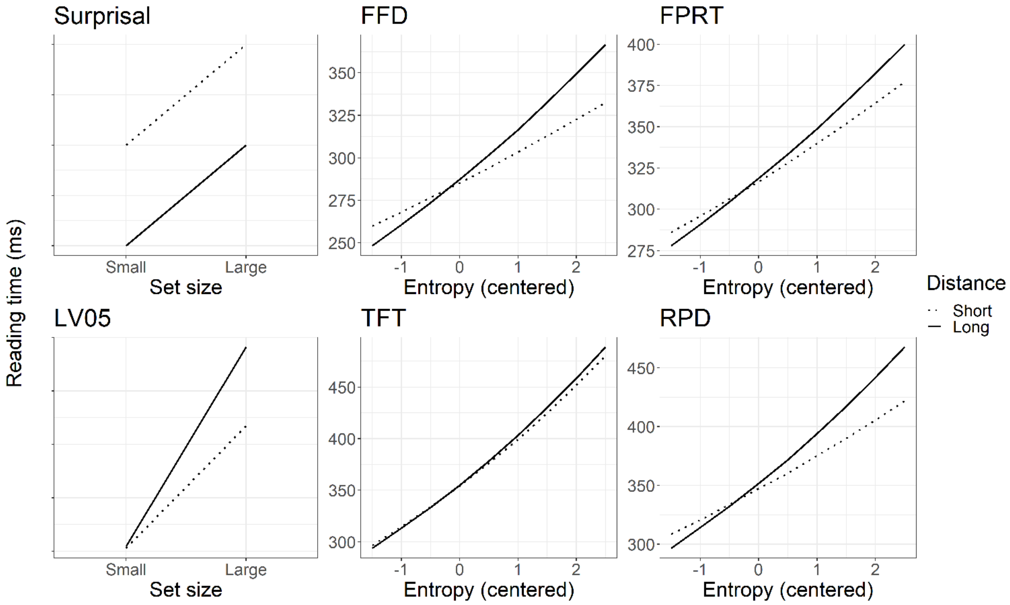

Figure 7. **Predicted versus modelled interaction of entropy and distance.** Note that this figure represents only predicted reading time patterns based on the model output and that there was no statistical support for the interaction of distance and entropy.

Predictability

The findings in the eye tracking data are somewhat consistent with evidence suggesting that the effects of predictability influence early stages of lexical processing and thus that its effects are more likely to be detected in early eye tracking measures (Staub, 2015), as well as gaze duration (Rayner, 1998). Somewhat inconsistent with this proposal was the fact that we observed a predictability effect in all four of our eye tracking measures, including regression path duration. However, this may have been due to the fact that first fixation durations were included in the computation of the remaining three measures, meaning that the primary source of the effect may have actually been first fixation duration. On the other hand, the effects of syntactic surprisal have been found in both early and late measures, including regression path duration (Boston et al., 2008). Although syntactic surprisal was not a factor in the current study, it is conceivable that the principle underlying the effect of syntactic surprisal on reading times would also apply to lexical surprisal. An argument against interpreting the effect in regression path duration, however, is the lack of evidence for an effect of predictability in self-paced reading times.

The lack of evidence for the predictability effect in self-paced reading was likely due to the fact that self-paced reading times reflect a combination of early and late processes, since readers are not able to regress to previous parts of the sentence. For this reason, self-paced reading times should arguably resemble regression path duration or total fixation times more than earlier measures such as first fixation duration. If it was indeed the case that the predictability effect in our regression path duration and total fixation measures was being driven solely by the inclusion of first fixation durations in their computation, this may explain why the effect was not also seen in self-paced reading.

Temporal decay

The lack of evidence for an effect of temporal decay in either self-paced reading or eye tracking is entirely consistent with findings suggesting that decay is not an important factor influencing reading 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. It would have been difficult to construct longer sentences without reintroducing these factors, which supports the idea that they are the source of processing difficulty in longer sentences, rather than temporal decay.

					BF_{10}:	
Measure	**Predictor**	$\hat{\beta}$ **(log ms)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
FFD	Intercept	5.66	$5.55, 5.76$	-	-	-
	Entropy	0.08	$0.03, 0.13$	23.88	4.65	2.15
	Distance	0.01	$-0.05, 0.07$	0.28	0.06	0.03
	Entropy x Distance	0.04	$-0.04, 0.11$	0.32	0.01	0.00
FPRT	Intercept	5.76	$5.61, 5.90$	-	-	-
	Entropy	0.08	$0.03, 0.13$	17.71	4.49	1.86
	Distance	0.00	$-0.06, 0.07$	0.27	0.06	0.03
	Entropy x Distance	0.02	$-0.06, 0.10$	0.19	0.00	0.00
TFT	Intercept	5.87	$5.70, 6.04$	-	-	-
	Entropy	0.12	$0.04, 0.21$	24.65	4.77	2.78
	Distance	0.00	$-0.06, 0.07$	0.32	0.07	0.04
	Entropy x Distance	0.01	$-0.08, 0.09$	0.22	0.00	0.00
RPD	Intercept	5.85	$5.67, 6.02$	-	-	-
	Entropy	0.10	$0.03, 0.18$	12.58	2.91	1.18
	Distance	0.01	$-0.05, 0.08$	0.35	0.07	0.03
	Entropy x Distance	0.04	$-0.06, 0.12$	0.41	0.01	0.00

Table 12. Eye-tracking model estimates with entropy used as a continuous predictor. $\hat{\beta}$ represents the model's estimated effect for each of the predictors on the log scale. The log transform means that estimates with a positive sign indicate slower reading times and that readers who are slower on average will be more affected by the manipulation than faster readers. The 95% credible interval gives the range in which 95% of the model's samples fell. Bayes factors are presented for a range of β priors including, from left to right: more informative than the prior used in the planned analysis, $N(0, 0.1)$; the prior used in the planned analysis, $N(0, 0.5)$; and more diffuse than the prior used in the planned analysis, $N(0, 1)$. BF_{10} indicates the Bayes factor for the full model (1) against a reduced model (0). BFs of less than $\frac{1}{3}$ indicate evidence for the reduced model, while BFs greater than 3 suggest evidence for the full model.

Particle preactivation at the verb

In spite of the lack of evidence for an effect of decay, the effect of lexical predictability at the particle is nonetheless interesting. As all words in all sentences were identical except for the verb, the only information influencing uncertainty at the particle site was the verb. This supports the possibility that particle options were preactivated at this point of the sentence. Alternatively, if preactivation did not occur at the verb, it may have resulted from the combination of the verb and direct objects immediately adjacent; for example, *...**spülte** sie die Teller...* (she **rinsed** the plates) should be sufficient to anticipate the most likely verb-particle combinations. The preactivation of particles is unlikely to have been triggered by information between the direct object and the particle site (e.g. *in der Küche*, in the kitchen), since this region did not add any information about the identity of the particle. It is therefore possible to conclude from the results that lexical preactivation occurred well before the particle was seen.

One final feature of interest in the data and perhaps in further support of particle preactivation at the verb is the fact that base verbs associated with higher entropy at the particle were higher in frequency, and yet were not read faster. High word frequency is strongly correlated with faster reading time (Kliegl et al., 2004; Rayner and Duffy, 1986). 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 may have led to slower reading, cancelling out the expected speed-up due to higher frequency.

It has previously been proposed that particles are not preactivated at all at the base verb, but rather that verbs that take particles are maintained in working memory to facilitate retrieval when the particle is finally encountered (Piai et al., 2013). Our findings offer a potential contradiction to this hypothesis. If particles had not been preactivated in the current study, there should have been no effect of entropy at all at the particle, since there is no reason to think that the base verbs associated with higher entropy would have required more resources to retrieve than base verbs associated with lower entropy, or vice versa. The

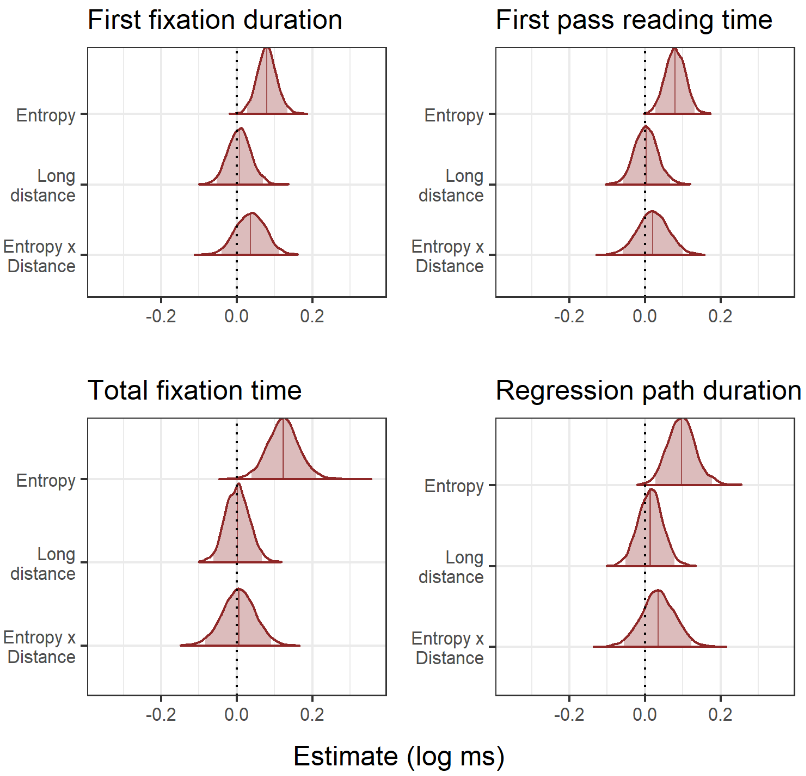

Figure 8. Changes in reading time for each eye-tracking measure using entropy as a continuous predictor. Now that entropy is a continuous predictor, the posterior represents the change in reading time for the average reader elicited by a 1-unit increase in entropy. The log transformed reading times mean that posteriors shifted to the right of zero indicate slower reading. Error bars show the 95% credible intervals.

possible cancelling out of the expected frequency effect at the verb may be further evidence against a non-preactivation account. A future test of this hypothesis would be to hold the verb and particle constant, and manipulate other regions of the sentence. This exact design has been tested using event-related potentials and will be presented in forthcoming work. However, in the current experiments, maintenance of the verb in working memory would not explain why low entropy particles should show faster reading times in eye tracking measures than high entropy ones.

CONCLUSIONS

The surprisal account would predict that delaying the appearance of a verb particle should have sharpened expectation and sped up reading times (Levy, 2008). In contrast, the LV05 account would predict that delaying the particle may result in temporal activation decay, but that highly lexically predictable particles would be more resistant to its effects (Lewis and Vasishth, 2005). Contrary to both these hypotheses, we found no evidence that distance had any effect on reading times. We did find evidence that higher predictability facilitated reading times, but only in eye-tracking measures. There was no evidence for an effect of predictability in any direction in self-paced reading. Since distance in the current study was induced with information that neither hinted at the identity of the upcoming verb particle nor increased interference or working memory load, our results suggest that the surprisal-based speed-ups observed at long distance in previous research may be due to the additional intervening information confirming lexical expectations. Our results also support previous modelling findings that temporal working memory decay is not a strong influence on reading times; at least not in simple, grammatical sentences.

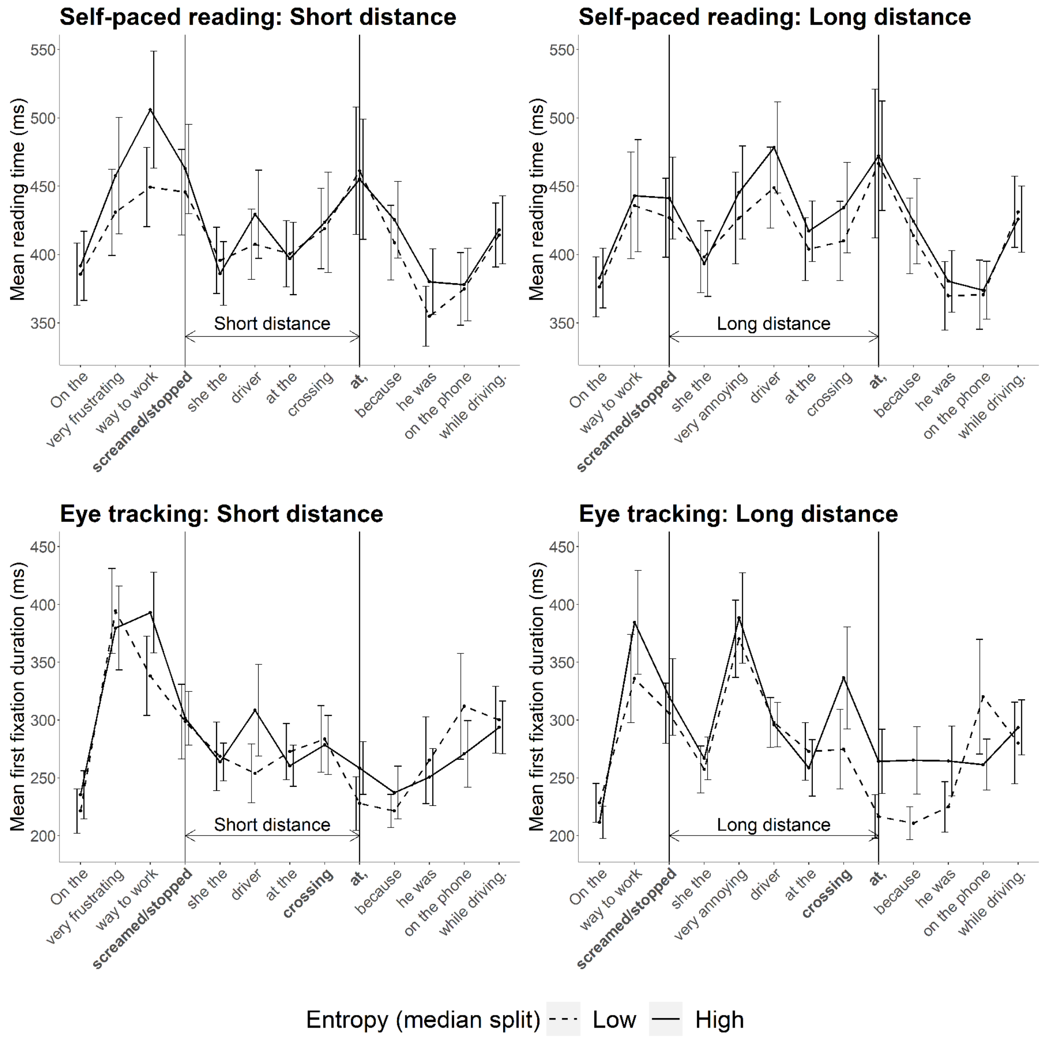

Figure 9. **Comparison of self-paced reading and eye tracking times plotted across the sentence.** Error bars show 95% confidence intervals.

	Verb only		Verb+particle	
Condition	**Mean**	**95% CI**	**Mean**	**95% CI**
Low entropy	0.17	0.11, 0.28	0.04	0.03, 0.07
High entropy	0.42	0.26, 0.69	0.04	0.03, 0.07

Table 13. **Mean verb and particle verb frequency per 1000 words for high and low entropy.** Entropy was categorised via median split.

Appendix 1 Data and code. All data and code necessary to reproduce our analyses are available here: https://osf.io/xwcvp/

ACKNOWLEDGMENTS

This work was supported by a scholarship from the Universität Potsdam Graduate School. The authors also wish to acknowledge the role of valuable feedback from reviewers and colleagues at AMLaP 2017 and CUNY 2018 in shaping the final paper.

REFERENCES

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1).

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Buerkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1).

Drummond, A. (2016). Ibex: Software for psycholinguistic experiments.

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

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.

Gibson, E. (2000). The Dependency Locality Theory : A Distance -Based Theory of Linguistic Complexity. In Marantz, A., Miyashita, Y., and O'Neil, W., editors, *Image, language, brain*, pages 95–126. MIT Press.

Gibson, E. and Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2):125–155.

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.

Hale, J. (2006). Uncertainty About the Rest of the Sentence. *Cognitive Science*, 30(4):643–672.

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.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

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.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–45.

Konieczny, L. and Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of iccs/ascs*, pages 13–17.

Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge.

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.

Linzen, T. and Jaeger, T. F. (2016). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, 40(6). ISBN: 1551-6709 (Electronic)\r0364-0213 (Linking).

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

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (ACL) system demonstrations*, pages 55–60.

Martens, S. (2013). TüNDRA: A Web Application for Treebank Search and Visualization. In *Proceedings* *of The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144, Sofia.

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. (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. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

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.

Roark, B. and Bachrach, A. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1(August):324–333. ISBN: 9781932432596.

Rohde, D. (2003). Linger: A flexible platform for language processing experiments.

Safavi, M. S., Husain, S., and Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates : Evidence against the expectation-based account. *Frontiers in Psychology*, pages 1–21.

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Team, R. C. (2018). R: A Language and Environment for Statistical Computing.

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. and Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, pages 767–794. Publisher: JSTOR.

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. and Nicenboim, B. (2016). Statistical methods for linguistic research : Foundational Ideas - Part II. *Language and Linguistics Compass*, pages 1–23. arXiv: ubmit/1447449.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. preprint, PsyArXiv.

von der Malsburg, T. and Angele, B. (2016). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94:119–133. arXiv: 1504.06896.

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.