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, a reader must keep track of dependent relationships between words, such as between a verb and its particle (e.g. *take* the trash *out*). Increasing the distance between such dependent elements may either facilitate or hinder reading of the distant word: the surprisal account of sentence processing predicts that increasing distance makes the position and identity of the distant word more predictable, facilitating its reading. This is known as an antilocality effect and may result from the predictive preactivation of probable upcoming words in memory. In contrast, increasing distance may slow reading via interference, working memory load, and temporal activation decay; this is known as a locality effect. Locality effects induced by interference and working memory have been much studied in long-distance dependency processing; however, the effect of temporal activation decay is more difficult to test in isolation. In one self-paced reading and one eye tracking experiment, we investigated the opposing effects of predictability and temporal activation decay by delaying the appearance of a verb particle that varied in lexical predictability. Crucially, the delay-inducing information contained no information about the identity of the upcoming particle and no new discourse referents, which are a well-studied source of interference and working memory load. Under the assumption that highly predictable particles may be associated with stronger preactivation, we hypothesised that stronger preactivation would make highly predictable particles more resistant to the effects of decay than less predictable particles. The self-paced reading study provided evidence against an effect of temporal decay, predictability, or and interaction. The eye tracking experiment provided evidence that higher predictability sped up early and total reading times, but evidence against an effect of either decay or the interaction of predictability and decay. In sum, delaying the verb particle did not appear to either facilitate or hinder reading times. This finding is consistent with accounts suggesting that reading time speed-ups over distance may be due to additional intervening information that confirms lexical expectations, as well as with accounts suggesting that temporal activation decay may not be a useful predictor of 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, Cambridge, Massachusetts, United States**

Corresponding author:
Kate Stone[1]

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

ABSTRACT

To make sense of a sentence, a reader must keep track of dependent relationships between words, such as between a verb and its particle (e.g. *take* the trash *out*). Increasing the distance between such dependent elements may either facilitate or hinder reading of the distant word: the surprisal account of sentence processing predicts that increasing distance makes the position and identity of the distant word more predictable, facilitating its reading. This is known as an antilocality effect and may result from the predictive preactivation of probable upcoming words in memory. In contrast, increasing distance may slow reading via interference, working memory load, and temporal activation decay; this is known as a locality effect. Locality effects induced by interference and working memory have been much studied in long-distance dependency processing; however, the effect of temporal activation decay is more difficult to test in isolation. In one self-paced reading and one eye tracking experiment, we investigated the opposing effects of predictability and temporal activation decay by delaying the appearance of a verb particle that varied in lexical predictability. Crucially, the delay-inducing information contained no information about the identity of the upcoming particle and no new discourse referents, which are a well-studied source of interference and working memory load. Under the assumption that highly predictable particles may be associated with stronger preactivation, we hypothesised that stronger preactivation would make highly predictable particles more resistant to the effects of decay than less predictable particles. The self-paced reading study provided evidence against an effect of temporal decay, predictability, or and interaction. The eye tracking experiment provided evidence that higher predictability sped up early and total reading times, but evidence against an effect of either decay or the interaction of predictability and decay. In sum, delaying the verb particle did not appear to either facilitate or hinder reading times. This finding is consistent with accounts suggesting that reading time speed-ups over distance may be due to additional intervening information that confirms lexical expectations, as well as with accounts suggesting that temporal activation decay may not be a useful predictor of reading times.

INTRODUCTION

Keeping track of dependent relationships between words in a sentence is a crucial step in understanding meaning. For example, to understand the full meaning of a particle verb such as *take out*, a reader must recognise that these two words form a dependency, even when they are separated by other sentence material, e.g. *take* the trash *out*. Separating dependent words may affect their processing in different ways: The surprisal account proposes that additional information separating dependent words sharpens expectation for the syntactic and even the lexical features of the distant word, speeding up its reading once it is encountered (Levy, 2008; Hale, 2001). This reading speed-up may occur because probable distant words and their features are preactivated in memory in advance of their being seen, facilitating their processing once encountered (Kuperberg and Jaeger, 2016). However, several accounts propose that activation decays over time; under this assumption, decay of preactivated words may mean that reading of the distant word is negatively impacted (Van Dyke and Lewis, 2003; Ferreira and Henderson, 1991;

Gibson, 1998; Lewis and Vasishth, 2005; Vasishth and Lewis, 2006). In this paper, we test the contrasting effects of predictability and temporal activation decay using long-distance verb-particle dependencies in German.

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).

However, while surprisal is able to capture a broad range of observed reading time behaviour, some evidence suggests that surprisal may only be a good predictor of reading times in contexts where upcoming words are highly predictable (Husain et al., 2014; Konieczny, 2000; Levy and Keller, 2013). In low predictability contexts, the limits of working memory capacity may outweigh any facilitation gained from context. For example, in German, it was found that reading times at the clause-final verb of a relative clause were faster when a single dative argument preceded the verb than when more distance was created by adding an extra adjunct (Levy and Keller, 2013). The faster reading times observed in the shorter distance dependency contrast directly with the predictions of surprisal. The authors concluded that the relative infrequency of adding an adjunct (according to a corpus analysis) actually posed a greater working memory load and that surprisal may therefore be a better predictor of reading times only when working memory load is low. The results also hinted at a potential role of verb predictability, as the corpus analysis also suggested the probability of the verb was higher in the shorter dative-only than in the longer dative-plus-adjunct condition. Casting doubt on these results, however, is a replication attempt finding only a reading time slow-down at longer distance, regardless of what information preceded the verb (Vasishth et al., 2018).

More direct tests of the predictability/distance interaction have been 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 contrast, in comparable constructions in Persian, additional distance slowed reading of the distant verb, regardless of its predictability, although 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. The Persian study used a relative clause and a prepositional phrase as the intervener (Safavi et al., 2016). Both relative clauses and prepositional phrases introduce new 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 new 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 instead increase evidence for the position and lexical identity of the upcoming verb (Hale, 2001; Levy, 2008). Together, these results suggest that facilitation in the reading times of a distant word at long distance may only occur when that word is highly predictable from the context.

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.

However, while activation decay may occur, there is evidence to suggest that it is not a useful predictor of reading time (Van Dyke and Johns, 2012; Engelmann et al., 2019; Vasishth et al., 2019), and that longer word recall times and reduced accuracy over time are better explained by interference than decay (Lewandowsky et al., 2009). On the other hand, much of this evidence comes from computational modelling based largely on data from experiments testing interference rather than specifically testing decay. 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). One empirical study has demonstrated the effects of decay over and above that of interference (Van Dyke and Lewis, 2003), although the authors later attribute these results to interference (Van Dyke and Johns, 2012).

Nonetheless, a basic account of temporal activation decay would predict that the longer the distance between two dependent words in a sentence, the greater the activation decay and the slower the reading. A reading slow-down may be even more likely if the lexical identity of the distant word is less predictable. This is in direct contrast to the surprisal account, which predicts that the further away the dependent word, the faster it should be read. Although not explicitly modelled by unlexicalised surprisal (Levy, 2008), a reading speed-up may be even more likely if the lexical identity of the distant word is more predictable (Collins, 2003; Charniak, 2001).

The current experiments

We tested the decay/predictability 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). 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). 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. 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. To induce decay between the verb and its particle, we manipulated distance with a neutral adjectival modifier. Critically, the modifier added no interference or working memory load through the introduction of new discourse referents (Gibson, 1998, 2000; Lewis and Vasishth, 2005), and did not provide semantic clues about the lexical identity of the dependency resolution. Any effects of the intervener on reading time were therefore attributable to temporal decay alone.

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. *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). 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 6 particles, and *large set* verbs that took at least 10 particles. Using a cloze test, we confirmed that each sentence required a particle. 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 at the particle less likely to be affected by decay.

We tested the hypotheses in self-paced reading and eye tracking, both to confirm that any effects seen were not limited to a particular experimental method, but also because the two methods 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 and re-read. 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 underlying slower or faster reading at a particular word and complements observations made in self-paced reading.

Predictions

Our first prediction concerns surprisal. The context of our particle verb sentences generates a strong expectation for a particle, confirmed with a cloze test. Under surprisal, delaying the appearance of the expected particle should mean that reading speed will become faster the longer it is delayed. The reading speed-up should be even greater if the lexical identity of the particle is highly predictable from the context, meaning that there should be an interaction between predictability and distance. We attempted to quantify these predictions by computing surprisal values for the particles; however, despite attempts with 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. 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).

Our second prediction concerns temporal activation decay. To quantify the effect of decay on reading time, we conducted a simulation using the decay parameter of the LV05 model Lewis and Vasishth, 2005. Note that the full LV05 model was not used as it is primarily a model of interference, which we were not testing in the current study. To quantify predictability in the simulation, we assumed a finite pool of spreading activation for all of the plausible particle continuations. Dividing the finite pool of spreading activation among a small set of particles therefore meant a higher starting activation per particle in the small set condition (higher predictability) and vice versa for a large set of particles (lower predictability). Our simulations suggested that decay over distance would make the long distance condition more sensitive to predictability of the particle than the short distance condition. Code for the simulation is included in the R script in the paper's OSF repository, see Appendix 1. 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.

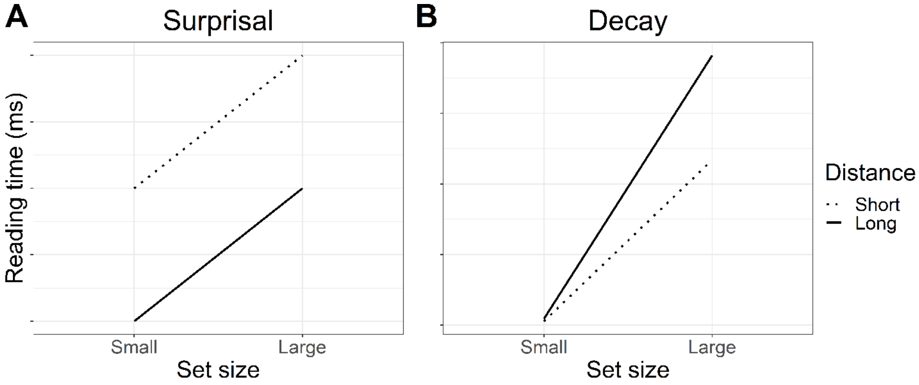

Figure 1. Predicted interaction of lexical predictability and distance. A. Informal predictions of the surprisal account. **B.** Predictions based on a simulation using the decay parameter of the LV05 model.

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 reading or language production 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 for this particular study.

Materials

The study had a 2×2 design with *set size* (small vs. large) and *distance* (short vs. long) as factors. 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. Each experimental item was a quartet of four sentences in which the context required a particle for the sentence to be grammatical. In the example experimental item below, the bolded 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 modifier was introduced between the verb and its particle (underlined). Crucially, the adjectival modifier did not introduce any new discourse referents or other features that could interfere with the particle's retrieval (Gibson, 1998, 2000; Lewis and Vasishth, 2005). 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 shifted to appear before the verb.

Example item:

a) **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.*

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

d) **Large set/long distance:**
Mit dem Lappen **spülte** sie die neu gekauften Teller in der Küche **ab**, um Platz zum Kochen zu schaffen.
*With the rag, she **rinsed** the newly bought plates in the kitchen **off** to create 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

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. 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. German native speakers provided completions for the truncated sentences in a paper-and-pencil cloze test (N = 126, 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 target sentences were randomly interspersed with 63 filler sentences, giving a total of 111 sentences per cloze test. Participants were instructed to complete each truncated sentence with the word or words that first came to mind.

The results of the cloze test yielded 24 items that achieved the required experimental manipulation; that is, a particle was always elicited and more particles were elicited in the large than in the small set condition. 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.

Condition	Cloze probability Mean	 95% CI	Entropy Mean	 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. **Cloze statistics for the final set of 24 items.**

Wherever possible, however, the highest cloze particle was used. Means and 95% confidence intervals of Beta distributions corresponding to the cloze probabilities for each factor level are presented in Table 1.

Cloze probabilities provided a measure of how predictable the target particles in each condition were. To determine whether the cloze probability of the particle differed between small and large set conditions, a logistic mixed model was fit in *brms* (Bürkner, 2017) in R (R Core Team, 2018) 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. Regularising 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 found in the code provided, see Appendix 1. 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 *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:

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

To determine whether uncertainty (and thus entropy) was higher in the large set condition, a lognormal regression model was fitted to the entropy values with the same contrast coding as for the cloze probability analysis. The *brms* hurdle lognormal family was used for the likelihood function to account for zeros in the data. Regularising 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.

This analysis raised an immediate problem with the experimental design. The categorical predictor *set size* used in the planned analysis was intended as a proxy for entropy and predictability, 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 licensed by each base verb, the results of the cloze test suggested they did not represent the range of particle completions provided by readers at the particle site. This can be seen in Figure 3: although the *average* entropy was higher in the large set than in the small set condition, both conditions contained high and low entropy sentences. We therefore present an analysis of entropy as a continuous predictor instead, since this maps better to our planned manipulation

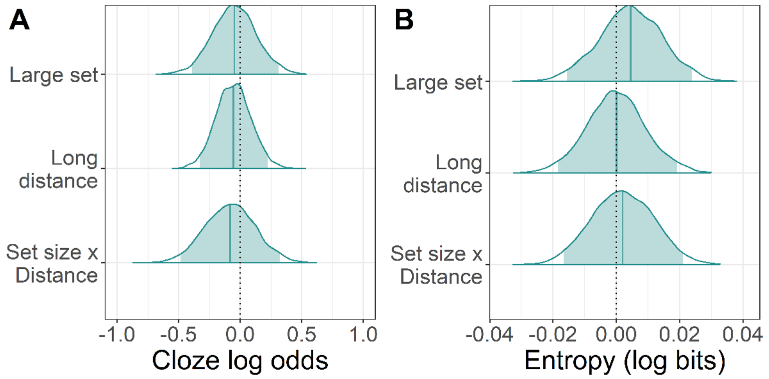

Figure 2. Change in cloze log odds and entropy of the target particle associated with each predictor. A. The posterior distributions for the effect of large set size and long distance on cloze probability relative to the grand mean of each condition (the dotted line). The posteriors for the small set 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. **B.** Posteriors for the effect of large set size and long distance on entropy.

of predictability (high entropy = low predictability and vice versa). For transparency, we present both the planned "categorical" analysis and the exploratory "continuous" analysis.

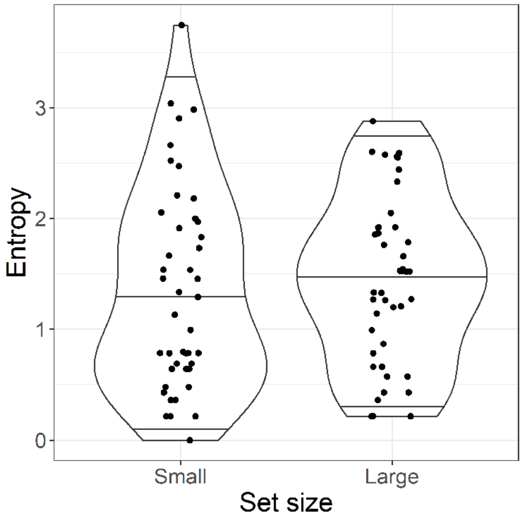

Figure 3. By-item entropy within small and large set categories. Violin plots show the median and 95% quantiles.

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 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 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.

Data 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* (Bürkner, 2017) in R (R Core Team, 2018). Reading times of less than 100 ms were excluded. The dependent variable was reading time at the particle with a 1000/y reciprocal transform as suggested by the Box Cox procedure (Box and Cox, 1964). 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 4. 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:

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

The prior distribution of the intercept was determined using domain knowledge that mean reading time is approximately 3 words per second 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. We assumed that the expected effect of set size would most likely be to either increase or decrease reading speed by, at most, 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 an η of 2 or more, the LKJ prior represents a distribution ranging from -1 to $+1$, but favours 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 code accompanying the article, see Appendix 1.

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

Question response accuracy and reaction times

Mean accuracy and reaction times to responses to comprehension questions in all four conditions are set out in Table 2.

Planned analysis

Set size as a categorical predictor

Mean self-paced reading speed by condition are shown in Table 3 and the model estimates in Table 4. The 95% credible intervals of each of the posteriors contain zero, suggesting that there was uncertainty

Condition	Accuracy (%)		Reaction time (ms)	
	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 2. **Summary of question response accuracy and reaction times for comprehension questions in the self-paced reading experiment.**

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.

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 3. **Mean self-paced reading speed by condition.**

Predictor	$\hat{\beta}$ (words/sec)	95% CrI	BF_{10}: 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 4. **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.

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. Descriptive statistics for reading times in each distance condition are shown in Table 5. Mean reading times according to entropy have been split into high and low categories by median-split for summary purposes, but entropy was used as a continuous predictor in the statistical model.

Mean reading times across the whole sentence for both experiments are plotted in Figure 4. 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*.

Condition	Mean reading time (ms)	95% CrI
(a) Low entropy, short distance	443	$420, 466$
(b) Low entropy, long distance	438	$416, 461$
(c) High entropy, short distance	433	$413, 455$
(d) High entropy, long distance	443	$422, 466$

Table 5. Mean self-paced reading speed by condition. For the purpose of these summary statistics only, the continuous entropy predictor was sorted into high and low categories via median-split.

The priors and model specification remained the same as for the planned analysis. The model coefficients are summarised in Table 6. As can also be seen in Figure 5, 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 6. 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.

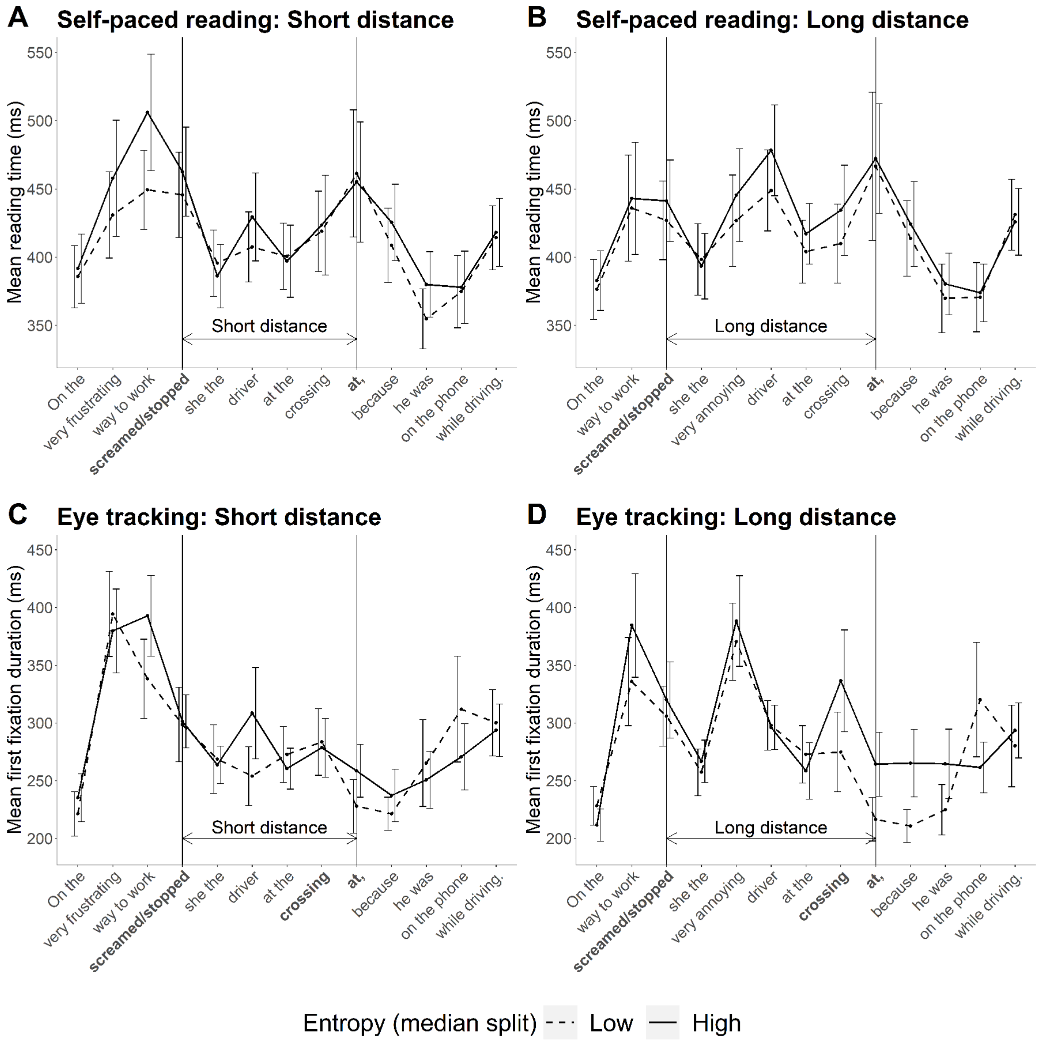

Figure 4. **Mean reading times across the sentence. A-B.** Mean reading times observed in the self-paced reading experiment. Error bars show 95% confidence intervals. **C-D.** Mean total fixation times observed in the eye tracking experiment.

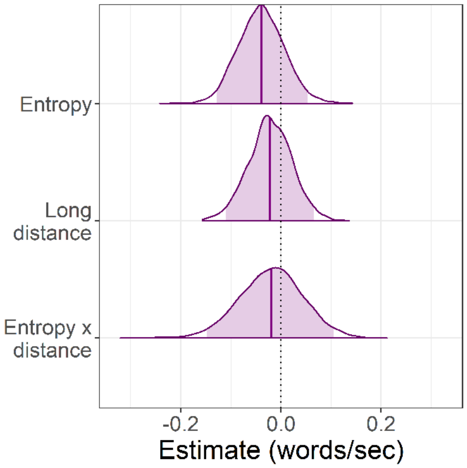

Figure 5. **Change in self-paced reading speed at the particle with entropy as a continuous predictor.** Now that entropy is a continuous predictor, the posterior represents the change in reading time elicited by a 1-unit increase in entropy. Due to the reciprocal transform, a shift in the posterior 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.

Reading speed predicted by the model is plotted in Figure 6. The numerical pattern suggests an interesting mix of the two hypotheses; that is, when predictability was high (low entropy), reading speed was faster at long distance in line with the surprisal account. In contrast, when predictability was low (high entropy), the pattern more closely resembles that predicted by decay. However, these patterns are not further interpreted as the outcome of the statistical analysis did not support an interaction.

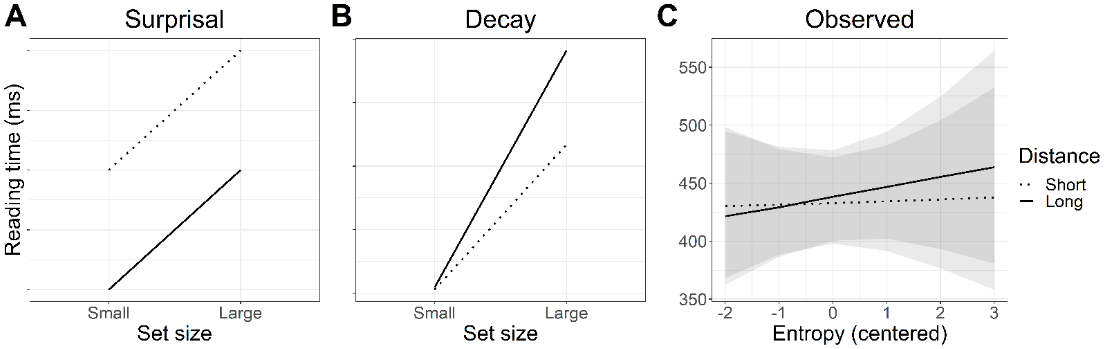

Figure 6. Predicted versus modelled self-paced reading times. A-B. Predicted interaction. **C.** Observed self-paced reading time pattern. Shaded areas indicate 95% confidence intervals.

Interim discussion

Neither the planned nor the exploratory analyses were consistent with the predictions in Figure 6. With respect to the planned analysis, one potential explanation may lie in the very small differences in cloze probability 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 rest of the sentence, while ours were based on lexical entropy at the particle. Nonetheless, while the small between-category difference in entropy may explain why we did not see a statistical difference between large and small set particle reading times, it does not explain why we still saw no difference when entropy was used as a continuous predictor. We turn now 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. Predictability has been shown to affect reading times in both early and total eye tracking measures (Staub, 2015; Rayner, 1998) and the revision of disconfirmed expectations, a higher rate of regressions (Clifton et al., 2007; Frazier and Rayner, 1987). Revision of disconfirmed expectations should occur more frequently when predictability is low and the probability of pre-integrating the “wrong” particle increases; we therefore analysed early and total reading times, as well as a measure of regression time. For each of these measures, we 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 reading or language production disorders, 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). Trials containing blinks or track loss were excluded. Linear mixed-effects models with full variance-covariance matrices estimated for the random effects of participant and item were fitted using *brms* (Bürkner, 2017) in R (R Core 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. As for Experiment 1, the spillover region was not analysed, but mean reading times across the whole sentence are presented in Figure 4. The preceding rather than the following word was chosen because the target particle was at the right clause boundary. The dependent variables were FFD, FPRT, TFT, and RPD 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 accompanying code, see Appendix 1.

RESULTS

Question response accuracy and reaction times

Mean response accuracy and reaction times for the comprehension questions in all four conditions are set out in Table 7.

Condition	Accuracy (%) Mean	Accuracy (%) 95% CI	Reaction time (ms) Mean	Reaction time (ms) 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 7. Summary of question response accuracy and reaction times in the eye tracking experiment.

Planned analysis

Set size as a categorical predictor

Observed reading times per condition are summarised in Table 8. The model estimates for each reading time measure are shown in Table 9. 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. A possible limitation of our approach using Bayes factor analyses is that we are evaluating multiple measures, without any correction for family-wise error (von der Malsburg and Angele, 2016). 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.

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 8. 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. Descriptive statistics for each reading time measure are shown in Table 10. Mean reading times according to entropy have been split into high and low categories by median-split for summary purposes, but entropy was used as a continuous predictor in the statistical model.

The model estimates can be seen in Table 11 and the model posteriors in Figure 7. The Bayes factor

					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 9. 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.

(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, regardless of which prior was used.

Measure	Condition	Mean reading time (ms)	95% CrI
FFD	(a) Low entropy, short distance	279	265, 295
	(b) Low entropy, long distance	264	250, 279
	(c) High entropy, short distance	293	277, 311
	(d) High entropy, long distance	317	299, 335
FPRT	(a) Low entropy, short distance	317	297, 338
	(b) Low entropy, long distance	287	270, 306
	(c) High entropy, short distance	321	300, 343
	(d) High entropy, long distance	357	334, 381
TFT	(a) Low entropy, short distance	357	332, 385
	(b) Low entropy, long distance	321	299, 346
	(c) High entropy, short distance	376	348, 407
	(d) High entropy, long distance	416	385, 449
RPD	(a) Low entropy, short distance	354	329, 382
	(b) Low entropy, long distance	325	301, 351
	(c) High entropy, short distance	358	332, 386
	(d) High entropy, long distance	402	373, 433

Table 10. Mean eye-tracking reading times by condition for the exploratory analysis. For the purpose of these summary statistics only, the continuous entropy predictor was sorted into high and low categories via median-split.

					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 11. 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.

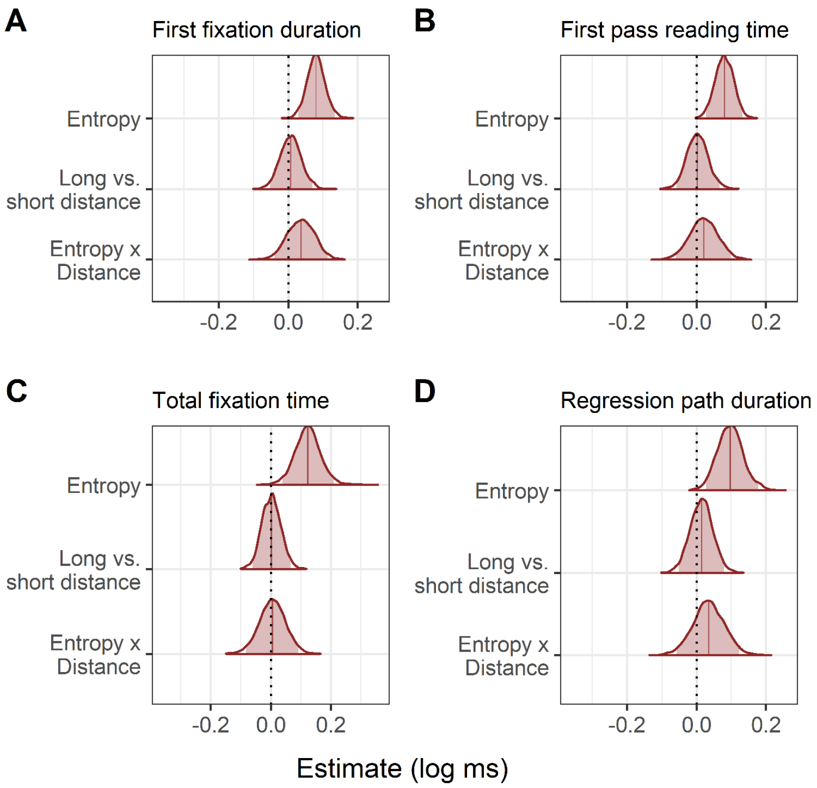

Figure 7. Changes in reading time for each eye-tracking measure using entropy as a continuous predictor. Now that entropy is a continuous predictor, the posterior represents the change in reading time for the average reader elicited by a 1-unit increase in entropy. The log transformed reading times mean that posteriors shifted to the right of zero indicate slower reading. Error bars show the 95% credible intervals.

The predicted versus observed interactions of distance and entropy are plotted in Figure 8. Numerically, the pattern of reading times again appeared to be a mixture of the surprisal and LV05 predictions. However, the results of the statistical analyses did not support an interaction of entropy and distance, and so this pattern is not further interpreted.

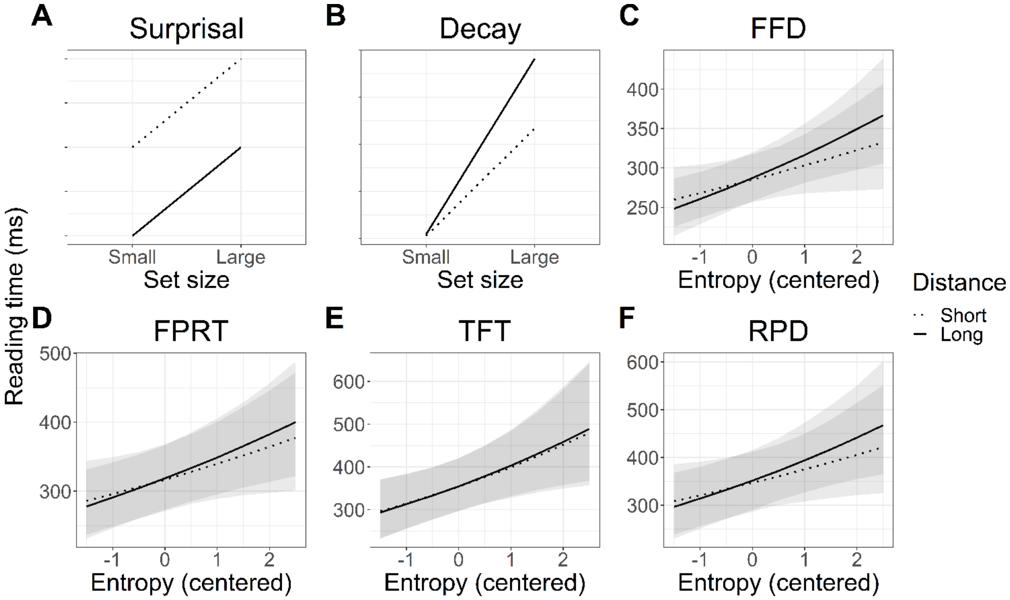

Figure 8. Predicted versus modelled interaction of entropy and distance on reading times in each eye tracking measure. A-B. Predicted interaction. **C-F.** Observed reading time patterns. Shaded areas represent 95% confidence intervals.

Interim discussion

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 evidence that temporal decay did not influence reading times, either alone or in interaction with entropy.

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 temporal activation decay. 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 decay.

Predictability

The findings in the eye tracking data are 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). Our results may appear inconsistent with this proposal in 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 (Vasishth et al., 2013). The effects of syntactic surprisal have elsewhere also been found in both early and late measures (Boston et al., 2008).

On the other hand, one possible mechanism by which effect of lexical surprisal may affect regression times is via the reanalysis of a mispredicted particle in high entropy (low predictability) sentences, rather than faster early lexical access in low entropy (high predictability) sentences (Clifton et al., 2007; Frazier and Rayner, 1987). However, if it were the case that the regression path slow-down was driven by a predictability effect, then the same slow-down should have been observed in self-paced reading times; this was not the case. 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. Therefore, self-paced reading times should arguably resemble regression path duration or total fixation times more than earlier measures such as first fixation duration. Instead, 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 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.

Another likely explanation for not having detected a decay effect is that the difficulty in creating experimental items meant there were only 24 experimental items in total. In the Latin square design, this meant that each participant saw only six target trials per condition. If the effect of decay is indeed very small, more trials per participant may be necessary in order to detect the effect.

Particle preactivation at the verb

In spite of the evidence against 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. One possible conclusion from our results is 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 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.

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 possible cancelling out of the expected frequency effect at the verb could be further evidence against a non-preactivation account, although this hypothesis requires testing. Such a test could be to hold the verb and particle constant, and manipulate other regions of the sentence. 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

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); in contrast, temporal activation decay should mean that delaying the particle would result in a locality effect, stronger in low vs. high predictable particles. Contrary to both hypotheses, we found no evidence that delaying the particle 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. Crucially, the particle in the current study was delayed with information that neither hinted at the upcoming particle's identity, nor increased interference or working memory load. The evidence against an effect of delaying the particle therefore suggests that the surprisal-based antilocality effects observed in previous research may be due to the additional intervening information that confirms lexical expectations, and that temporal activation decay is not a strong influence on reading times.

Appendix 1

Data and code

All data and code necessary to reproduce our analyses are available here: https://osf.io/yg5wx/

Appendix 2

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 A1, while the frequencies of the verb+particle constructions were comparable, frequency of the base verb was notably higher in the high entropy condition.

	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 A1. **Mean verb and particle verb frequency per 1000 words for high and low entropy.** Sentences were divided into high and low entropy categories via a median split.

ACKNOWLEDGMENTS

We thank Genevieve McArthur, Stefan Frank, Aya Meltzer-Asscher, and Michael Hahn for their detailed and thoughtful reviews of the manuscript, which greatly improved the final paper. We also wish to acknowledge the valuable feedback from reviewers and colleagues at AMLaP 2017 and CUNY 2018.

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.

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

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.

Clifton, C., Staub, A., and Rayner, K. (2007). Chapter 15 - Eye movements in reading words and sentences. In Van Gompel, R. P. G., Fischer, M. H., Murray, W. S., and Hill, R. L., editors, *Eye Movements*, pages 341–371. Elsevier, Oxford.

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.

Frazier, L. and Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26(5):505–526.

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.

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.

Lee, M. 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. (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.

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

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.

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.

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.

Van Dyke, J. A. and Johns, C. L. (2012). Memory Interference as a Determinant of Language Comprehension. *Language and Linguistics Compass*, 6(4):193–211. arXiv: NIHMS150003 ISBN: 0009-2665.

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., Malsburg, T. v. d., and Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *WIREs Cognitive Science*, 4(2):125–134. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1209.

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*.

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.