Predicting long distance lexical content in German verb-particle constructions

Doctoral Thesis submitted to the Faculty of Human Sciences at the University of Potsdam in

partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

by

Kate Stone

Defended: 18 August 2020

Supervisors:
Prof. Dr. Shravan Vasishth
Dr. Franklin Chang

Reviewers:
Prof. Dr. Shravan Vasishth
Prof. Dr. Stefan Frank

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-47679
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-476798

Contents

List of Figures

List of Tables

Erklärung

Hiermit erkläre ich, dass ich bei der Abfassung der vorliegenden Arbeit alle Regelungen guter wissenschaftlicher Standards eingehalten habe. Weiter erkläre ich, dass ich die voliegende Arbeit selgständig verfasst habe und über die Beiträge meiner Koautoren hinaus, welche in der beiliegenden Erklärung über die Beiträge zu Gemeinschaftsveröffentlichungen spezifiert sind, keine Hilfe Dritter in Anspruch genommen habe.

Kate Stone
Potsdam, den 15. Dezember 2019

Abstract

A large body of research now supports the presence of both syntactic and lexical predictions in sentence processing. Lexical predictions, in particular, are considered to indicate a deep level of predictive processing that extends past the structural features of a necessary word (e.g. noun), right down to the phonological features of the lexical identity of a specific word (e.g. /kite/; DeLong, Urbach, & Kutas, 2005). However, evidence for lexical predictions typically focuses on predictions in very local environments, such as the adjacent word or words (DeLong et al., 2005; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004). Predictions in such local environments may be indistinguishable from lexical priming, which is transient and uncontrolled, and as such may prime lexical items that are not compatible with the context (e.g. Kukona, Cho, Magnuson, & Tabor, 2014). Predictive processing has been argued to be a controlled process, with top-down information guiding preactivation of plausible upcoming lexical items (Kuperberg, 2016). One way to distinguish lexical priming from prediction is to demonstrate that preactivated lexical content can be maintained over longer distances.

In this dissertation, separable German particle verbs are used to demonstrate that preactivation of lexical items can be maintained over multi-word distances. A self-paced reading time and an eye tracking experiment provide some support for the idea that particle preactivation triggered by a verb and its context can be observed by holding the sentence context constant and manipulating the predictabilty of the particle. Although evidence of an effect of particle predictability was only seen in eye tracking, this is consistent with previous evidence suggesting that predictive processing facilitates only some eye tracking measures to which the self-paced reading modality may not be sensitive (Rayner, 1998; Staub, 2015). Interestingly, manipulating the distance between the verb and the particle did not affect reading times, suggesting that the surprisal-predicted faster reading times at long distance may only occur when the additional distance is created by information that adds information about the lexical identity of a distant element (Grodner & Gibson, 2005; Levy, 2008). Furthermore, the results provide support for models proposing that temporal decay is not major influence on word processing (Lewandowsky, Oberauer, & Brown, 2009; Vasishth, Nicenboim, Engelmann, & Burchert, 2019).

In the third and fourth experiments, event-related potentials were used as a method for detecting specific lexical predictions. In the initial ERP experiment, we found some support for the presence of lexical predictions when the sentence context constrained the number of plausible particles to a single particle. This was suggested by a frontal post-N400 positivity (PNP) that was elicited when a lexical prediction had been violated, but not to violations

when more than one particle had been plausible. The results of this study were highly consistent with previous research suggesting that the PNP might be a much sought-after ERP marker of prediction failure (DeLong, Troyer, & Kutas, 2014b; Delong, Urbach, Groppe, & Kutas, 2011; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). However, a second experiment in a larger sample experiment failed to replicate the effect, but did suggest the relationship of the PNP to predictive processing may not yet be fully understood. Evidence for long-distance lexical predictions was inconclusive.

The conclusion drawn from the four experiments is that preactivation of the lexical entries of plausible upcoming particles did occur and was maintained over long distances. The facilitatory effect of this preactivation at the particle site therefore did not appear to be the result of transient lexical priming. However, the question of whether this preactivation can also lead to lexical predictions of a specific particle remains unanswered. Of particular interest to future research on predictive processing is further characterisation of the PNP. Implications for models of sentence processing may be the inclusion of long-distance lexical predictions, or the possibility that preactivation of lexical material can facilitate reading times and ERP amplitude without commitment to a specific lexical item.

Acknowledgements

I would first like to thank Shravan Vasishth for not only giving me the opportunity to work in his lab, but also for working hard to create a group that fosters rigorous science, and a healthy and inclusive working environment. In doing this, he has created a team of scientists who take joy in discussing ideas, doing good science, and improving each other's work. This is no small feat and it has been a privilege to be a part of the group and to learn from it the value of slow, methodical, thoughtful science. To me, the most important member of this group has been my co-advisor Titus von der Malsburg, whose enthusiasm for and dedication to psycholinguistics encompasses not just the science itself, but also the improvement and health of the field, and the teaching of the next generation. I am lucky to have benefited personally from his guidance. Thanks must also go to Franklin Chang for kindly agreeing to co-supervision from afar, even if things didn't work out as we planned.

In addition to my main advisors, this research was a team effort with input both intellectual and emotional from the lab members I've had the enormous fortune to work alongside over the last few years, in particular my 'Doktorschwestern' Serine Avetisyan, Anna Laurinavichyute, and Daniela Mertzen; my unofficial mentor and official collaborator Sol Lago; past and present lab members Felix Engelmann, Lena Jäger, Bruno Nicenboim, and Dario Paape; and more recently Paula Lissón, Dorothea Pregla, Daniel Schad, and Garrett Smith. I would also like to thank Harald Clahsen and Anna Jessen for getting me here in the first place.

I am greatly indebted to our lab manager Johanna Thieke and her assistants Chiara Tschirner, Romy Leue, Elna Haffner, Marie de la Fuente, Alexandra Lorson, and others, for the many hours spent painstakingly developing experimental stimuli and collecting data, and for generally being so patient with my attempts at German. Thanks must also go to Fynn Dobler, Franziska Hauer, and Robin Schäfer for reviewing experimental stimuli at various stages of development.

I am extremely grateful to the Potsdam Graduate School for providing me with a scholarship that fully supported my studies, and particularly for their flexibility in accommodating unexpected challenges and opportunities. Major thanks must also go to Annett Eßlinger and Michaela Schmitz, without whom the Linguistics Department would simply not function.

Finally, and most importantly, none of this would have been possible without the unwavering support of Warwick, Katja, Heike, Tamie, Clare, and Athalia, who are the reason I keep going.

Chapter 1

1 Introduction

Reading is a cognitively demanding activity. The human sentence parser must simultaneously draw on multiple sources of knowledge to comprehend incoming input while already planning and executing eye movements. To minimise the demand of reading on working memory, various strategies are employed. These include pausing, re-reading, and, the topic of this thesis, predictive processing. By predicting what information may appear next, the parser can begin accessing this information in memory, thereby reducing workload once the predicted information is seen. Defining the mechanisms that permit predictive processing in language comprehension is important for determining how the features of words are interconnected in the brain, and how we propagate activation through these connections to understand language.

One of the outstanding questions in the domain of predictive processing is how to convincingly establish that it is a controlled, strategic process, as distinct from uncontrolled, automatic spreading activation. One way to demonstrate this distinction is to show that the effects of predictive processing are sustained over longer distances than those of automatic spreading activation (priming), which should be transient and non-strategic (Huettig, 2015; Kuperberg & Jaeger, 2016). In this thesis, I attempt to show that predictions for specific lexical items (words) can be generated and maintained across a sentence. While evidence for long distance predictions is not necessarily evidence against lexical priming, the maintenance of lexical predictions over distance, in the face of other known processing limitations such as activation decay, interference, and resource limitations, strengthens the argument for a controlled process.

1.1 A definition of predictive processing

There are various terms employed to describe predictive processing in language processing, including forward inference (McKoon & Ratcliff, 1992; Singer & Ferreira, 1983), anticipation (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003), expectation (Hale, 2001; Levy, 2008), commitment (W.-Y. Chow, Lau, Wang, & Phillips, 2018; Lau, Holcomb, & Kuperberg, 2013), and preactivation (Kuperberg & Jaeger, 2016). Throughout this thesis, I will use the term prediction to refer to situations where a commitment to a specific syntactic or lexical item has occurred. I will use *preactivation* to refer to situations where a number of syntactic or lexical items have been strategically activated, but where a single item has not been committed to. Kuperberg & Jaeger (2016) defines preactivation as any kind of strategic activation of events, words, structure, morphological, phonological, or other units of language representation that cannot be explained by priming alone. Note that the term *strategic* does not mean that the preactivation is conscious, only that some process of control (e.g. inhibition or suppression) results in certain representations being preactivated to the exclusion of others.

1.2 Structure of the thesis

The key question addressed by this thesis is whether lexical predictions can be generated over long distances. The purpose of addressing this question is to contribute evidence distinguishing predictive processing from priming. In the current chapter, Chapter 1, I review the literature on predictive processing, establishing that conclusive evidence distinguishing prediction from priming does not yet exist, and describing current knowledge about the cognitive mechanisms subserving predictive processing.

Chapters 2 and 3 provide a short introduction to and discussion of German verb-particle constructions, e.g. *Das Fest* ***ging los*** (the party started), and the empirical methods used in the subsequent experimental chapters of the thesis.

Chapter 4 presents two reading time experiments using particle verbs. Distance between the verb and particle was manipulated, as well as predictability of the particle's lexical identity. Reading times were measured at the particle. Crucially, the experimental sentences in all conditions were identical except for the verb, meaning that any changes in reading time observed at the particle could only be attributed to differences in processing that occurred well before the particle's appearance. In early and total eye tracking measures, increased predictability was associated with reading times, suggesting that both the structure of the verb-particle dependency had been preactivated, as well its probable lexical content. Delaying the particle did not affect reading times in either experiment. Overall, the results suggest that structural and lexical content can be preactivated and maintained over distance to facilitate processing. These results have been submitted to the open-access journal, *PeerJ*.

Having established that long-distance structural and lexical content may be preactivated in verb-particle constructions, Chapter 5 presents a test of whether preactivation can crystallise into commitment to a *specific* verb particle (a lexical prediction). The results of this event-related potential (ERP) experiment suggested that specific particles were indeed being predicted, but that even a small amount of uncertainty about the particle's lexical identity may have discouraged commitment. While these results were consistent with previous research, statistical evidence was inconclusive. A larger sample replication attempt is therefore presented in Chapter 6. The results of the replication attempt contradicted the main findings of the first experiment and were inconclusive about the presence of specific lexical predictions, but did offer new insights into an ERP component recently touted as an index of the cost of prediction failure, the post-N400 positivity. The ERP experiment and its replication are currently in preparation for submission to *Neurobiology of Language*, a new MIT Press open access journal.

Chapter 7 places the experimental findings in the larger context of prediction research and discusses implications for future research and current models of sentence processing.

1.3 History and central debates in prediction research

The influence of prior knowledge and current context on the production of sentences has long been recognised (Goldman-Eisler, 1958; Howes & Osgood, 1954; Taylor, 1953). Early studies of context in linguistic processing examined the facilitation effect on lexical judgements provided by semantically related versus unrelated word pairs (D. E. Meyer & Schvaneveldt, 1971). Others investigated the effect of context on the generation of inferences, both backward inferences about the event already described and forward or predictive inferences about the events that could occur next (McKoon & Ratcliff, 1992; Singer & Ferreira, 1983). Inferences were defined as the derivation of information, either simple or complex, that was not explicitly presented in a text (McKoon & Ratcliff, 1992). The purpose of inferencing was to establish connections between referential elements in the immediate neighbourhood of a text and relied upon easily accessible cues from the context of the sentence (McKoon & Ratcliff, 1992).

In some of the first experiments to test the effects of context on eye movements, it was found that words congruent with their context appeared to require less processing effort than words that were not. Words predictable in a given context were read faster and were less likely to be fixated than unexpected words (Ehrlich & Rayner, 1981). In some cases, contexts were so strong that readers reported having seen a word despite it not having appeared (Ehrlich & Rayner, 1981). That highly predictable words are read faster and skipped more often has been a consistent finding in subsequent research (Farmer, Yan, Bicknell, & Tanenhaus, 2015; Husain, Vasishth, & Srinivasan, 2014; Kliegl, Nuthmann, & Engbert, 2006; Konieczny, 2000; Matsuki et al., 2011; N. J. Smith & Levy, 2013; Staub, 2015; Van Berkum et al., 2005).

The relationship between predictability and reading speed is not straightforward, however. This is especially true when distance is added between words in dependent relationships, such as between a subject and its verb. Over longer distances, predictable words may be read slower or faster, depending on the type of information that separates them (Husain et al., 2014; Levy & Keller, 2013; Safavi, Husain, & Vasishth, 2016). The frequency of syntactic and lexical co-occurrence of words can also determine reading times (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; McDonald & Shillcock, 2003; although see Staub, 2015), as can working memory capacity (Just & Carpenter, 1992; Nicenboim, Logacev, Gattei, & Vasishth, 2015; Waters & Caplan, 1996) and pre-processing of the next word in the parafovea (McConkie & Rayner, 1975; Rayner, 2014). Recent evidence suggests that the reading time effects produced by a word's predictability may only be present when the expected upcoming word is visible in the parafovea (Staub & Goddard, 2019). When the target word in the parafovea is not consistent with expectations, the facilitatory effect of predictability disappears, even when the word in the parafovea is switched to the expected word once fixated, using the boundary paradigm (Staub & Goddard, 2019). This suggests that predictability may facilitate processing only at the very early stages of word recognition. Disentangling the time-course of preactivation using eye-movements is difficult, however, since any underlying process can only be inferred from fixation times, skipping behaviour, and regressions. To address this issue, researchers of predictive processing also used the more temporally sensitive methodology of event-related brain potentials (ERPs).

The N400 is an ERP component whose amplitude is sensitive to the congruency of a word in its context; the less congruent a word is, the larger the N400 (Kutas & Hillyard, 1980b, 1984). The N400 has been reliably linked with a stimulus' meaning in a strictly linguistic context (Kutas & Federmeier, 2011). It is not elicited by surprising stimuli in other domains such as music (Besson & Macar, 1987). The linguistic context may be as simple as word pairs, not necessarily a fully formed sentence (Kutas & Federmeier, 2011). Sentential context does have an effect on the N400, however, such that it becomes smaller for each new word in a sentence, so long as that word is congruent with the preceding context (Payne, Lee, & Federmeier, 2015). However, the *strength* of a context does not influence the N400's amplitude (Van Petten & Luka, 2012). Reduced amplitude of the N400 has therefore been taken as a further indication that predictable words require less processing. Importantly, however, just because a word is congruent does not mean that it is more predictable, and thus the N400 should not be interpreted an index of predictability (Kutas & Federmeier, 2011). A more detailed discussion of the N400 follows in Chapter 2, *Methods in Prediction Research.* In sum, while reduced N400 amplitudes and faster reading times do suggest that context facilitates processing, they do not conclusively prove that words are being predicted. An alternative interpretation is that context facilitates the *integration* of congruent words.

1.3.1 Prediction or integration?

The first task in demonstrating predictive processing was to show that the facilitation seen at predictable words did not simply reflect the ease of integrating that word into the existing sentence representation. Existing integration accounts explained these facilitative effects as words being easier to integrate the more consistent they were with the preceding context (Gernsbacher, 1991). Similarly, a larger N400 ERP component at an unexpected word was argued to reflect the difficulty of integrating the unexpected word into the preceding context, rather than lower preactivation of that word (Brown & Hagoort, 1993; Hagoort, Baggio, & Willems, 2009; Van Berkum, Hagoort, & Brown, 1999). This inspired a range of attempts to demonstrate the facilitative effect of a predicted word *before the predicted word had been seen* (Ferreira & Chantavarin, 2018).

A series of studies demonstrated that an unseen but highly predictable word could affect processing already at the previous word. Eye movements were demonstrated to be directed towards more likely pictures before the corresponding word had been heard (Altmann & Kamide, 1999; Kamide et al., 2003). The weakness of the visual world paradigm, however, was that it offered only a limited set of targets to which eye movements could be directed. For example, upon hearing the sentence *The boy will eat the...*, subjects' eyes moved toward *cake* before it was heard. However, this may be explained by the fact that *cake* was the only visible target that was a plausible object of *eat* (Altmann & Kamide, 1999, 2007; Kamide et al., 2003).

A series of N400 studies then demonstrated that violating the constraints of a predicted but unseen word elicited a larger N400, despite the "violation" being congruous with the preceding sentence. The classic example is of the sentence (DeLong et al., 2005):

(1) The day was breezy so the boy went outside to fly...

In this sentence, *a kite* is the most expected continuation as indicated by a fill-the-gap test. The next word should therefore be *a*. Presenting readers with *an* instead should have been surprising if they had already predicted the form *kite*, even though it is perfectly licensed by the sentence (perhaps the boy went outside to fly *an aeroplane*). A larger N400 at the unexpected determiner *an* was taken to demonstrate that prediction of *kite's* full word form, right down to its phonological form, had occurred (DeLong et al., 2005). Studies in Spanish, Dutch, and Polish have observed similar effects at gender markings that violate the gender of the most expected upcoming word (Szewczyk & Schriefers, 2013; Van Berkum et al., 2005; Wicha et al., 2004). However, the gender marking manipulations have yielded inconsistent results. A larger P600, but no N400, was observed at Spanish articles whose gender declination did not match a high-cloze noun (Wicha et al., 2004). In contrast, only a small, early, positive deflection at Dutch adjectives was seen if their gender declination was inconsistent with the gender of high-cloze noun (Van Berkum et al., 2005). Attempts to replicate these studies, either directly or conceptually, have also proven unsuccessful (Ito, Martin, & Nieuwland, 2016; Kochari & Flecken, 2019; Nieuwland et al., 2018). However, a more recent meta-analysis of open access data from these replication studies has demonstrated that a small difference in the N400, larger for unexpected determiner marking, is always present (Nicenboim, Vasishth, & Rösler, 2019). The effect was too small ever to reach statistical significance in a single study, but was always positive. An effect with a consistently positive sign would not be expected if the true effect was distributed around zero; that is, if there were truly no effect, the sign would sometimes be negative. The inconsistency of findings with respect to eliciting an N400 or a later positivity (or both) may relate to research on N400 blindness in thematic role reversals, finding that time may play a role in generating the N400, as discussed in Chapter 2 (W.-Y. Chow et al., 2018; A. Kim & Osterhout, 2005). On the whole, however, this line of research has pointed to a facilitative effect of as-yet-unseen words which is difficult to reconcile with an integration account. This is not to say that integration is not an important factor in sentence processing, only that it does not account for such findings.

1.3.2 Prediction or priming?

The second and arguably more pervasive argument with regards to predictive processing is that it is difficult to distinguish from priming. Priming is generally considered to be automatic, non-targeted, short-lasting, and involuntary; an initial stage of activation spread that occurs before higher levels of processing kick in (Huettig, 2015; Kuperberg & Jaeger, 2016). In contrast, predictive processing assumes that feedback from higher levels of processing facilitates the processing of upcoming words by pre-emptively increasing their activation at various levels of representation (Kuperberg & Jaeger, 2016; Lau et al., 2013). This controlled preactivation via feedback from upper levels does not necessarily entail conscious control, but rather reflects some mechanism of selection or commitment to a particular word or structure over others (Lau et al., 2013). Some evidence suggests that competition between lexical items may play a role in controlling which words or features 'win' the preactivation race (Staub, Grant, Astheimer, & Cohen, 2015), while others have tried to demonstrate the role of suppression at early stages of processing (Gaston, Lau, & Phillips, 2019).

Another approach to distinguishing prediction from priming has been to show that when the same words appear in different contexts, it is the higher level context that determines the correct interpretation. Anomalous words in low-context sentences containing the same key words as a high-context sentence have been shown to elicit a larger positive shift ERP amplitude (M. Otten & Van Berkum, 2008). For example, the context-incongruent word *stove* presented after 2a below elicited a post-N400 positivity (PNP), because the context pointed to a particular word. When the same key words were used in 2b, whose context did not point to any particular word, the PNP was not seen:

(2) a. Sylvie and Joanna really feel like dancing and flirting tonight. Therefore they go to a...

b. After all the dancing, Joanna and Sylvie really don't feel like flirting tonight. Therefore they go to a...

This was taken to suggest that the meaning of the sentence had generated a prediction in the high-constraint sentence which was not generated by simple word-based priming in the low-context sentence. While this does suggest that word-based priming was not a predictor of ERP amplitude at the anomalous word, this does not rule out other sources of priming. Priming can occur at multiple levels of representation, from morphological/phonological (Frost, Deutsch, Gilboa, Tannenbaum, & Marslen-Wilson, 2000), to syntactic (Bock, 1986b), lexical (Bock, 1986a), and possibly even general knowledge (Chwilla & Kolk, 2005; von der Malsburg, Poppels, & Levy, 2019). A second approach to distinguishing prediction from priming has been to demonstrate that prediction can be encouraged by increasing the likelihood that a prediction will be correct (Brothers, Swaab, & Traxler, 2017; Lau et al., 2013). The argument here is that the degree of automatic priming should not be affected by strategic considerations. In one study, when the number of semantically related prime-target pairs in an experimental block was increased, the difference in the N400 between related and unrelated pairs was smaller (Lau et al., 2013). This was taken to mean that the N400 was driven more by predictive processing than semantic association. However, it is conceivable that as the proportion of related prime-target pairs increased, there was also less time for activation levels to decay between trials. This would mean that an additive effect of semantic priming could account for the apparent up-regulation of predictive processing.

Also difficult to explain with a predictive processing account are findings that activation may spread to words not licensed by or improbable in the given context. Studies have shown that semantic category members appear to receive some activation in contexts where they are unlikely, such as the category member *pines* where *palms* is the most expected completion in the following sentence (Federmeier & Kutas, 1999; Metusalem et al., 2012):

(3) They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of...

In a similar vein, transient priming of word-relevant but not context-relevant continuations has been demonstrated (Kukona et al., 2014). In this visual world study, subjects hearing the sentence *The boy eats a white...* briefly looked more at the non-target *car* as well as at the target *cake* more than at other distractor nouns, despite *car* being semantically implausible in the given context. However, *white car* is a temporarily plausible adjective-noun combination. These findings suggest that, at least in very early stages of processing, prediction may not be distinguishable from priming. Thus, demonstrating that predictions can be sustained beyond the very next word is key.

Predictive processing beyond the next word has been demonstrated in structural and discourse expectations. Evidence for the structural prediction of a distant gap is seen in filler-gap dependencies (Fiebach, Schlesewsky, & Friederici, 2001; Ness & Meltzer-Asscher, 2017; Phillips, Kazanina, & Abada, 2005). Maintaining the possibility of a verb particle in particle verb constructions may also lead to a higher working memory load (Piai, Meyer, Schreuder, & Bastiaansen, 2013). Predictions about discourse set up by the connective *even so* have also been shown to increase the amplitude of the N400 if violated (Xiang & Kuperberg, 2015), although *even* is apparently not integrated rapidly enough to reduce the effect of unpredictable words on reading times in sentences such as *The cat [even] chased a weasel last night* (Mayer, Dillon, & Staub, 2019). Much more scant, however, is evidence, that predictions about more detailed elements such as full lexical items are generated and maintained over distance. If the effects in studies such as DeLong et al. (2005) demonstrate prediction of a specific lexical item and not automatic priming of that word, then it should follow that lexical predictions, as well as structural, should be sustainable over distance. Evidence exists that lexical items strongly favoured by the context receive the most facilitation (Husain et al., 2014; Levy, 2008), but as yet uncertain is whether these lexical items were predicted in advance. In order to determine whether this is the case, an important step is to precisely define the mechanism by which predictive processing occurs.

1.4 A mechanism for prediction: Hypotheses from computational models

How are multiple streams of information combined to generate predictions about the rest of a sentence? The preactivation account of predictive processing outlines how activation from sensory input flows through bottom-level to higher level representations (Kuperberg & Jaeger, 2016). Pre-activation of upcoming words is then thought to occur after the context of a sentence has triggered high-level associations such as event structure. This high-level information is used to predictively pre-activate lower level features such as semantic and morphosyntactic information (Federmeier & Kutas, 1999; Luke & Christianson, 2016; Metusalem et al., 2012). With sufficient contextual information, pre-activation may even spread to a probable lexical item(s), including its bottom-level features such as phonological form. Each new word input triggers preactivation and updating at multiple levels of representation (Kuperberg & Jaeger, 2016).

How is such an account to be tested? One approach is to develop a computational model. Computational models offer an advantage over verbal theories in that they are more easily communicated (e.g. with concrete computer code) and thus allow more consistent testing of theoretical predictions against reality (Farrell & Lewandowsky, 2018). Furthermore, different cognitive mechanisms can underlie the same behavioural observation. Modelling these mechanisms quantitatively can help to decide which most closely resembles the truth, and may lead to novel predictions about behaviour that can in turn be tested empirically (Farrell & Lewandowsky, 2018). At present, there is no computational model of predictive processing per se, although there are many that model individual aspects of the processes that are presumed to be involved.

An ideal model of predictive processing should account for how multiple sources of bottom-up and top-down information interact with each other at what level of cognition and at what point in time following a stimulus, and accurately predict the effect of this process on empirical measurements such as reading times. Crucially, the ideal model should provide a mechanism that distinguishes preactivation from priming, if indeed these are two separate processes. There are a number of existing computational models, each addressing a subset of these factors. For example, some models describe how context affects word processing difficulty, but do not propose a mechanism for how this occurs (Levy, 2008). Some propose a mechanism for how uncertainty affects word processing, but are agnostic about where (e.g. in working memory) or when it is implemented (Hale, 2001, 2006). Others, such as race models, model the accumulation of activation towards a lexical decision, but do not describe an internal mechanism for how activation accrues or from what sources (Staub et al., 2015). Connectionist models come closer by describing the accumulation of activation with respect to specific information streams such as orthography (McClelland & Rumelhart, 1981), orthography plus top-down information (Rumelhart & McClelland, 1986b), morphology (Rumelhart & McClelland, 1986a), orthographic and phonologic content (Seidenberg & McClelland, 1989), and syntactic structure (Chang, 2002; Elman, 1991). Other models describe a probabilistic environment created via multiple streams of information and the effect that this has on the processing of full lexical items, but not parts thereof (Jurafsky, 1996; Rabovsky, Hansen, & McClelland, 2018). Each of these models provides some insight into how context affects reading.

1.4.1 Incremental increases in context facilitate word processing

The context of a sentence sets up expectations about what structural and even what lexical information might appear next. This forms the basis of the expectation models of sentence processing, whereby the difficulty of parsing a word is equal to the negative log probability of that word appearing given its preceding context (Hale, 2001; Levy, 2008). The expectation models assume that parsing occurs in a parallel, incremental fashion, with each new, incoming word triggering an update of the probability distribution of likely sentence continuations. There is no commitment to any one particular sentence continuation; instead, a set of possible sentence continuations is held in parallel, ranked by their frequency in a probabilistic context-free grammar (PCFG). The degree of update required to the probability distribution induced by each new word input is proportional to the difficulty of processing the new word; that is, the greater the update, the greater the "surprisal". Surprisal is a function of the resource allocation cost posed by the *discarding* of potential syntactic parses at that word (Hale, 2001; Levy, 2008). 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 a structurally expected word. Conversely, at a structurally unexpected word, surprisal will be higher. For example, in the sentence (Hale, 2001):

(4) The horse raced past the barn fell.

according to a PCFG, the probability of the verb *raced* being the main verb is seven times higher than the probability of it being part of a reduced relative subject noun phrase with adjoining verb phrase. The higher-probability (but incorrect) reading is therefore the highest ranked with the largest probability mass. At encountering *fell*, the parser must then discard the large probability mass associated with the main verb reading, leading to long reading times at *fell*.

The expectation model makes accurate predictions as to reading times in sentences with ambiguity, subject preference, and wh-disambiguation (Levy, 2008). It also correctly predicts that, in head-final languages such as German, Hindi, and Japanese, clause-final verbs should be read faster when there is more intervening information. This has been empirically confirmed (Husain et al., 2014; Konieczny, 2000; Lewis & Vasishth, 2005; Nakatani & Gibson, 2008; Vasishth, 2003; Vasishth & Lewis, 2006). The expectation model explains this effect with the fact that more intervening information eliminates more potential structural candidates based on the parser's frequency-based knowledge of grammar. For example, a prepositional phrase is less likely to co-occur with another prepositional phrase based on a PCFG (Levy, 2008). This eliminates or down-ranks any candidate structure that contains a prepositional phrase, thus restricting the pool of potential candidates and redistributing probability mass among the remaining candidates. In this way, the probability of a verb appearing next (or soon) increases and the degree of update necessary to the probability distribution decreases.

Naturally, there are linguistic phenomena that surprisal does not account for. The first is English relative clauses, where object-extracted relative clauses (ORCs; 5b) are generally found to be more difficult to process than subject-extracted clauses (SRCs; 5a). Surprisal would predict that the higher corpus-based frequency of the SRC structure should make it easier than the ORC structure, which is indeed supported by faster reading times for SRCs over ORCs:

(5) a. The reporter who sent the photographer to the editor hoped for a good story.
b. The reporter who the photographer sent to the editor hoped for a good story.

However, while surprisal correctly predicts more difficulty in ORCs, the predicted location of the difficulty differs from the observed location (Grodner & Gibson, 2005; Hale, 2001; Levy, 2008; Levy & Gibson, 2013). Surprisal predicts that reading time slow-down will occur at the embedded ORC subject *the photographer* due to its lower frequency and therefore the parser's lower expectation for encountering it. In human data, however, the SRC advantage is actually observed at the verb *sent*. One study has noted difficulty at *the photographer*, although to a lesser degree than at the verb *sent* (Levy & Gibson, 2013; Staub, 2010). At *sent*, however, surprisal actually predicts that ORCs would be read faster since a broader range of syntactic structures are possible after *the reporter who* than after *the reporter who the photographer* (Levy & Gibson, 2013). The expectation model-based prediction does hold for a number of non-English languages, (Konieczny, 2000; Levy & Keller, 2013; Nakatani & Gibson, 2008; Vasishth & Lewis, 2006), but not for English or Russian relative clauses (Levy & Gibson, 2013; Levy, Fedorenko, & Gibson, 2013).

One thing lacking in the original expectation models was the effect of decay and interference over distance (Gibson, 1998; Lewis & Vasishth, 2005). A more recent iteration of surprisal attempting to address this issue is the noisy channel surprisal model, where the preceding context is presumed to have some degree of noise in its mental representation (Futrell & Levy, 2016). In this variant of the model, the reader is assumed to build noisy (i.e. potentially incorrect) sentence representations and generate expectations about the structure of the rest of the sentence based on these noisy representations. Surprisal remains the measure of processing difficulty of a word, but the probability of the next word is calculated assuming a noisy, rather than a perfect, representation of the sentence seen so far. The types of noise modelled are *erasure noise*, where a symbol in the sentence representation is replaced, and *deletion noise*, where a symbol is completely deleted. When symbols are erased and replaced, the comprehender is aware of which symbols have been affected. When symbols are deleted, the comprehender no longer knows how many symbols were previously there. The model has been empirically evaluated on task-induced structural forgetting, such as grammaticality illusions in double centre embeddings such as these examples from Vasishth, Suckow, Lewis, & Kern (2010):

(6) a. * The apartment that the maid who the cleaning service had sent over was well decorated.

b. The apartment that the maid who the cleaning service had sent over was cleaning every week was well-decorated.

In such sentences, English native speakers have been found to be more surprised by the third verb phrase in 5b *was well-decorated*, even though it is grammatical (Frazier, 1985; Vasishth et al., 2010). Conversely, the ungrammatical version (missing a verb phrase in 6a) is processed more easily. In contrast, German native speakers were correctly more surprised when the second verb phrase was absent in 6a (Vasishth et al., 2010), although other research has elicited the grammaticality illusion in German using stimuli with a lower processing load (Bader, 2012; Bader, Bayer, & Häussler, 2003; Häussler & Bader, 2015). Dutch native speakers did not show the grammaticality illusion in their own language, but both Dutch and German native speakers show the effect when reading in English (Frank, Trompenaars, & Vasishth, 2016). This was taken to demonstrate that memory resources can be shaped by the distributional statistics of a reader's native language. Noisy channel surprisal predicts these effects, at least qualitatively, via a parameter encoding the probability of a relative clause being verb-initial. This is relatively frequent in English, but never occurs in German relative clauses, where the verb must appear clause-finally. The higher probability of the structure of the German relative clause means it is less susceptible to forgetting in the noisy channel model than the English relative clause (Futrell & Levy, 2016).

In sum, expectation models show how iterative increases in context can lead to facilitation of word processing. The key feature missing from the expectation models is an account of the cognitive mechanism by which context generates expectations or reallocates resources. However, higher surprisal at a word has been associated with larger N400 amplitude (Frank, Otten, Galli, & Vigliocco, 2013) and is closely related to cloze probability (Levy, 2008), both of which have been associated with the relative activation of a word (Kutas & Federmeier, 2011; Staub et al., 2015). This suggests that surprisal may be describing the preactivation of structural and lexical candidates; although at least in terms of lexical preactivation, surprisal alone does not capture all empirical observations of the N400 (Frank et al., 2013; Rabovsky et al., 2018).

1.4.2 The facilitative effect of context may be driven by spreading activation

Connectionist models provide a mechanism by which sensory input triggers spreading activation that can facilitate processing of the next input. Connectionist models derive their name from the fact that cognitive processing occurs exclusively via activation spreading through interconnected units (Rumelhart & McClelland, 1986b). The activation can be excitatory or inhibitory, and must reach a certain threshold to be propagated, much like the human system of neurons and action potentials. Connectionist models differ from symbolic models such as the expectation models above, which are modularised and contain symbolic representations of language manipulated by a set of rules (MacDonald & Christiansen, 2002). Another examples of a symbolic model is the ACT-R model of reading to be discussed later in this chapter (Lewis & Vasishth, 2005). The ACT-R model also assumes spreading activation, but comprises separate modules for declarative memory, procedural rules, and working memory. In a connectionist model, in contrast, there are no separate modules for separate language functions such as syntax, semantics, and working memory; in the connectionist model, working memory "is the network itself" (MacDonald & Christiansen, 2002). A key contribution of connectionist models has been to demonstrate that human-like language performance can be achieved in the absence of any predefined assumptions about complex syntactic theory (Linzen, 2019; McClelland & Rumelhart, 1981).

One of the first major neural network models used to describe the process of word recognition was parallel distributed processing (PDP; Rumelhart & McClelland, 1986b). Its first instantiation, the model attempted to recover human behavioural observations that letters are more quickly recognised in word-like strings than alone or in random strings (McClelland & Rumelhart, 1981). Later iterations of the model included syntactic and semantic constraints (Rumelhart & McClelland, 1986b), past tense morphology (Rumelhart & McClelland, 1986a), and orthographic and phonologic content (Seidenberg & McClelland, 1989). Each model consisted of interconnected levels such as a visual feature level, a letter level, and a word level. Higher levels fed back input to lower levels. Interconnectivity within and between the layers meant all information (bottom-up and top-down) could be accessed and processed in parallel. Input from other units could be either excitatory or inhibitory, such that the activation of a unit at a given time point was the net sum of all excitatory and inhibitory activation received. Without further input, nodes would gradually return to their resting activation state. Resting activation differed depending on how often the nodes were activated, such that more frequent activation led to a higher level of resting state activation. Each of these models was able to replicate human phenomena such as the above findings about letter recognition (McClelland & Rumelhart, 1981), the temporary overgeneralisations of children learning the English past tense (Rumelhart & McClelland, 1986a) and the ability to pronounce novel words despite their having no representation in the lexicon (Seidenberg & McClelland, 1989).

One issue with these early neural networks was that words and letters were pre-assigned to units. Later distributed models utilised a simple recurrent network (SRN) that did not pre-assign words or concepts; these had to be learned by the model (Elman, 1991). Elman's model allowed forward propagation of activation through a unit and back-propagation to update its activation weights following feedback. The activation weights of one cycle were also copied so that they can be incorporated in the next cycle. Thus, the state of the model after a cycle was the current input plus the copied activation from the previous cycle. The SRN was trained on sentences with hierarchical and recursive relationships, for example:

(7) a. Girl feeds dogs.
b. Girls see.
c. Dog who chases cat sees girl.
d. Boys who girls who dogs chase see hear.

Its task was to predict the next word in a novel sentence. It was accurately able to predict agreement of the verb with the subject, even after an embedded relative clause, as well as use the verb's correct argument structure. More than that, a principle component analysis (PCA) revealed that the model distinguished nouns by role, grouping subjects and objects separately in the PCA space. Verbs were also differentially represented according to whether or not they took objects.

The model's main contribution to the understanding of predictive processing was to demonstrate its likely role in the acquisition of grammar, without the need for a pre-defined set of language rules. In using its current state to predict the next word and updating activation weights based on the accuracy of that prediction, the model was successfully able to acquire and apply grammatical concepts to sentences it had not seen before. Later, more nuanced models have also further suggested that prediction may be a key strategy by which the parser learns grammar. The dual path model has been able to recreate the effects of structural priming (Chang, 2002; Chang, Dell, & Bock, 2006) and provides an account of how children acquire structural ordering by trial-and-error (Twomey, Chang, & Ambridge, 2014). For example, the location-theme ordered construction *sprayed the wall with water* is specific to some types of verbs (cover, coat), while the theme-location ordered construction *sprayed water on the wall* is specific to others (pour, spill). Yet other verb classes can be used in either construction (spray, squirt). The dual-path model accurately predicted the way that children temporarily overgeneralise to other verbs such as *cover* and produce sentences such as **covered water on the wall*, until semantic verb classes are acquired. Both Elman's model and the dual-path model demonstrate how the statistical properties of words can be learned simply by trial-and-error and eventually categorised with other verbs that behave in the same way.

Criticism of Elman's model and of connectionist models in general has been that they contain too many degrees of freedom that can be arbitrarily set by the researcher, such as the number of layers and nodes, and the type of training set (Christiansen & Chater, 1999). For example, Elman's model was trained on a small context-free grammar and was then tested on the same structures it was trained on (Christiansen & Chater, 1999). Christiansen & Chater (1999) therefore used artificial languages exhibiting different types of human recursive structure to train their model and found that it was accurately able to replicate human reading behaviour in dealing with centre embedding, cross-dependency embedding, and right-branching recursion. Importantly, it captured reading times on single embedding after having been trained only on centre and cross-dependency embedding. Even more importantly, it demonstrated similar effects regardless of changes to the number of layers and nodes.

Although connectionist models have succeeded in demonstrating that a broad range of human behaviour can be captured with fewer pre-defined rules and functional units than symbolic models, they have not yet been able to capture all the nuance and complexity of human language. This nuance can be at the level of how we are able to correctly interpret the phrase *I saw the Grand Canyon flying to New York* (Rumelhart & McClelland, 1986a) to fine-grained grammatical complexities such as being able to correctly generalise all types of wh-island constraints (Wilcox, Levy, Morita, & Futrell, 2018). A further example is that, while SRNs are able to explain individual differences in working memory span simply by the amount of reading experience the network has had, this predicts a potentially infinite increase in performance over the lifespan, which is not a biological reality (MacDonald & Christiansen, 2002). That is, while the difference in reading performance between a human of 10 and 20 years of age would be quite large, the difference between 20 and 50 years would be relatively small, and, after a certain age, would start to decline (MacDonald & Christiansen, 2002).

Outstanding questions in the field of connectionist models include whether they are actually able to acquire human-like syntactic behaviour, or whether they simply develop clever strategies for "cheating the system" (McCoy, Pavlick, & Linzen, 2019). If the latter, this would mean that generalisations to human behaviour based on model predictions would be faulty, and thus unlikely to be observed empirically. Neural networks are also often tested on their ability to predict the next word, with the assumption that the word with the highest activation is that which will be predicted. However, some have argued that the most commonly predicted elements of a sentence are not full word forms, but rather parts thereof such as syntactic or semantic features (Luke & Christianson, 2016). Probing the internal structure of such models has shown how full word forms may be organised (Coenen et al., 2019; Elman, 1991; Tabor, 2000), which offers clues about how word features may be being associated by the model. It is conceptually possible that these feature groupings could be accessed to generate partial predictions about, for example, semantic category or thematic role, but this is not the approach usually taken by predict-the-next-word model assessment strategies. Finally, predictions for elements beyond the next word and how they are maintained in the face of factors known to affect sentence processing, such as resource constraints, activation decay, and interference are also not a feature of current connectionist models.

1.4.3 Distance may adversely affect activation levels

Predictions about upcoming sentence information may have to be maintained for some time before the predicted element or elements are encountered. As predicted dependencies between words are stretched over longer and longer distances, additional factors come into play. These include the amount of new information separating the dependent words (Gibson, 1998, Gibson (2000)), temporal decay, and whether or not the intervening information shares common cues with the dependent words (Lewis & Vasishth, 2005). These factors must be considered when arguing that structural and lexical predictions can be maintained over distance. While connectionist models do include inhibition from words with common features (McClelland & O'Regan, 1981), they do not give an explicit account of how this might affect sentence processing. A number of symbolic models have specifically addressed this question, however.

1.4.3.1 The limits of working memory in storing and integrating new information

The locus for maintaining linguistic predictions is generally assumed to be working memory (Gibson, 1998, 2000; Lau et al., 2013; Lewis & Vasishth, 2005). Working memory has limits, however, on processing dependent sentence elements over distance (Just & Carpenter, 1980, 1992). An operationalisation of working memory limitations on sentence parsing came from Sentence Processing Locality Theory (SPLT; Gibson, 1998) and Dependency Locality Theory (DLT; Gibson, 1998, 2000). Here, word activation was assumed to decay over time and the number of new discourse referents (objects or events) introduced between two dependent words was assumed to additively increase the cost of integration and storage of the final dependent element. While non-new referents may also take up resources, the energy expended on this was initially thought to be low enough to be ignored (Gibson, 2000), although later findings suggest this may not be the case (Gibson & Wu, 2013). In fact, a number of studies of Mandarin Chinese have found faster reading times for object relative than for subject-relative constructions, despite the fact that object relatives introduce more discourse referents (Gibson, 1998; Warren & Gibson, 2002).

Under DLT, processing difficulty of a particular word is calculated as a function of that word's storage and integration cost, in addition to other factors such as frequency, plausibility, and discourse complexity (Gibson, 2000). A slow-down in processing is proportional to the cost of integrating the new head into the building sentence parse and maintaining the parse plus its associated syntactic requirements in working memory. Potential grammatical parses are held in working memory in parallel, ranked by frequency and in line with the structural, lexical, discourse, and other constraints of the sentence seen so far. To be able to integrate new input, relevant structure(s) in working memory must be reactivated past a certain threshold. Since the working memory resources needed to perform these actions are assumed to be finite, the more potential structures there are, the longer reactivation will take. The longer a predicted structure must be kept in memory, the greater the cost of maintaining it. Cost is computed as a function of the amount of activation required to reach threshold. Likewise, the greater the distance between an incoming word and its attachment site, the greater the integration cost. A number of linguistic phenomena support this account.

DLT correctly predicts that integrating the embedded verb of a subject-extracted relative clause should incur less cost than the verb of an object-extracted relative clause, because two integrations have to take place in object-extracted relative clauses versus one in subject-extracted relative clauses. For example, in the following example from Gibson (1998):

(8) a. The reporter who attacked the senator admitted the error.
    b. The reporter who the senator attacked admitted the error.

the embedded verb *attacked* should be read faster in 1a than in 1b, because it must be attached as the verb for *the senator* as well as attaching an empty category for its object *the reporter*, co-referenced with the pronominal *who*. While each of these thematic roles is still present in 1a, at the embedded verb *attacked*, the second object referent *the senator* has not yet been introduced. Thus, integration of the two referents in 1b is costlier than in 1a. The "subject-relative advantage" has been attested in a number of studies and languages (Betancort, Carreiras, & Sturt, 2009; Traxler, Morris, & Seely, 2002; Ueno & Garnsey, 2008; Vasishth, Chen, Li, & Guo, 2013), although there is some evidence of the opposite effect in Chinese, possibly dependent on structural bias (Gibson & Wu, 2013; Hsiao & Gibson, 2003).

Other phenomena that DLT correctly predicts include that incomplete dependencies should increase storage cost and slow down matrix verb processing, such as where a subject is separated from its matrix verb by an increasingly long relative clause (Grodner, Gibson, & Tunstall, 2002). DLT also predicts a processing overload in complex constructions such as multiple embeddings, resulting in structural forgetting. This has indeed been observed in English and French, where a missing verb is not detected in sentences with three layers of embedding; a so-called "grammaticality illusion" (Gibson & Thomas, 1999; Gimenes, Rigalleau, & Gaonac'h, 2009). German and Dutch readers did not show the same overload in their native language, however, leading some researchers to speculate that working memory limitations may reflect language experience (Frank et al., 2016; Vasishth et al., 2010). In other measures, however, (reaction time and prediction error), German native speakers have been found to demonstrate the grammaticality illusion (Bader et al., 2003; Häussler & Bader, 2015).

One critique of the locality effect has been that it is largely observed in studies of English and rarely in head-final languages such as German and Hindi (Vasishth & Drenhaus, 2011). Whether this observation reflects real processing differences between head-initial and head-final languages or simply an English-centric research field is unclear. One study challenging the former hypothesis is of the head-final language, Persian, which observed exclusively locality effects, regardless of verb predictability (Safavi et al., 2016). Locality has also been noted in head-final structures in German (Levy & Keller, 2013; Vasishth, Mertzen, Jäger, & Gelman, 2018). Conversely, antilocality effects have been elicited in English in unpublished data creating distance an increasing number of prepositional phrases (Jaeger, Fedorenko, Hofmeister, & Gibson, 2008). It is thus clear from the variability in locality and antilocality effects that the number of discourse referents, as modelled by DLT, is not the only factor in dependency processing.

1.4.3.2 Similarity-based interference and temporal activation decay

In addition to resource constraints, two further factors that could affect the activation levels of predicted information in working memory are interference from elements with matching features and the decay of activation over time. In a model using the ACT-R framework, the effects of similarity-based interference and activation decay on the processing of long-distance dependencies have been demonstrated (Lewis & Vasishth, 2005). This activation-based model is a symbolic model comprising separate modules for working memory, the lexicon, and a set of production rules. Intermediate sentence structures and long-term lexical content are stored in declarative memory as chunks with feature-value pairs. Access to the lexicon is influenced by both frequency of use and the current context. Activation can fluctuate depending on how many times the structure is reactivated by incoming information, but will otherwise decay over time. When a word is encountered, a chunk is accessed in declarative memory, including its associated syntactic information such as argument structure. This syntactic information generates a syntactic expectation which, combined with the current working memory module contents, triggers retrieval cues for finding a prior constituent to attach to. Finding this constituent takes time and is subject to interference from information that may match some or all of the retrieval cues. Full lexical items can also be predicted and integrated, and receive additional activation if the context continues to support their prediction. The model's main proposal is that the time taken to retrieve the constituent to which new information will be attached is what determines word reading times.

The activation-based model correctly predicts both speed ups and slow-downs in sentences, in contrast to DLT and expectation-based models which predict either exclusively slow-downs or speed-ups, respectively (Lewis & Vasishth, 2005; Vasishth & Lewis, 2006). For example, in the garden path sentence *The assistant forgot the student...*, two parses are triggered; one analysing *the student* as a direct object and one analysing it as the head of a reduced relative clause (Van Dyke & Lewis, 2003). The subject-object structure is the preferred interpretation, as demonstrated by a cloze test. This is the parse that is therefore pursued, while the reduced relative clause interpretation is left to decay. If the sentence continues with the predicted subject-object structure, retrieval time will be reduced and reading time will be faster. However, when the next chunk, *was standing*, indicates a reduced relative clause, energy is needed to reactivate the decayed structure, leading to longer retrieval time. The longer the distance between the object and the disambiguation point, the more severe the decay and the more effort is required to reactivate and retrieve the less-expected parse. The model therefore predicts that increased distance can speed up processing if the most expected structure is encountered, and slow down processing if not.

In addition to usage and decay, a further important factor influencing retrieval time in the model is similarity-based interference (Lewis & Vasishth, 2005; Van Dyke & McElree, 2006). Similarity-based interference is the effect of competition during the retrieval of constituents from working memory. Matched features such as *plural* or *animate* provide cues that compete, leading to lower activation for both and, occasionally, retrieval of the wrong word or form, such as in agreement attraction errors (Wagers, Lau, & Phillips, 2009). Both syntactic and semantic cues can contribute interference (Van Dyke & McElree, 2011). Interference has been shown to be distinguishable from the effects of decay on dependency processing (Van Dyke & Lewis, 2003), suggesting that both are important factors in predicting reading times. On the other hand, others have argued that decay is not an important factor in predicting reading times, or that variation in reading times is better explained by interference (Lewandowsky et al., 2009; Vasishth et al., 2019).

The benefit of the activation-based model is that, in contrast to DLT and expectation-based models, it provides an account of word processing difficulties *and* an explicit underlying process. It accurately predicts a range of reading phenomena with minimal changes to the supporting ACT-R framework (Lewis & Vasishth, 2005). However, like all symbolic models, it requires the explicit setting of computational rules based on syntactic theory, which inevitably introduces researcher degrees of freedom. It also assumes that working memory is the sole locus of processing difficulty and that reading time is determined exclusively by retrieval time, while other models make more accurate reading time predictions using a different process (Nicenboim & Vasishth, 2018). Finally, the model gives an account of how upcoming structure and even lexical items may be retrieved. However, since it not a model of predictive processing, this forward structure building is only specified to the depth of cue matching and does not distinguish between alternative explanations such as integration and priming.

1.5 Conclusions and rationale for this thesis

Predictive processing is distinguishable from integrative processes by the finding that unseen words can affect the processing of words already seen. Less certain is whether prediction can be distinguished from priming. One way to demonstrate a distinction would be to show that lexical predictions are not just transient, as they would be in priming, but rather sustained over long distances. Research to date has demonstrated the likely existence of long-distance *structural* predictions; however, evidence for the prediction of distant *lexical* items is rare. This may be because other factors such as uncertainty, interference, activation decay, and resource constraints come into play once dependencies between words are stretched. In this thesis, I contribute to evidence distinguishing predictive processing from priming by demonstrating that lexical preactivation, and potentially even lexical predictions can be sustained over long distances. In the following chapter, I will detail the methods used to demonstrate this.

Chapter 2

2 Methods in prediction research

A range of empirical methods are employed in the study of predictive language processing ranging from sentence completion tasks to functional magnetic resonance imaging (fMRI). In this chapter, I introduce the methods used in this thesis; namely, cloze tests, self-paced reading, eye tracking, and event-related potentials (ERP). While each method has numerous variants and empirical measures that can be exploited by researchers, I limit the discussion to those utilised in this thesis.

2.1 The cloze test

The predictability of a particular word at a given position in a sentence is most commonly measured with a cloze test in which participants are asked to complete sentences cut off at a critical point (Taylor, 1953). While the cloze test is highly correlated with empirical measures such as reading times (N. J. Smith & Levy, 2013) and the N400 ERP component (Kutas & Federmeier, 2011), it is an offline task whose driving process is the subject of much debate. In particular, it not clear whether the distribution of cloze probabilities given by a sample of people represents the distribution of word activations within each respondent's lexicon, or the frequency of that word's use by the whole sample (Staub et al., 2015). An argument against the latter possibility is that cloze probabilities do not correspond well with corpus statistics (N. J. Smith & Levy, 2011). Staub et al. (2015) propose instead that verbal cloze response times can be described by an activation race model, where evidence for a number of potential responses independently accumulates until a "winner" reaches some threshold. If this were the case, Staub et al. (2015) suggest that cloze probability reflects the level of activation that alternative responses have at a given point in the sentence.

2.2 Self-paced reading

The self-paced reading (SPR) paradigm used in this thesis involved experimental participants pressing on a keyboard key to reveal the next word of a sentence. Only the current word was ever visible on the screen (Ferreira & Henderson, 1990; known as the moving-window or non-cumulative design, Just, Carpenter, & Woolley, 1982). When a reader slows down their rate of key pressing, this is thought to indicate an increased cognitive processing load at the corresponding words, or, more commonly, spilling over to the subsequent words (Just et al., 1982; Mitchell, 1984).

The main advantage of SPR to the experiments presented in this thesis is that the target region of the experimental stimuli is always very small (often 2-3 letters), making it very likely to be skipped in natural reading (Rayner, 2009). SPR ensures that every word is presented to the reader. The disadvantage of SPR is that it does not give any precise clues as to the nature of slowed key pressing at or after a target word. For example, has the slow-down occurred because readers are having difficulty retrieving the word from working memory (due to e.g. interference from the experimental manipulation)? Or because they are reactivating their internal sentence representation to check if they have parsed it incorrectly? Individual subjects will also differ in which of these processes is more affected, and even in whether processing load effects are seen at the target word or in the 'spillover' region (Just et al., 1982; Mitchell, 1984). Furthermore, the motor activity required to plan and deploy button pressing may also obfuscate cognitive processing (Just et al., 1982). On the other hand, SPR does not require specialised equipment, and is cheap and simple (Mitchell, 1984). It can also be deployed online, which has allowed researchers to reach large sample sizes with a broader spectrum of backgrounds and abilities than the average lab-based experiment.

2.3 Eye tracking

The pattern and duration of eye fixations during reading has been tied to various cognitive aspects of language processing (Rayner, 1998). Tracking eye movements while reading can fill the gap left by self-paced reading: when a difficult region of a sentence is encountered, it is possible to infer from eye movements what a reader does to resolve the difficulty. They may regress to earlier parts of the sentence, spend more time looking at the target word, or skip forward to later parts of the sentence. Each of these categories of movement has been linked to stages of linguistic processing.

Predictive processing has been linked to the very earliest stages of lexical access (Rayner, 1998; Staub, 2015). Its effects are therefore more likely to be seen the first time a reader fixates on a particular word and how long they spend looking at it, reflected in the measures first fixation duration and first-pass reading time. The effects of predictability on these early measures are often also observed in total gaze duration (Rayner, 1998). Skipping a word entirely, particularly if it is a content word, is also strongly linked that word's predictability (Ehrlich & Rayner, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004). Other eye movements that are not usually linked to predictability, but provide useful information about how processing difficulty may be being resolved are regressions. Regressions of the eye to earlier parts of the sentence suggests readers are reanalysing their internal sentence parse (Pickering & Traxler, 1998) and have been linked to event-related potential (ERP) components highly associated with syntactic reanalysis (Metzner, von der Malsburg, Vasishth, & Rösler, 2017).

Eye-tracking has its own disadvantages, of course, including variation in preprocessing methods between labs and a large number of measurements seducing researchers into multiple comparisons (von der Malsburg & Angele, 2016). The issue of individual differences in reading strategies also exists, as it does for SPR (Rayner, 1998).

2.4 Event-related potentials (ERP)

Each time a new word in a sentence is encountered, the brain's activity in response to that word can be picked up by electrodes on the scalp. By timelocking the activity to a word's onset and subtracting background noise, an event-related potential (ERP) can be computed at each word. Each ERP comprises a number of components, each of which has been associated with different cognitive processes. The components of primary interest to the current thesis are the N400 and the post-N400 positivities.

2.4.1 The N400

The N400 is a negative deflection in the ERP over the posterior scalp that onsets from around 200 ms after a word is presented, peaks at around 400 ms, and lasts until about 600 ms (Kutas & Federmeier, 2011). The N400 is elicited by meaning-based incongruities in an incoming stimulus such as words semantically incongruent with a given context (Kutas & Hillyard, 1980b, 1984), but it can also be elicited by non-word stimuli with some linguistic meaning, such as unexpected phonological and orthographic information in pronounceable pseudowords (Bentin, McCarthy, & Wood, 1985; Deacon, Dynowska, Ritter, & Grose-Fifer, 2004), and line drawings of objects unexpected in a given context (Wicha, Moreno, Kutas, Jolla, & Related, 2003). In contrast, the N400 is *not* elicited by surprising information that does not have linguistic meaning, such as unexpected font size changes (Kutas & Hillyard, 1980b), or unexpected geometric patterns and musical notes (Besson & Macar, 1987).

The most consistent finding in N400 research is that the component is inversely correlated with a word's cloze probability; that is, its amplitude becomes larger the lower the number of readers who would expect to see that word (Kutas & Hillyard, 1980b, 1984). Its amplitude also becomes lower at each new word in a sentence, as long as that word is congruent with its preceding context (Payne et al., 2015; Van Petten & Kutas, 1990). While it is clear that the

N400 reflects the processing of meaning derived from a context, there are differing accounts about what stage of processing the component indexes.

The two main accounts of the cognitive generators of the N400 are the *prediction* and the *integration* views. The prediction view proposes that semantic cues from a linguistic context preactivate lexical representations of plausible words, facilitating their processing once encountered and reducing the N400's amplitude (Kutas & Federmeier, 2011; Lau, Phillips, & Poeppel, 2008). It should be noted that while the term *prediction view* is in common use, the account also encompasses views that weight lexical access over lexical predictions (Kutas, 2018). The *integration view* of the N400, in contrast, proposes that the component reflects the combinatorial effort of integrating a new word into an existing sentence representation (Brown & Hagoort, 1993; Hagoort et al., 2009; Van Berkum et al., 1999). Evidence for and against these theories is discussed here.

2.4.1.1 The prediction view

In a synthesis of several decades of N400 research, Kutas & Federmeier (2011) propose that the N400 reflects the stage where multiple streams of information from primary sensory processing spread to the broader linguistic network. At this stage, the current activation state of context-plausible lexical entries changes with new evidence accumulating from both bottom-up and top-down input. The greater the activation change for a particular word, the larger the N400. An input word with low cloze probability therefore elicits a large amount of activation change for that word's lexical entry and a correspondingly larger N400. One strong argument for the role of the broader linguistic network in N400 amplitude is that to elicit an N400 requires violation of more than just surface-level word meaning. Evidence for this comes from N400 'blindness' to thematic role reversals.

In a series of studies, syntactically correct words with implausible meanings were found not to elicit the expected N400, but rather only the P600 traditionally associated with syntactic anomalies (A. Kim & Osterhout, 2005; Kuperberg, 2007; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003). For example, in *The hearty meal was devouring...* (A. Kim & Osterhout, 2005), a P600 was elicited at *devouring*, even though incongruity is only detected via the verb's meaning and not by its form; i.e. it is not the *-ing* form of the verb that is causing the problem, since *cooling* or *waiting* would be perfectly acceptable alternatives. Therefore, the meaning of *devour* must be the source. The absence of an N400 was taken to mean that the strong semantic attraction between *hearty meal* and *devour* had made the N400 blind to the incongruity, even though the meaning of *devour* was enough to trigger conflict in the syntactic system (A. Kim & Osterhout, 2005; Kuperberg et al., 2003). What this also suggests is that lexical access can occur fast enough to inform syntactic computations, but not deeply enough to elicit an N400.

More recent studies have supported this interpretation by showing that increasing the amount of time before the incongruity is encountered can reverse N400 blindness (W.-Y. Chow et al., 2018; Momma, Slevc, & Phillips, 2016). In addition to time, it has also been proposed that closer associations of particular nouns with verb-argument roles (e.g. *bull-gorer* versus *villager-hauntee*) may allow the N400 to be generated even in short-distance role reversals (Ehrenhofer, Lau, & Colin Phillips, 2019). The fact that time and the strength of lexical association are essential to the N400 supports the idea that, rather than being sensitive to primary lexical access processes, the N400 reflects the transition to a secondary stage of access where meaningful associations are built (Kutas & Federmeier, 2011).

Under the umbrella of prediction views of the N400 also comes the probabilistic update or the "sentence gestalt" account, in which the N400 reflects the degree to which incoming input updates an existing sentence representation (Rabovsky et al., 2018). This account differs slightly to traditional accounts of sentence processing, in that words are not retrieved from memory and integrated into a building sentence construction, but rather a probabilistic representation of the sentence (of agnostic form) is generated from the statistics of the comprehender's experience (Rabovsky et al., 2018). The degree to which this probabilistic representation must be updated following new input is positively correlated with N400 amplitude; that is, the larger the update, the larger the N400. A neural network implementation of the model has successfully captured a variety of N400 effects (Rabovsky et al., 2018).

One caveat of the prediction view of the N400 is that it predicts that a word's current activation should be higher if a context points more strongly towards it. The N400 is famously insensitive to contextual constraint, however (DeLong et al., 2014b; Federmeier, Wlotko, Ochoa-Dewald, & Kutas, 2007; Kuperberg & Wlotko, 2019; Kutas & Hillyard, 1984; Van Petten & Luka, 2012). In other words, a low-probability word in a strong context will elicit the same N400 amplitude as a low-probability word in a weak context. In contrast, the effect of contextual constraint has been found to affect measures in other modalities that are well correlated with N400. For example, cloze probability is strongly correlated with the N400, and yet cloze test responses are faster in strong than in weak contexts (Staub et al., 2015). However, the effect of constraint on the process driving offline cloze responses is clearly different to that driving the N400; what this process is, and how it could be reconciled with the prediction view of the N400, is still unclear.

2.4.1.2 The integration view

The main alternative view of the N400 is the integration view, which proposes that the N400 reflects the difficulty of integrating a word into a sentence or discourse context (Brown & Hagoort, 1993; Hagoort et al., 2009; Van Berkum et al., 1999). The more difficult as word is to integrate, the larger the N400. One source of support for the integration account of the N400 was a study showing that although semantically unrelated, masked semantic primes (presented at a speed not consciously visible) could slow down lexical decision times, they had no more effect on the N400 than visible primes (Brown & Hagoort, 1993). This was taken to mean that the invisible primes could elicit spreading semantic activation, but that this activation did not affect the processes reflected by the N400. Thus, the N400 could *not* be a product of lexical access. However, this effect is not unlike the thematic role blindness phenomenon described above, and is therefore not incompatible with the prediction view of the N400.

There are several other challenges to the integration view of the N400, including that in order to integrate a word into its context, lexical access of that word must have already occurred (Lau et al., 2008). Directly contradicting this, however, is the fact that pseudowords are able to modulate N400 amplitude, despite not having lexical representations (Deacon et al., 2004; Holcomb, 1993). For similar reasons, N400s elicited by phonographic/orthographic manipulations are equally inexplicable with an integration account (Rugg & Barrett, 1987).

The integration account also predicts that the only factor influencing N400 amplitude should be how congruous a word is with its preceding context. However, congruity alone does not explain all variance in N400 amplitude (Lau, Namyst, Fogel, & Delgado, 2016). In this study, word combinations such as *runny nose* (congruous/predictable), *dainty nose* (congruous/unpredictable), *yellow bag* (congruous/unpredictable), and *innocent bag* (incongruous/unpredictable) elicited large N400s in the unpredictable conditions, regardless of congruity. In contrast, incongruous words had a much smaller effect on N400 amplitude, whether predictable or not. The final piece of evidence against the integration account is the finding that sentence-congruous determiners and adjectives whose only incongruity is with as-yet unseen word can still elicit an N400 (DeLong et al., 2005; Nicenboim et al., 2019; Szewczyk & Schriefers, 2013), although the effect is not undisputed (Ito et al., 2016; Kochari & Flecken, 2019; Nieuwland et al., 2018).

An attempt at incorporating both the prediction and the integration views proposed the idea that semantic congruity and predictability may have effects on the N400 at varying latencies (Nieuwland et al., 2019). In this study, higher noun predictability decreased the amplitude of the N400 from 200 ms with a peak at 330 ms, while higher noun congruity decreased the amplitude in a longer lasting time window, beginning later at 350 ms. The difference in timing was hypothesised to reflect an effect of preactivation on the early upward flank of the N400 and a later effect of integration on its downward flank. This hypothesis is yet to undergo confirmatory analysis, however.

2.4.2 Post-N400 positivities

Anomalous words often elicit positive deflections in the ERP beginning at around 500 ms. The positivity is typically long-lasting, although the 600-900 ms window is “typical” (Van Petten & Luka, 2012). In recent years, evidence has suggested this late positivity may be spatially divisible into two separate components: a posterior P600 and an anterior post-N400 positivity (DeLong et al., 2014b; Kuperberg & Wlotko, 2019; Van Petten & Luka, 2012). While the posterior P600 was originally thought to be driven primarily by syntactic anomalies (Osterhout & Holcomb, 1992), subsequent findings have highlighted a possible role of semantic input (A. Kim & Osterhout, 2005; Kuperberg et al., 2003). Also debated is whether these components are separate to, or merely extensions of the P300 component (Coulson, King, & Kutas, 1998; Osterhout, 1999; Sassenhagen & Fiebach, 2019). The argument surrounding the P300 is beyond the scope of this thesis, however.

2.4.2.1 The P600

The P600 was first described in an early study linking a late positive component to words that revealed that an initial syntactic analysis of a sentence was incorrect (Osterhout & Holcomb, 1992). For example, in the sentence $^{*/?}$*The broker hoped to sell the stock was. . .*, the reader realises at *was* that an active analysis of *The broker hoped to. . .* is not correct. This elicited a large positive deflection in the ERP which was named the P600. Subsequent research associated the P600 with other syntactic processing such as garden path disambiguations (S. M. Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Osterhout, Holcomb, & Swinney, 1994) and noun-verb number agreement violations (Hagoort, Brown, & Groothusen, 1993; Hagoort, Wassenaar, & Brown, 2003).

Surprisingly, the P600 was later found to be elicited by syntactically intact but semantically incongruent words (A. Kim & Osterhout, 2005; Kuperberg et al., 2003). This only appeared to occur, however, when there was a high degree of semantic association within the dependency; for example, a P600 and no N400 was seen at *devouring* in *The hearty meal was devouring. . .*, while the reverse was true in *The dusty tabletops were devouring. . .* (A. Kim & Osterhout, 2005). It was assumed that *devouring* triggered a thematic role assignment to *hearty meal* (actor) that violated the thematic role assignment suggested by the syntax (theme). The fact that thematic role violations could elicit a P600 at all was particularly surprising , since thematic role assignments were typically thought to be lexico-semantic in nature (Kuperberg et al., 2003). In their interpretation of the findings, A. Kim & Osterhout

(2005) proposed that the strong semantic association between *hearty meal* and *devour* was so strong that it triggered syntactic reanalysis of the sentence rather than difficulty processing the semantic implausibility. Various subsequent accounts of the phenomenon have since been given (Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Fitz & Chang, 2019; Kuperberg, 2007), but one important contribution of these studies was to provide evidence against syntax-first models of sentence processing by suggesting that semantic information can operate in parallel to syntax (A. Kim & Osterhout, 2005; Kuperberg, 2007).

A point of some contention has been the precise nature of the cognitive process indexed by the P600. Kuperberg et al. (2003) proposed that the P600 reflected the discrepancy between the thematic roles assigned by the semantic context and the roles assigned by the critical verb. For example, in *For breakfast the eggs...*, semantic knowledge leads the parser to assign a theme role to *eggs* as part of a passive construction, but the appearance of *eat* requires a subject resulting in *eggs* being assigned an actor role. The P600 then reflects the discrepancy in this role assignment. Others proposed that it was the inanimate subject that elicits the P600 (Hoeks, Stowe, & Doedens, 2004). Yet others argued that the P600 was generated in a processing step where a separate stream computing the semantic plausibility of *meal + devour* conflicted with the morphologically-driven, actor-action argument structure suggested by *the eggs...* (Bornkessel-Schlesewsky & Schlesewsky, 2008). Each of these accounts involves the computation of semantic and syntactic computation in parallel but separate streams.

A variation of the dual-stream account has modelled the P600 as being correlated with low activation of the *-ing* verb form, resulting from both semantic and syntactic evidence converging on a prediction for an *-ed* verb form (Fitz & Chang, 2019). In contrast to these dual-stream accounts, the retrieval-integration model successfully accounts for thematic role violation effects with a single computational stream in which the N400 reflects lexical retrieval and the P600 syntactic integration (Brouwer et al., 2017). Generally accepted by all accounts, however, is that the P600 reflects some aspect of syntactic computation.

2.4.2.2 The anterior post-N400 positivity (PNP)

A seminal review of ERP studies on the N400 at incongruous words noted that it is often (about 30% of the time) followed by a second, positive ERP component beginning at around 600 ms for incongruent words, often with a fronto-spatial distribution (Van Petten & Luka, 2012). This has been subsequently been distinguished from the P600 and referred to as the post-N400 positivity (PNP) (DeLong et al., 2014b; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). Unlike the P600, the PNP appears to be sensitive to contextual constraint

(DeLong et al., 2014b; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). This has led some to propose that the PNP may offer an index of prediction failure (DeLong et al., 2014b; Delong et al., 2011).

One proposed source of the difference between the PNP and the P600 is the nature of the incongruity that triggers each component. It has been noted that completely anomalous words that are impossible to integrate into the given sentence elicit differences in the P600 but not the PNP, whereas the reverse is true of unexpected but plausible words (DeLong, Quante, & Kutas, 2014a; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). For example, the verb *look* would be a relatively unexpected continuation of the sentence *The children went outside to. . .* where the most expected word might be *play*, but *look* is not implausible (Federmeier et al., 2007). This has led to speculation that the PNP reflects a process of *successful* update of the probabilistic representation of a sentence (Kuperberg & Wlotko, 2019). While further research is needed on the precise conditions that elicit the PNP, current findings point to its utility as a much-desired index of prediction failure.

Chapter 3

3 German particle verbs

The research in this dissertation focuses on the predictive processing of German particle verb constructions. German particle verbs present an ideal test case for investigating predictive processing as they form a strong semantic and syntactic dependency. Delaying the appearance of the particle and manipulating its predictability can therefore allow us to probe the conditions under which readers may utilise predictive processing to facilitate reading. In this chapter, I briefly introduce the particle verb construction and discuss the main debate surrounding the syntactic analysis of particle verbs; namely, whether they form a single functional unit in the lexicon or whether verb and particle are combined at the level of syntax. While this is not a question addressed by the research in this dissertation, understanding how particle verbs may be cognitively represented is important to understanding the dependent relationship of verb and particle, and to understanding previous psycholinguistic research on particle verbs.

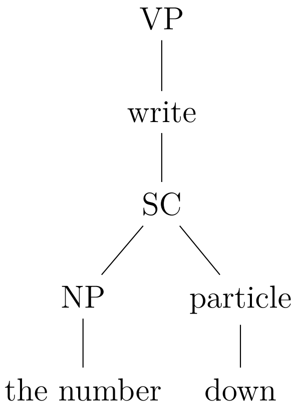
3.1 Definition of a particle verb

The most generally accepted definition of a particle verb is that it is a construction comprising a verb and a particle that can be separated by other elements of a sentence not usually licensed in the middle of a word (Dehé, Jackendoff, McIntyre, & Urban, 2002; Falk & Öhl, 2010; Müller, 2002). In German main clauses, V2 word order dictates that the verb appears in second position. The particle usually appears at the right main clause boundary, although there are exceptions (Müller, 2002). Particles, at least in Germanic languages, are formally and in some cases, semantically, related to prepositions, and form a close relationship with a verb similar to an affix (Dehé et al., 2002). However, what constitutes a “true” particle is somewhat controversial (Dehé et al., 2002). Some accounts propose that particles in German can be formed by any word class including prepositions *untergehen* (to undergo), nouns *achtgeben* (to pay attention to, literally: attention give), adjectives *schönreden* (to sugarcoat, literally: gloss speak), adverbs *sitzenbleiben* (to remain seated, literally: sit stay), and verbs *kennenlernen* (to meet [someone], literally: know learn) (Falk & Öhl, 2010; Müller, 2002). Other accounts are stricter, allowing only prepositions, nouns, and adjectivals such as *ernstnehmen* (to take something seriously, literally: serious take) and excluding morphologically complex, pronominal adverbs such as *hinab* and *heraus*, and the deictic *hin*/*her* prefixes (Zeller, 2001). For the purposes of this thesis, the less strict definition of particle-hood is adopted.

3.2 Syntactic analyses of particle verbs

Further controversy lies in how to syntactically analyse particle verbs, mainly because they exhibit behaviour that resembles both single and multiple words. Dehé et al. (2002) divide the various existing analysis approaches into two broad categories: those that treat the base verb and particle as two separate functional units, and those that treat them as a single functional unit.

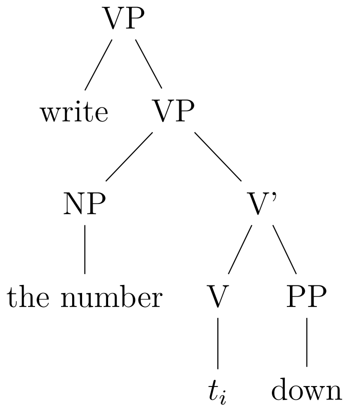
The main argument for analysing particle verbs as separate functional units is the fact that other elements of the sentence can separate the verb and the particle (Booij, 2002). This directly violates the *lexical integrity principle*, which states that a single word does not allow other words to be placed word-medially (Chomsky, 1970). It has therefore been proposed that particle verbs function as secondary predicates, where the particle forms a constituent with the direct object of the base verb (see review in Dehé et al., 2002). For example:



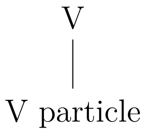
As an argument against a secondary predicate analysis of particle verbs, Müller (2002) describes the construction above as a depictive secondary predicate, similar in relation to an adjunct combined at the level of syntax. Particle verbs, he argues, more closely resemble resultative predicates, which are better analysed as complex predicates because their combination is licensed by lexical rules. For example, in example 9a below, *raw* depicts the state of the meat, a predicate that is the result of combining the elements *meat* and *raw*. In 9b, *black* is licensed by the main verb *paint* and describes the result of that action.

(9) a. Er isst das Fleisch roh.
He eats the meat raw.
b. Sie streicht die Tür schwarz.
She paints the door black.

The syntactic complex predicate approach assumes that verb and particle are a single functional unit where the particle forms a constituent with the verb instead of with the direct object (Müller, 2002). The verb and particle are processed separately until syntactic processing joins them as a phrasal constituent (reviewed in Dehé et al., 2002):



Others view the single verb-particle functional unit as a morphological unit combined pre-syntactically, perhaps in the lexicon (reviewed in Dehé et al., 2002); for example:



One argument in support of particles as bound morphemes is that they can influence the argument structure of their verb in the same way that morphemes can, such as by changing the verb's transitivity (Booij, 2002). However, particles exhibit many more behaviours that are not morpheme-like. For example, particles in German appear at the right sentence boundary, which is not a feature of other German morphological objects (Müller, 2002). It is also not possible to have more than one particle per verb as it is with morphological marking; for example, *los* in *weil Maria loslacht* (because Maria starts to laugh) and *an* in *weil Maria Karl anlacht* (because Maria smiles at Karl) cannot be combined to form *weil Maria Karl losanlacht* (intended: because Maria starts to smile at Karl; Müller, 2002). Furthermore, the base verb of a particle verb can be deleted, whereas the same is not true of a affixed verb: **weil Jens übertreibt und Hans unter* (*because Jens overstates and Hans under: affixed verb) versus *weil Peter einsteigt und Hans aus* (because Peter gets in and Hans out: particle verb). A final argument against a morphological analysis is that particles can be nouns, prepositions, and adverbs (among other things) which, in contrast to morphemes, can be standalone lexical items (Müller, 2002; Zeller, 2001).

An additional argument for treating particle verbs as a single functional unit is their ability to undergo derivational processes - a phenomenon usually restricted to single words and idioms (Cappelle, Shtyrov, & Pulvermüller, 2010). For example, *show off* and *fix up* can be nominalised: *a show-off, a fixer-upper* (Cappelle et al., 2010), while regular verb/adverb combinations cannot: *an arrive early, *a come later, although *a latecomer* is acceptable. In German, the particle verb *wirft jemandem etwas vor* (to accuse someone of something, literally: throw something before someone) behaves in the same way as the prepositional phrase construction in *wirft etwas in den Briefkasten* (throw something in the letterbox; Cappelle et al., 2010). However, *Vorwurf* (accusation) can be derived from *vorwerfen* (to accuse), while *in den Briefkasten-Werfung (in the letterbox throwing) cannot be derived from its verbal phrase construction. The derivational properties of particle verbs also raise a new set of questions about whether a single *functional* unit is also represented by a single *lexical* unit.

One possibility is that there are actually different kinds of particle verbs: “idiomatic” and “combinatorial” (Cappelle et al., 2010; Fraser, 1976). An example of a combinatorial particle verb is *walk in*, whereby something occurs (walking) with the result that you are/get “in” (Cappelle et al., 2010). An idiomatic particle verb would be *give up*, to which the same logic cannot be applied (i.e. something occurs (giving) so that you are “up”). In German, particle verbs can also be distinguished into “true” particle verbs and “particle verbs in a broader sense” (Müller, 2002). True particle verbs are proposed to be “idiomatic”; those where the meaning does not equal the sum of the parts, for example *absagen*, “to cancel” (literally: “off say”). Particle verbs in a broader sense are compositional; for example, *aussagen*, “to state” (literally: “out say”). Müller (2002) proposes that true (idiomatic) particle verbs such as *absagen* are lexicalised whereas combinatorial particle verbs such as *aussagen* are compositionally understood or produced.

The first difficulty in separating particle verbs into two such categories is where to draw the line. In cases such as *figure out* or *slow down*, this line is difficult if one considers the Lakoff & Johnson (1980) conceptual metaphors (summarised in Cappelle et al., 2010). For example, *figure out* could be conceptualised as a solution locked in a box and being calculated “out” of it in order to be seen. With this in mind, the boundary between idiomatic and combinatorial particle verbs becomes blurred. A second potential argument against distinct categories is that the high frequency of combinatorial particle verbs means that they are likely to become lexicalised anyway, as is the case for other high-frequency phrasal constructions (Müller, 2002; Ullman, 2004). One way to determine how particle verbs may be cognitively represented is, of course, with psycholinguistic research methods.

3.3 Psycholinguistic research on particle verbs

Broadly speaking, the main question raised by the syntactic analyses above is whether particle verbs have their own lexical entry or whether they are composed of separate lexical units combined by syntax. This question has been the main focus of empirical research on particle verbs. Priming studies have found that there was no difference in lexical decision times to verbs between primes with idiomatic versus combinatorial meaning, e.g. *mitbringen-Geschenk* (bring with-gift) versus *umbringen-Mord* (kill-murder). Similarly, N400s at target verbs were found to be equally reduced by idiomatic and combinatorial verb primes, although it should be noted that this was not a study of particle verbs per se and included affixed verbs such as *erziehen* (Smolka, Gondan, & Rösler, 2015). An MEG study of the early mismatch component found that participants showed equally large mismatch negativities to unlicensed idiomatic verb-particle combinations (eg. *rise down instead of rise up) and combinatorial combinations (e.g. *fall up instead of fall down), from which it was concluded that both idiomatic and combinatorial types were lexicalised (Cappelle et al., 2010). Similar conclusions have been drawn from analogous structures in Norwegian (Kush, Dillon, Eik, & Staub, 2019).

The above results suggest that both idiomatic and combinatorial particle verbs have lexical representations. However, idiomatic German particle verbs have been shown to demonstrate difficulty at the syntactic as well as the lexical level (Czypionka, Golcher, Błaszczak, & Eulitz, 2019). In this study, non-licensed verb-particle combinations (that is, combinations that cannot have had a lexical entry), elicited late positive ERP components, suggesting possible attempts to combine them at the syntactic level. This suggests that particle verbs may be represented as a single lexical unit, but that it is also possible to combine the two elements at the level of syntax. This is consistent with a study of English particle verbs proposing a gradient between syntactic and lexical representations of English particle verbs (Brehm & Goldrick, 2017).

In this dissertation, rather than examining the cognitive representation of particle verbs, their properties are instead exploited to examine predictive processing. Such an attempt has been made using Dutch particle verbs, which were used to investigate the presence of long-distance predictions (Piai et al., 2013). This study is discussed in more detail elsewhere within the thesis, but in essence, determined that plausible verb-particle options were not preactivated by verbs that could take particles. Instead, the findings suggested that verbs identified as potentially taking a particle were maintained in working memory to facilitate retrieval of the particle should it be encountered. A valuable outcome of this and other ERP work on particle verbs was to demonstrate that verb particles are able to elicit N400s in the same way as content words such as nouns (Czypionka et al., 2019; Piai et al., 2013). This result was not necessarily a given, since there is some doubt as to whether the N400 can be elicited by function words (Brown, Hagoort, & Keurs, 1999; Frank, Otten, Galli, & Vigliocco, 2015; but cf. Van Petten & Kutas, 1991). On the other hand, syntactic analyses of particle verbs propose that particles do actually behave as content words. While this has interesting consequences for the syntactic theory of particle verbs, I leave that to future research and concentrate in the following experiments on the predictive processing of separated particles.

Chapter 4

4 Preactivation and decay in long-distance verb-particle dependencies

The contents of this chapter have been submitted for publication in the journal PeerJ in collaboration with Titus von der Malsburg and Shravan Vasishth.

4.1 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.

4.2 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 & 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; Husain et al., 2014; Lewis & Vasishth, 2005). 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 (Ferreira & Henderson, 1991; Gibson, 1998; Lewis & Vasishth, 2005; Van Dyke & Lewis, 2003; Vasishth & 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 (Ness & Meltzer-Asscher, 2019; e.g. Xiang, Dillon, Wagers, Liu, & Guo, 2014) and an empirical study has demonstrated its effects over and above that of interference (Van Dyke & 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, Jäger, & Vasishth, 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 & Vasishth, 2005) or new discourse referents (Gibson, 1998, Gibson (2000)). In addition, we tested whether higher lexical predictability may make a word more resistant to decay.

The LV05 model (Lewis & 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 & 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 & 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 & Vasishth, 2005), although discourse referents may not be the only source of slowing in longer dependencies (Gibson & 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. *spannen*, to tense'', which can take at least seven particles) should involve a larger demand on working memory than verbs with a small set size (e.g \textit{kleuren},to colour”, which can take only two particles). 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 & 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 (Konieczny & Döring, 2003; Levy, 2008). 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 & 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.

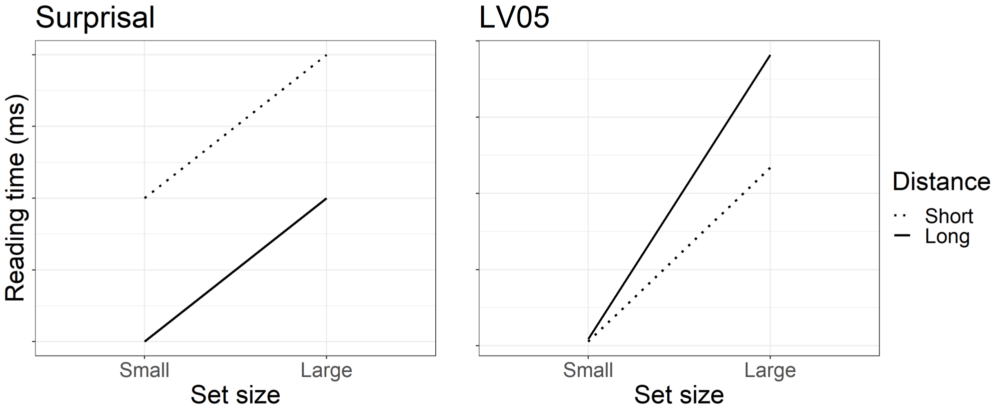


Figure 1: Interaction of set size and distance predicted by the surprisal and LV05 models.

4.3 Experiment 1

4.3.1 Methods

4.3.1.1 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.

4.3.1.2 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 10 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:

(10) a. Small set/short distance:

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

4.3.1.2.1 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 95% confidence intervals of the 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):

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

	Cloze probability		Entropy	
Condition	**Mean**	**95% CI**	**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: Results of the cloze test for the final set of 24 items.

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 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.

4.3.1.2.2 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

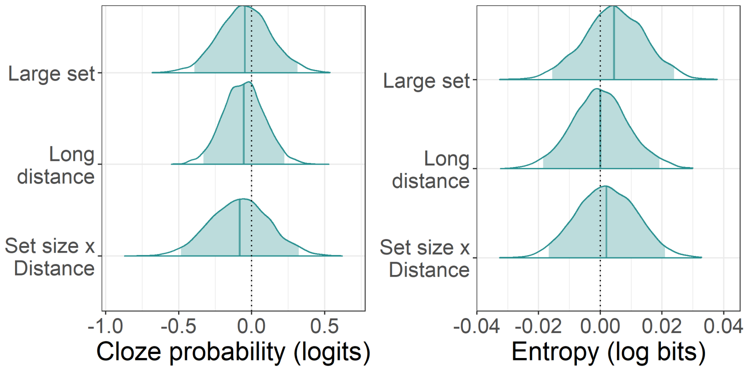


Figure 2: Change in cloze probability and entropy associated with condition. The posterior distributions are those for large set size and long distance relative to the grand mean of each condition (the dotted line). The posteriors for the small set size and short distance conditions can therefore be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

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.

	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.

4.3.1.2.3 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. 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:

(11) a. Small set:

Die Hausfrau sagte, dass sie mit dem neu gekauften Lappen die Teller
The housewife said, that she with the newly bought rag the plates
in der Küche ***abschrubbte****, um Platz zum Kochen zu schaffen.*
in the 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
The housewife said, that she with the newly bought rag the plates
in der Küche ***abspülte****, um Platz zum Kochen zu schaffen.*
in the 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.

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.

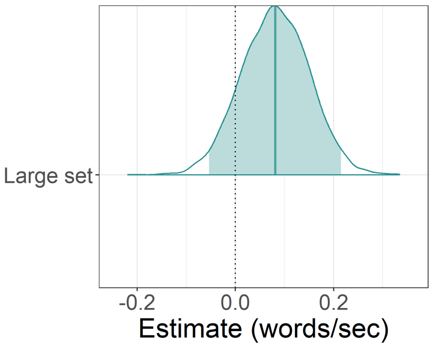


Figure 3: Change in self-paced reading time in the online norming study. The curve is the posterior distribution associated with the large set condition relative to the grand mean of large and small set conditions (dotted line). Due to the reciprocal transform, a shift in the posterior to the right of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

4.3.2 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 10 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.

4.3.3 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 & 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, Kambe, & Duffy, 2000) might interact with reading times in the spillover region. Instead, we present mean reading times across the sentence in Figure 9, where no spillover effect is apparent. 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 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.

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 (M. Lee & Wagenmakers, 2013).

4.3.4 Results

4.3.4.1 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.

4.3.4.2 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.

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 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 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).

4.3.4.3 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

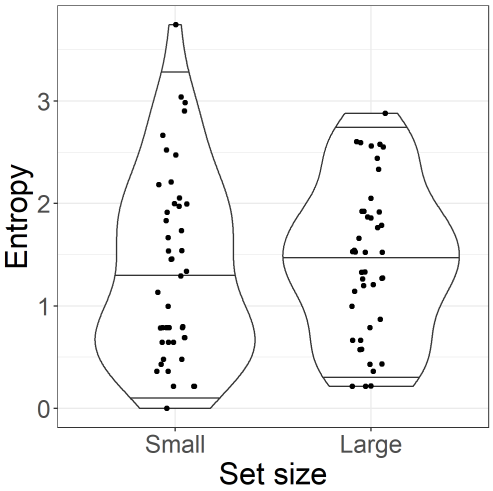


Figure 4: By-item entropy within small and large set categories. Violin plots show the median and 95% quantiles.

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.

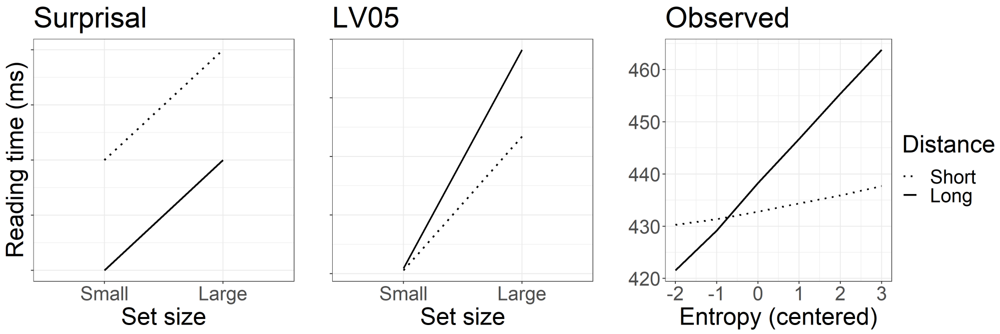


Figure 5: Predicted versus modelled self-paced reading times. Note that this figure represents only a predicted reading time pattern based on the model output and that there was no statistical support for the interaction of distance and entropy.

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.

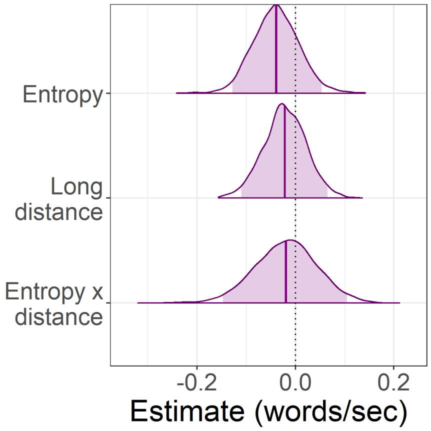


Figure 6: Change in self-paced reading speed at the particle with entropy as a continuous predictor. Now that entropy is a continuous predictor, the posterior represents the change in reading time elicited by a 1-unit increase in entropy. Due to the reciprocal transform, a shift in the posterior to the 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.

4.3.5 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 & 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, 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.

4.4 Experiment 2

4.4.1 Methods

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.

4.4.1.1 Participants

In line with the power analysis reported in the self-paced reading section above, 60 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.

4.4.1.2 Materials

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

4.4.1.3 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.

4.4.1.4 Analysis

Sampled data were exported from DataViewer (SR Research) and pre-processed in R using the `em2` package (Logačev & 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 (Boston et al., 2008; Kliegl et al., 2004), 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. As for the self-paced reading experiment, the spillover region was not analysed, but is plotted in Figure 9. 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:

$$\beta_0 \sim Normal(5.7, 0.5)$$
$$\beta_{1,2,3} \sim Normal(0, 0.5)$$
$$\upsilon \sim Normal(0, \sigma_\upsilon)$$
$$\gamma \sim Normal(0, \sigma_\gamma)$$

$$\sigma_v, \sigma_\gamma \sim Normal_+(0,1)$$
$$\rho_v, \rho_\gamma \sim LKJ(2)$$
$$\sigma \sim Normal_+(0,1)$$

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.

4.4.2 Results

4.4.2.1 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.

4.4.2.2 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 & 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.

4.4.2.3 Exploratory analysis: 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* × *distance*) were in favour of the null hypothesis.

				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.

4.4.3 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.

4.5 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

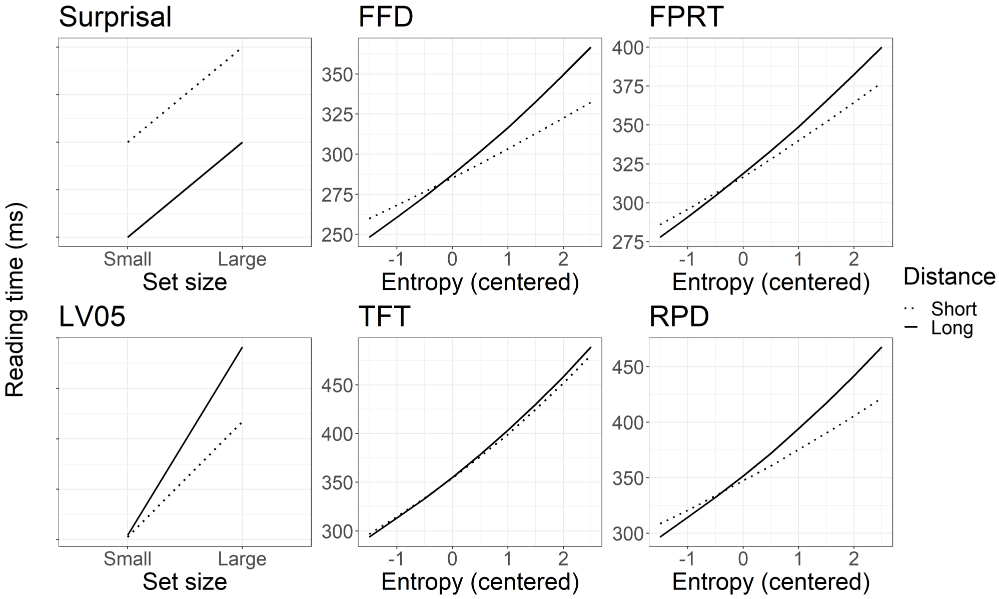


Figure 7: Predicted versus modelled interaction of entropy and distance. Note that this figure represents only predicted reading time patterns based on the model output and that there was no statistical support for the interaction of distance and entropy.

entropy in eye tracking. 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 & 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.

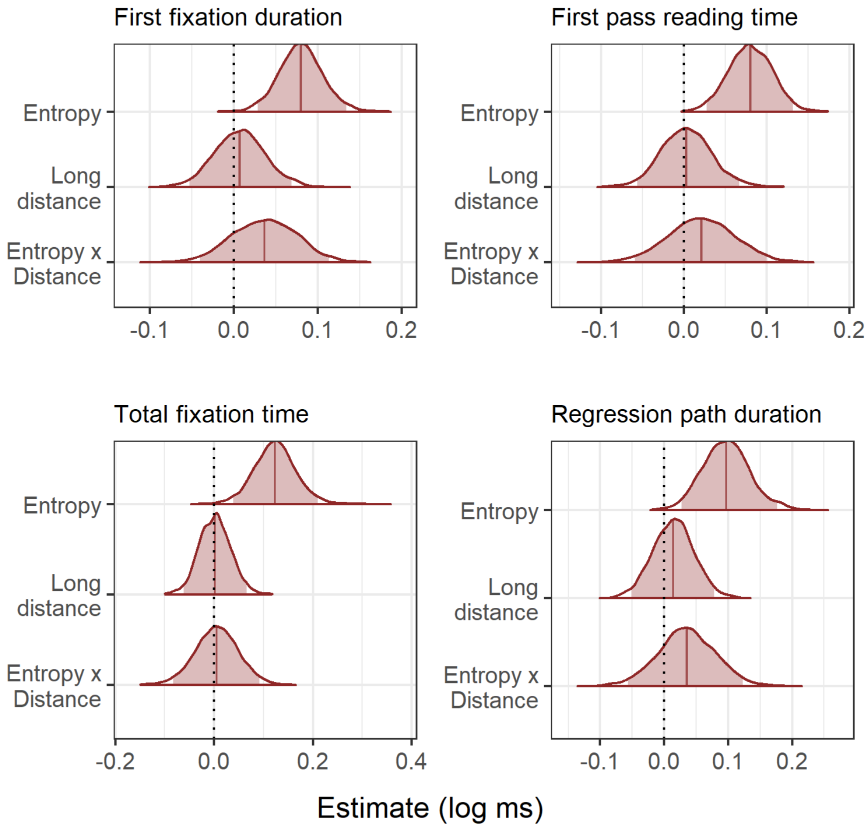


Figure 8: Changes in reading time for each eye-tracking measure using entropy as a continuous predictor. Now that entropy is a continuous predictor, the posterior represents the change in reading time for the average reader elicited by a 1-unit increase in entropy. The log transformed reading times mean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

4.6 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 & 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.

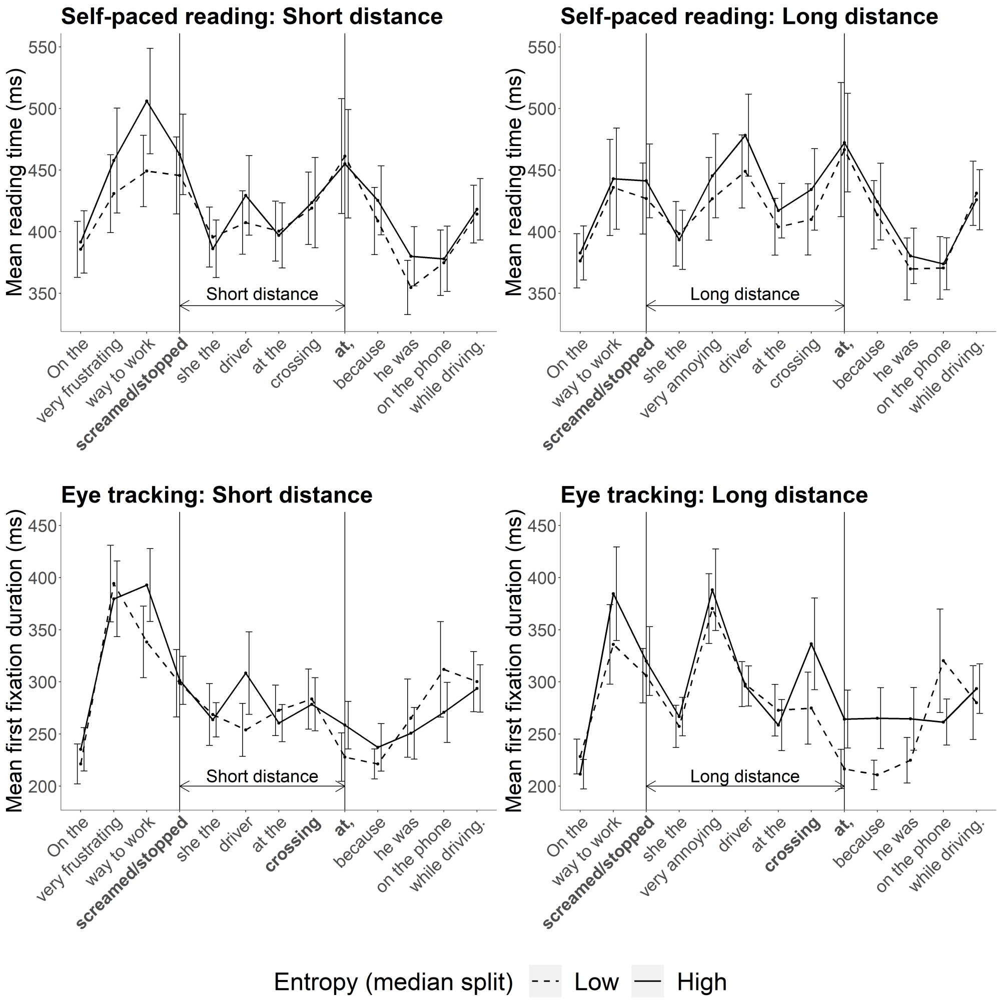


Figure 9: Comparison of self-paced reading and total fixation times plotted across the sentence. Error bars show 95% confidence intervals.

4.6.1 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.

4.6.2 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 (Engelmann et al., 2019; Lewandowsky et al., 2009; 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.

4.6.3 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 & 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 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.

4.7 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 & 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.

				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.

	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.

Chapter 5

5 Long-distance lexical predictions in verb-particle constructions

5.1 Abstract

In the previous chapter, an eye tracking experiment suggested that when native German readers encounter a verb that is likely part of a verb-particle construction, they may preactivate the lexical entries of plausible particles. However, the eye tracking experiment did not allow us to conclude whether readers committed to a *specific* particle (made a lexical prediction). In the current chapter, this question was investigated using event-related potentials (ERPs). ERP has the advantage over reading time studies of being able to measure cognitive processes that occur even before eye movements are planned. As discussed in *Chapter 3*, several ERP components have been linked to language processing which be utilised to interrogate the presence of lexical predictions. Previous ERP evidence for lexical predictions has been provided by studies showing larger amplitude N400s for determiners that do not match a predicted next word (e.g. seeing *an* when *kite* is expected). However, the bulk of ERP evidence for lexical prediction is limited to predictions about adjacent words. While there is evidence that structural (syntactic) predictions can span longer distances, evidence for specific lexical predictions over long distances is limited. In this chapter, an ERP experiment tested the hypothesis that in German sentences where a particle, e.g. *auf* (up), appeared at some distance downstream from its verb, e.g. *räumen* (tidy), the identity of the particle could be predicted at least two words in advance, but only if its cloze probability was high. The study found that violations of particles that were almost 100% probable were associated with a larger re-analysis cost (late post-N400 ERP positivity) than violations of particles that were highly probable, but whose identity was made uncertain by the presence of a strong competitor particle. Statistical evidence for the late positivity was inconclusive, however; a possible reason for this is discussed and a solution proposed.

5.2 Introduction

Predicting upcoming linguistic information allows us to more quickly assimilate that information once it appears and aids processing in noisy environments. Characterising the way linguistic predictions are generated offers a key insight into how the features of words are associated and how propagating activation through these features may facilitate comprehension. Some evidence suggests that linguistic predictions may even influence our perception of a word once it is seen (Lupyan & Clark, 2015). The precise depth of linguistic predictions and the mechanisms that control predictive processing during reading, however, remain largely unclear. In the current study, we attempt to demonstrate that predictions can be made for specific words over long distances during reading, but that such predictions may be discouraged if the chance of success is not sufficiently high.

The ERP components most often used in the study of predictive processing are the N400 and the late positivity. The N400 is a negative spike in the ERP occurring at around 250-500 ms after a word is first seen and is highly correlated with a word's congruency within the sentence context (Kutas & Federmeier, 2011). The N400 becomes smaller at each new, context-congruent word in a sentence (Payne et al., 2015; Van Petten & Kutas, 1990). Conversely, words not congruent with a given context increase the amplitude of the N400 (Kutas & Federmeier, 2011; Van Petten & Luka, 2012). This dichotomy has led to the hypothesis that sentence context allows the preactivation of probable words that are then easier to process once encountered, and that non-preactivated words incur a processing cost (Kutas & Federmeier, 2011). Following the N400 for unexpected words, a positive deflection in the ERP is sometimes observed, peaking at around 600-900 ms (Van Petten & Luka, 2012). When this positivity has a posterior spatial distribution, it is referred to as the P600 or the late positivity and is associated with syntactic violations, possibly reflecting failed attempts at analysis (DeLong et al., 2014a; Federmeier et al., 2007; A. Kim & Osterhout, 2005; Osterhout & Holcomb, 1992; Van Petten & Luka, 2012). When the positivity has a frontal distribution, it has been referred to the post-N400 positivity (PNP). The PNP has been associated with the cost of reanalysing an unpredicted but contextually congruent word (DeLong et al., 2014a; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Berkum et al., 2005; Van Petten & Luka, 2012). Taken together, these findings on the N400, the P600, and the PNP suggest a processing advantage of preactivation and a processing cost for non-preactivated words.

Preactivation of upcoming words is thought to occur after the context of a sentence has triggered high-level associations such as event structure (Kuperberg & Jaeger, 2016). This high-level information is used to predictively preactivate lower level features such as semantic and morphosyntactic information (Federmeier & Kutas, 1999; Luke & Christianson, 2016; Metusalem et al., 2012). With sufficient contextual information, preactivation may even spread to probable lexical items, including bottom-level features such as phonological form. Evidence that bottom-level, form-based features have been preactivated is therefore taken as a sign that prediction of a full lexical item has been made (DeLong et al., 2005; Szewczyk &

Schriefers, 2013; Van Berkum et al., 2005; Wicha et al., 2004). For example, a larger N400 observed at the unexpected determiner *an* when the most predictable next word is *kite* is taken as evidence that the lexical item *kite* has been predicted because its phonological form has already been activated and expectations about the determiner generated (DeLong et al., 2005). Such an effect has also been found when using unexpected gender and animacy inflections on an adjective preceding a predicted noun (Szewczyk & Schriefers, 2013; Van Berkum et al., 2005; Wicha et al., 2004). Evidence supporting preactivation of low-level form and morphological features has recently been contested, however (Kochari & Flecken, 2019; Nieuwland et al., 2018), although support for a smaller effect size consistent with the original claim has been proposed (Nicenboim et al., 2019). There is therefore some evidence that lexical predictions triggered by context facilitate not only integration of the predicted word, but also the words immediately preceding it.

A facilitatory effect of preactivation on the determiners and adjectives of a predicted noun demonstrates the processing benefit of prediction in a very localised environment. Presumably, the benefit of prediction extends further than this, especially when one considers that not all dependent elements of a sentence appear adjacently. Studies of context-based anticipation have shown that discourse constraints may allow the pre-activation of words across sentence boundaries (M. Otten & Van Berkum, 2008; M. Otten, Nieuwland, & Van Berkum, 2007). It is difficult to distinguish in these cases, however, whether anticipation involves the preactivation of specific words, a 'lexical prediction', or simply the features of a plausible word category. In contrast to feature preactivation, a lexical prediction could allow a predicted word to be integrated into the sentence parse long before the word itself is seen. This could enable other chunks of the sentence appearing before the predicted word to be fully interpreted earlier than if the parser made no prediction at all. Evidence that predictions enable such early processing may include the recovery cost observed for disconfirmed predictions (DeLong et al., 2014a; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Berkum et al., 2005; Van Petten & Luka, 2012). In these studies, it is unlikely that the cost of recovering from a misprediction was simply that the predicted word alone was discarded, since this should be relatively inexpensive. If, however, predictions are not just maintained but also *used* to aid interpretation of the sentence, predicting the wrong word would pose a much higher cost. Despite this risk, in some circumstances early integration of a predicted word could be of great benefit to the processing load of the parser, especially in long-distance dependencies such as the German verb-particle construction in (12):

(12) *Der Professor* ***fuhr*** *mit seinem Vortrag trotz regelmäßiger Störungen*
The professor **carried** with his lecture despite regular interruptions

fort.
on.
The professor carried on with his lecture despite regular interruptions.

The exact meaning of the base verb *fuhr* (to carry) is dependent on the particle *fort* (to carry on). Other particles congruent with the beginning of the sentence *Der Professor fuhr...* could include *ab* (to drive off), *zusammen* (to startle), or *zurück* (to reverse something or cut something back). The verb *fuhr* frequently appears in combination with a particle in German, which may mean that seeing *fuhr* triggers preactivation of its licensed particles. A strong expectation for the exact identity of the particle may be generated later in the sentence; for example, it becomes clear at the object *his lecture* that the most likely particle is *fort.* A lexical prediction for the item *fort* could therefore be made, giving early access to the verb semantics of *carry on with [something].* The desire for early access to verb semantics is high in languages in which a verb canonically appears in second position. This makes separable verb-particle constructions an excellent test of long-distance predictions because unless the particle is predicted in advance, the reader is forced to wait until the end of the sentence to know the exact meaning of the verb.

Particle verbs have previously been used to investigate the presence of long-distance predictions in Dutch (Piai et al., 2013), as have similar complex predicate constructions in Hindi and Persian (Husain et al., 2014; Safavi et al., 2016). In reading time studies of Hindi and Persian, it was found that distance, the strength of the predictability of the head-final verb, and the type of intervening information may determine whether or not a lexical prediction facilitates reading (Husain et al., 2014; Safavi et al., 2016). An ERP study of Dutch particle verbs included the hypothesis that Dutch verbs that can take a large number of possible particles (e.g. *spannen* ‘to tense’, which can take seven particles) may place a larger demand on working memory than verbs with a small set size (e.g. *kleuren* “to colour”, which takes only two particles; Piai et al., 2013). In other words, pre-activation of a larger number of particles at the base verb could result in a larger working memory load for large-set verbs as more particles must be maintained in working memory until the dependency can be resolved. A larger left anterior negativity (LAN) was observed at verbs that took particles versus verbs that never took a particle, but there was no evidence that the number of particles a verb took affected LAN amplitude. This was interpreted as evidence that the particles themselves were not preactivated at the base verb, but rather that verbs identified as potentially taking a particle were maintained in working memory to facilitate retrieval of the particle should it be encountered. In the current study, we approach the question of long-distance verb particle predictions from a different angle.

The current study design is informed by Piai et al.’s (2013) study of Dutch particle verbs;

however, we extend it in a number of ways. A valuable outcome of Piai et al.'s (2013) study was the demonstration that verb particles were able to elicit N400s in the same way as content words such as nouns. This result was not necessarily a given, since there is some doubt as to whether the N400 can be elicited by function words (Frank et al., 2015). On the other hand, syntactic (but not morphological) analyses of particle verbs propose that particles do actually behave as content words (for a discussion of particle verb analyses see Dehé et al., 2002). In either case, Piai et al.'s findings allow us to focus on the ERP components more robustly studied in relation to prediction, the N400 and the late positive component. Additionally, we use the particle instead of the verb as the target region and compare only target regions that contain the same word. The target regions are also preceded and followed by identical words between conditions. This means that the target region will not be differentially affected by frequency, length, or lexical associations, or by ERPs from the previous or the next word. The example item (13) below sets out the experimental design.

We constructed sentences using German particle verbs in which the base verb and its particle were separated. With this design, we sought not only test whether preactivation of the verb particle was occurring, but whether a *lexical prediction* for the identity of the particle was being made. We therefore compared two prediction-encouraging sentences and manipulated uncertainty about the identity of the particle. This was achieved by constructing sentences that constrained the set of plausible particles to either just one (condition a/b), or to a small set of two or more particles (condition c/d). The example item (13) shows the base verb and its particle(s) in bold font. The pre-critical and spillover regions are underlined. Critically, at least two of the plausible particles in (c/d) had comparable cloze probability, i.e. there was close competition for the identity of the particle. We hypothesised that readers would be more likely commit to a lexical prediction when there was only one highly probable particle (a/b) than when there were multiple particles (c/d).

(13) a. 1 plausible particle/plausible:

Der ordentliche Professor ***fuhr*** *mit seinem Vortrag trotz regelmäßiger*
The orderly professor **carried** with his lecture despite regular
Störungen immer ordnungsgemaß ***fort****, da er für seine Unaufgeregtheit*
interruptions always properly **on**, as he for his unflappability
bekannt war.
known was.

The orderly professor carried on with his lecture despite regular interruptions always as directed, as he was known for being unflappable.

b. 1 plausible particle/violation:

Der ordentliche Professor **fuhr mit seinem Vortrag trotz*
The orderly professor **carried** with his lecture despite
*regelmäßiger Störungen immer ordnungsgemaß **wahr**, da er für seine*
regular interruptions always properly **true**, as he for his
Unaufgeregtheit bekannt war.
unflappability known was.

The orderly professor carried true with his lecture despite regular interruptions always as directed, as he was known for being unflappable.

c. 2+ plausible particles/plausible:

*Der ordentliche Buchhalter **fuhr** seinen zuverläßssigen Computer bei der*
The orderly accountant **turned** his reliable computer at
*Arbeit immer ordnungsgemaß **herunter (hoch)**, da er für seine korrekte*
work always properly **off/on**, as he for his correct
Arbeitsweise bekannt war.
work practices known was.

The orderly accountant turned on/off his reliable computer at work always as directed, as he was known for his faultless work practices.

d. 2+ plausible particles/violation:

Der ordentliche Buchhalter **fuhr seinen zuverläßssigen Computer bei*
The orderly accountant **turned** his reliable computer at
*der Arbeit immer ordnungsgemaß **wahr**, da er für seine korrekte*
work always properly **true**, as he for his correct
Arbeitsweise bekannt war.
work practices known was.

The orderly accountant turned true his reliable computer at work always as directed, as he was known for his faultless work practices.

The most obvious approach to assessing the ERPs elicited by the particle would have been to test for a 2 × 2 interaction showing that there was a greater difference in ERP amplitude for (a) vs. (b) than for (c) vs. (d). However, using the grammatical conditions (a/c) in an analysis would be problematic because the particles in these two conditions were not matched for identity, which would have introduced frequency, length, and other semantic confounds. Second, the cloze probability of the single plausible particle in condition (a) would naturally be higher than the cloze probability split between the two or more plausible particles in condition (c). Cloze probability is known to be closely associated with the amplitude of the N400 and the late positive component (Kutas & Federmeier, 2011; Van Petten & Luka, 2012). Since the research question concerned the recovery cost of disconfirmed predictions, we therefore compared only the two violation conditions (b/d) where the presented particle was not compatible with the context.

The ‘violation’ particles were carefully selected so that they were completely implausible in the given context. Their cloze probability was therefore matched at zero, as was their ability to be integrated into the sentence parse. Any difference in ERP amplitude at the violation particles could therefore have only been due to some aspect of dealing with the violation that differed due to processing that occurred *before* the particle. Furthermore, since at least two words before the particle were matched, any difference observed at the particle cannot be due to processing occurring in the immediately preceding region. While this design did not allow us to pinpoint exactly where in the sentence a lexical prediction may have been made, a difference in ERP amplitude at the particle allowed us to infer that a prediction was committed to at the latest before the pre-critical region.

5.2.1 Predictions

The pre-registered hypotheses and predictions concerned the N400. We predicted that the particle violation would cause greater surprise and a larger N400 when commitment to a specific lexical prediction was violated (b) than when presumably no commitment had been made (d). However, the 1 vs. 2+ particle manipulation was similar to constraint manipulations in previous research that have indicated the N400 is sensitive only to the congruency of a word in its context and not to the strength of that context. This would have predicted no difference in the N400 between our conditions (b) and (d). On the other hand, previous studies of constraint did not quantify the amount of competition between specific lexical items in lower constraint conditions as we did, and we hypothesised that competition could have dampened the N400. These predictions and the analysis plan were pre-registered on OSF: https://osf.io/qbna2

Based on previous research linking it to failed predictions in strongly constraining contexts, the ERP component of more interest was actually the post-N400 positivity (DeLong et al., 2014a; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Berkum et al., 2005; PNP, Van Petten & Luka, 2012). The PNP has been found to be larger for disconfirmed predictions in higher than lower constraint sentences, meaning it would be expected to be larger in our commitment condition (b) than our no-commitment condition (d), as (b) is the higher constraint sentence. The analyses of these hypotheses are therefore presented as the pre-registered analysis of the N400 and an exploratory analysis of the PNP.

5.3 Experiment 3

5.3.1 Methods

5.3.1.1 Participants

All participants were recruited via an in-house, online recruitment database to ensure participants who participated in the cloze test were not re-recruited. For reasons of practicality (e.g. the need to use this database, ease of access to the campus, university regulations requiring physical signatures to provide reimbursement), the majority of participants were university students. In line with university policy, all participants were reimbursed for their time either financially or in the form of credit points toward their studies. All participants were right-handed German native speakers, with no history of developmental or current language, neurological, or psychiatric disorder. All participants provided written consent to participation in the study.

54 participants were recruited who matched the inclusion criteria. Four participants were excluded because >75% of their target EEG segments were contaminated by excessive muscle and/or blink artefact. This left a total of 50 subjects (6 male), with a mean age of 25 years (range = 17 to 40 years, SD = 5 years).

5.3.1.2 Materials

For each particle verb, two sentences were constructed. The base verb appeared 3-4 words from the beginning of the sentence, while the particle appeared further downstream, as can be seen in (2). The position of the base verb and particle in each sentence pair was matched. The particle formed the target region where ERPs would be measured. Within each sentence pair, at least two words before the particle were identical. 103 sentence pairs were constructed and presented as a cloze test to 30 German native speakers (mean age 25 years, SD 6 years, range 18-41 years) on a desktop computer in our in-house lab using Ibex software (Drummond, 2016). The sentence pairs were divided into two lists, such that each participant only saw one sentence from each item. The particle of the sentence was replaced by a gap, which participants were asked to fill with the first word that came to mind.

The items were then ranked in terms of how well they fulfilled the criteria that condition (a) elicited only one particle and condition (c) elicited at least two particles with relatively similar probability. To rank the items, cloze probabilities and 50% highest probability density intervals (HPDIs) were calculated for each particle completion. Other kinds of completions were grouped into categories (e.g. prepositional phrases, adjectives, nouns) and a cloze probability and HPDI was calculated for each category. The HPDI was used instead of a confidence interval as some cloze probabilities were close to 100%, meaning that their distributions may have been left skewed, making a confidence interval somewhat misleading. HPDIs calculate the area where, in this case, the highest 50% of the probability density lies, regardless of distribution. Items were then ranked by entropy among the responses (lowest to highest condition a; highest to lowest condition c). For condition (c), further weight was given to items where the first two particles given were closer in cloze probability. This ranking scheme left a final set of 44 plausible items fulfilling the criteria of the experiment. The cloze statistics are summarised in Table 14.

	Target particle		**Difference between 1st- and 2nd-BC**	
Condition	Mean	95% CrI	Mean	95% CrI
1-particle	0.90	0.74, 1.00	0.73	0.64, 0.85
2+particle	0.54	0.53, 0.56	0.29	0.16, 0.39

Table 14: Cloze probability summary statistics for the plausible conditions. 1st- and 2nd-best completions (BC) refer to the highest and second-highest cloze particles at the target site.

To create the violation conditions, two German native speakers selected particles that were not possible in the sentence context, including illicit verb-particle combinations. Care was taken to ensure that, either through regional differences or creative thinking, the violation particles were not able to be integrated into the sentence. These violation particles were never elicited by the cloze test. The same particle was used in both violation sentences within each item (b and d), meaning that word length, frequency, and cloze probability (i.e. zero) were matched between the violation conditions. The example item (13) above shows all four target conditions. The base verb and its particle(s) are in bold font, and the pre-critical and spillover regions are underlined.

In addition to the 44 target items, 62 more general filler sentences were randomly interspersed. These filler sentences comprised of sentences with a similar syntactic structure to the target sentences but where the verb and particle were adjacent, sentences of a similar length with no particle verbs, and a small proportion of short, simple sentences. Each participant therefore saw a total of 108 sentences during the testing session. After each target or filler sentence, participants answered a yes/no question about the proposition of sentence. These questions targeted different regions of the sentence.

The final 44 items were split into four lists in a Latin square design, such that each participant only saw 1 out of the 4 conditions for each item. The order of presentation of sentences within each list was fully randomised by the presentation software, Open Sesame (Mathot, Schreij, & Theeuwes, 2012).

5.3.1.3 Procedure

Participants were seated in a shielded EEG cabin at distance of approximately 60 cm from a 56 cm presentation screen. The experimental paradigm was built and presented to participants using Open Sesame (Mathot et al., 2012). Each experimental session began with an instruction screen advising participants that they would read sentences presented word-by-word and that after each sentence, they would answer a question. Yes/no answers were given via two respective buttons on a video game controller. Each question displayed a reminder as to which button corresponds with Yes (left finger) and which with No (right finger). Participants were instructed to answer as quickly and accurately as possible. Each experimental session began with four practice trials.

Each trial in the experiment began with a 500 ms fixation cross in the centre of the screen followed by a blank screen jittered with a mean of 1000 ms and standard deviation 250 ms. Each sentence was presented word-by-word for a duration of 190 ms per word plus 20 ms for each letter. The target word, however, was presented for 700 ms regardless of length. The inter-stimulus interval was 300 ms. After each sentence was completed, a yes/no question appeared; for example, *Verlief der Vortrag ungestört?* (did the lecture proceed uninterrupted?). Answering the question via the video game controller triggered the beginning of the next trial. The order of presentation of sentences within each list was fully randomised by the presentation software. Breaks were offered after every 30 sentences. The testing session including EEG setup lasted approximately two hours.

5.3.1.4 EEG recording and preprocessing

The EEG recording was made in the Department of Linguistics at the University of Potsdam, Germany, in a purpose-built EEG cabin using a 32-lead system and electrodes arranged on the head according to the international 10-20 system. EEG was recorded at a sampling rate of 512 Hz and online filtered with a band pass of 0.01-30 Hz.

Raw EEG recordings were downsampled offline in BrainVision Analyzer 2, Version 2.1.2, to 500 Hz for ease of interpretation. Zero phase shift IIR Butterworth filters were applied at a low pass of 0.01 Hz (order of 2, time constant of 15.92) and a high pass of 30 Hz (order of 2, no time constant). A notch filter was applied at 50 Hz. The full recording was then segmented into epochs from sentence onset to question onset. Ocular correction was then applied to the sentence epochs using automatic independent component analysis (ICA) with a meaned slope algorithm. The reference electrodes were two electrodes placed at the left outer canthus and above the left eye to record horizontal and vertical eye movements, respectively. All channels were corrected except the two mastoids and the two ocular movement channels using restricted Infomax. The bound number of blinks was 60 with a convergence bound of 1E-07. The number of ICA steps was 512. Components were found using sum of squared correlations with the horizontal and vertical ocular electrodes. The total value to delete was 30%.

The corrected segments were further segmented into 1200 ms epochs representing a period of 200 ms before the onset of the target word (the particle), and 1000 ms after onset. EEG segments with muscle artefact or irreparable eye-blink or -movement artefact were automatically marked for 200 ms before and after each respective artefact, defined as exceeding:

- a maximum voltage step of more than 50 μV,
- a maximum absolute difference of 200 μV in a 100 ms interval,
- a minimum amplitude of -100 μV,
- a maximum amplitude of 100 μV,
- and a minimum low activity of 1 μV in a 100 ms interval.

Marked segments were then reviewed and manually discarded if they appeared to be muscle-related or to reflect a technical issue such as gel-bridging or poor electrode contact. Data from individual electrodes was excluded if impedance exceeds 20 $k\Omega$ during the experiment. Whole participants were excluded if they demonstrated a lack of concentration throughout the experiment as indicated by chance-level accuracy to experimental questions, due to technical problems, or if more than 75% of their recording was affected by muscle artefact. The data were then exported and baseline correction and statistical analysis conducted using the R package `eeguana` (Nicenboim, 2018).

5.3.1.5 Pre-registered analysis

As discussed above, only the two violation conditions were analysed, i.e. conditions (b) vs. (d). A linear mixed effects model with full variance-covariance matrices estimated for the random effects of subject and item was fitted using the `brms` package for R (Buerkner, 2017). The dependent variable for the N400 model was mean amplitude in the time windows 250 to 500ms at electrode *Pz*. It is admittedly more common in ERP analysis to average amplitude over a number of electrodes or to test all electrodes individually and correct for multiple comparisons. The former approach increases the chance of detecting a ‘significant’ effect by averaging out variance, but in doing so, ignores real sources of variation in the data. The latter approach is highly susceptible to Type 1 error and potentially less sensitive, depending on how strict the correction is. All three approaches (including ours) have their disadvantages, however we felt that fitting models with full random effects structure at a single electrode was a good compromise. The predictor 'number of candidate particles' was effect contrast coded: -0.5 (1-particle, condition b), 0.5 (2+particle, condition d). Correlated varying intercepts and slopes were modelled by subject (γ) and by trial (υ). The priors were as follows:

$$\begin{gathered}
\beta_0 \sim Normal(0, 10) \\
\beta_1 \sim Normal(0, 5) \\
\upsilon_{0,1} \sim Normal(0, \sigma_\upsilon) \\
\gamma_{0,1} \sim Normal(0, \sigma_\gamma) \\
\sigma_\upsilon, \sigma_\gamma \sim Normal_+(0, 5) \\
\rho_\upsilon, \rho_\gamma \sim LKJ(2) \\
\sigma \sim Normal_+(0, 5)
\end{gathered}$$

Selection of the model priors was made on the basis that the effect size of the N400 can be quite varied between ERP studies. Also taken into account was the fact that, at the time of pre-registration, it was not common practice in ERP research to report error terms of individual effect size estimates, as noted by Van Petten and Luka (2012); although more recent studies have begun to report such parameters (Kochari & Flecken, 2019; Nicenboim et al., 2019; e.g. Nieuwland et al., 2018). We therefore chose an uninformative, regularising prior for the effect of number of particles, β_1. This prior reflected the assumption that the difference in amplitude between conditions was unlikely to be more than 10 μV in either direction. The standard deviation of by-subject and by-item adjustments to the slope and intercept were represented by the priors for υ and γ. Correlations between the intercept and slope adjustments were denoted by the prior for ρ, with an eta of 2 favouring small correlations. Finally, any remaining variance not captured by the other parameters was reflected in the prior for σ.

A prior predictive check, i.e. simulating data using the selected priors (Schad, Betancourt, & Vasishth, 2019), indicated that the priors were able to generate plausible estimates. A posterior predictive check of the model, i.e. simulating data using the posterior estimates, gave a good fit to the observed data. The prior and posterior predictive checks are presented in Appendix 5.6.

The final model specified in `brms` is displayed below Note that 'part' refers to the predictor 'number of particles' and 'gram' to the grammaticality condition, where 'gram == 0.5' means that only the violation conditions have been subset:

	Accuracy		Response time	
Condition	Mean (%)	95% CI	Mean (ms)	95% CI
a	91	89, 93	1,806	1,736, 1,878
b	90	87, 92	1,907	1,829, 1,990
c	96	94, 97	1,885	1,814, 1,959
d	91	89, 93	1,842	1,772, 1,915

Table 15: Mean accuracy and response times.

```
m.n400 <- brm(formula = mean.amp ~ part + (1+part|recording) + (1+part|item),
              subset(df_N400, gram==0.5),
              family = gaussian(),
              prior = priors,
              iter = 2000,
              chains = 4,
              save_all_pars = TRUE,
              control = list(adapt_delta = 0.99)
              )
```

5.3.2 Results

5.3.2.1 Accuracy and reaction times

Mean accuracy and response times for questions about the target sentences were comparable across conditions, as can be seen in Table 15.

5.3.2.2 EEG data exclusion

Of the 500 target trials used in the statistical analysis from the 50 included subjects, 6.37% were excluded due to artefact and 0.46% were excluded due to question response times over 10 seconds (indicating a technical problem). Three target trials (<0.01% of all trials) were not recorded due to experimenter error.

5.3.2.3 Pre-registered N400 analysis

All analyses, planned and exploratory, are summarised for conciseness together in Table 16. The model posteriors are likewise plotted together in Figure 14. A summary of the results for each analysis follows. The final participant pool and number of items included in all analyses deviated slightly from the pre-registration:

1. A subject pool of 40 participants was pre-registered, however recruitment was faster than expected and a total pool of 54 datasets were collected (of which 4 were rejected). No data were pre-processed or analysed before deciding to extend recruitment.

2. The 44 target items were initially selected by computing mean cloze probabilities and highest density intervals, then visually selecting the items that met the study design criteria. Improved computational methods at the stage of tidying data unfortunately revealed that visual inspection had erroneously included 4 items where a particle was not the best or second-best completion in the cloze task. These 4 items were excluded, making the total number of items included in the final analysis 40.

Figure 10 shows that plausible particles did not elicit an N400, but that there was a slight negative shift in the ERP for the plausible 2+particle condition. This was to be expected, since the plausible 2+particles had a lower cloze probability than the plausible 1-particles, and was not further analysed. Violation particles showed a clear N400, but the amplitude of the N400 did not appear to differ between the 1- and 2+particle conditions (b vs. d). This was supported by the statistical analysis presented in Table 16. Figure 14 shows the posterior in visual form. In the pre-registration we stated that the presence of an effect would be assessed by determining whether the 95% credible interval contained zero and whether the proportion of the posterior greater than zero was at least 95%. Subsequent information indicated that these proportions are not a good measure of whether an effect is not zero, since the tails of the posterior distribution have a lot of variability, meaning that what is in the 95% credible interval in one set of posterior samples may not be in the next set. The proportion of posterior samples greater than zero also does not give evidence that the effect is not zero, since a null effect would have a distribution around zero. For these reasons, we present the posteriors but base our decisions on the presence or absence of an effect on Bayes factors.

Evidence for the presence of an effect against the null hypothesis was assessed using Bayes factors. The Bayes factor gives a ratio of the marginal likelihoods of two models. We compared models with the predictor ‘number of particles’ (referred to as model 1) versus a reduced model without this predictor (referred to as model 0), i.e. BF_{10}. We assessed the strength of the evidence with reference to the Bayes factor classification scheme in (Jeffreys, 1939). A Bayes factor of approximately 1 would indicate no evidence in favour of either model. A BF of greater than 3 (when the comparison is BF_{10}) would mean there was evidence in favour of the model with the predictor, and a Bayes factor of less than $\frac{1}{3}$, evidence in favour of the null hypothesis. Since the Bayes factor is very sensitive to the prior used, we computed Bayes factor not only using the planned priors, but also for models with more and less informative priors (M. Lee & Wagenmakers, 2013). The Bayes factor analysis of the planned prior for the N400 effect found moderate evidence of just under 6:1 in favour of the null hypothesis that the N400 did not differ between the violation conditions.

			BF_{10} **for** β **priors:**		
Predictor	**Estimate (μV)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
N400					
Intercept	-0.14	-0.87, 0.60	-	-	-
2+ particles	-0.54	-1.53, 0.48	0.74	0.18	0.09
PNP					
Intercept	1.31	0.42, 2.16	-	-	-
2+ particles	0.83	-0.36, 2.07	1.14	0.32	0.16
P600					
Intercept	1.63	0.47, 2.55	-	-	-
2+ particles	0.34	-0.91, 1.62	0.71	0.17	0.08

Table 16: Model estimates and Bayes factors for the planned and exploratory analyses. The priors for the Bayes factor analyses were: Informative N(0,1), Planned N(0,5), Diffuse N(0,10).

5.3.2.4 Exploratory analysis of the post-N400 positivity (PNP)

After plotting our grand-averaged waveforms for the pre-registered analysis, we noted a distinction between conditions in the post-N400 period of the ERP (i.e. after 500 ms) at multiple electrodes, see Figure 11. The waveforms were positive-going relative to the N400, with a maximal difference between conditions apparent in fronto-central electrodes and a positive peak in parietal electrodes. The higher amplitude positivity was, in both regions, associated with violations in the 1-particle condition. That is, violations of the 1-particle condition elicited larger positivities than violations of the 2+particle condition. No such difference was visible in the plausible conditions.

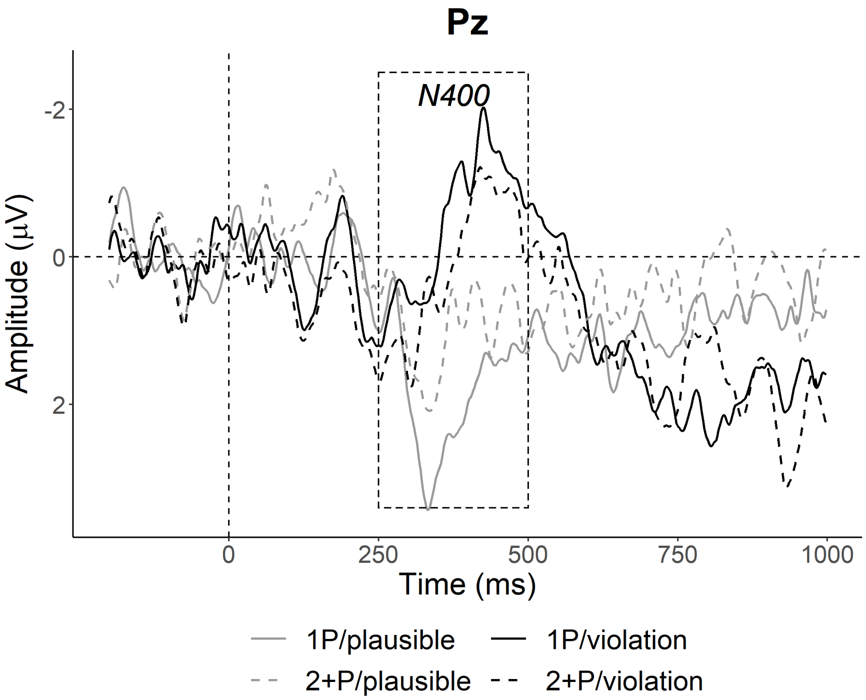


Figure 10: Results of the pre-registered analysis. An N400 is seen in the violation condition for both the 1-particle and 2+particle conditions. The plot provides visual confirmation that there was no N400 in the plausible condition, although a more negative deflection can be seen for the 2+particle condition, as is expected due to its lower cloze probability.

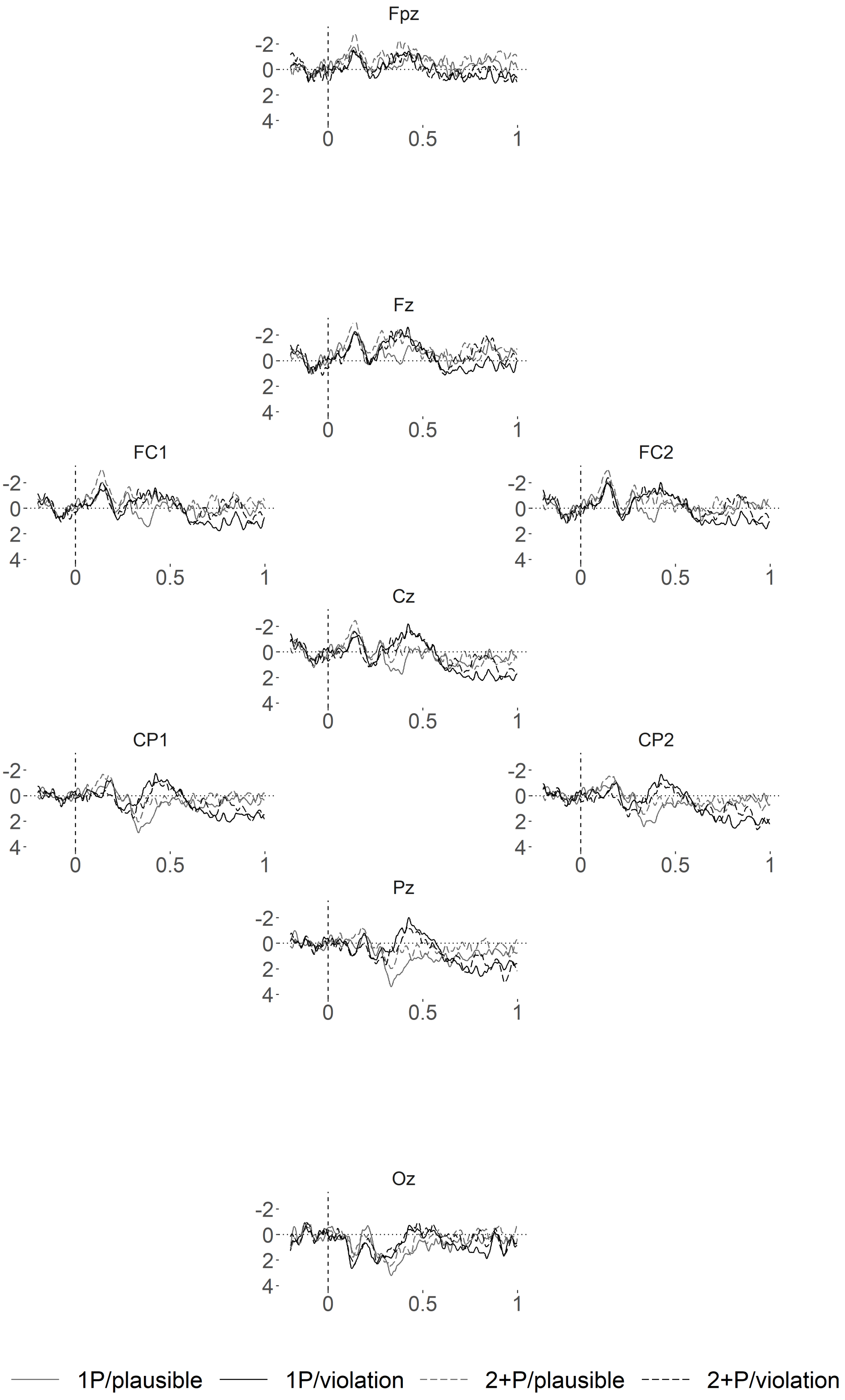


Figure 11: Grand average ERPs at selected electrodes.

A positive deflection following the N400, larger for unexpected than expected words and typically between 600 and 900 ms, is described in around 30% of ERP studies on predictive processing (Van Petten & Luka, 2012). Subsequent research has suggested that the maximum difference in the amplitude of this late positivity may differ in scalp distribution depending on the type of unexpected word. Unexpected but still plausible words have been associated with a more anteriorly distributed maximal difference, while unexpected and completely implausible words have been associated with a more posteriorly maximal difference (DeLong et al., 2014a; Kuperberg & Wlotko, 2019). Our unexpected words were carefully controlled such that they were completely implausible; however, the spatial distribution of our late positive component is best described as fronto-central. Figure 12 shows the topography of the late positivities by condition in the 600-900 ms time window.

We therefore made the data-driven selection of an electrode in the fronto-central region, *Cz*, and analysed average amplitude in the window 600-900 ms (Van Petten & Luka, 2012). Figure 13 displays the window of analysis. A linear mixed effects model was again applied in `brms` (Buerkner, 2017), also comparing only the violation conditions and using the same priors as for the N400 analysis. The model estimates showed that amplitude in the violation condition with just one plausible particle (b) was more positive than in the violation condition with two or more plausible particles (d), as can be seen in Table 16. A visual plot of the posterior can be seen in Figure 14.

The Bayes factor moderately favoured the null hypothesis by approximately 3:1. Evidence for the null hypothesis here is most likely related to the choice of prior. Relative to the effect size predicted by the original experiment (approx. $1\mu V$), a prior with a standard deviation of $\pm 5\mu V$ (i.e. 95% of observations would fall within $\pm 10\mu V$ of the mean) would be biased to favour the null, since there would be little evidence that effect sizes were much bigger than $1\mu V$. However, even with a prior with standard deviation $1\mu V$, the Bayes factor was still inconclusive, favouring neither the null nor the alternative hypothesis.

5.3.2.5 Exploratory analysis of the P600

The posterior P600 has been associated with syntactic violations; in particular, violations that cannot be repaired (DeLong et al., 2014a; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). While both particle violations elicited P600s in relation to their plausible counterparts, the amplitude of these components did not differ according to the 1-particle/2+particle manipulation, as can be seen in Table 16. The Bayes factor analysis found moderate evidence in favour of the null hypothesis that there was no difference in the amplitude of the two P600s.

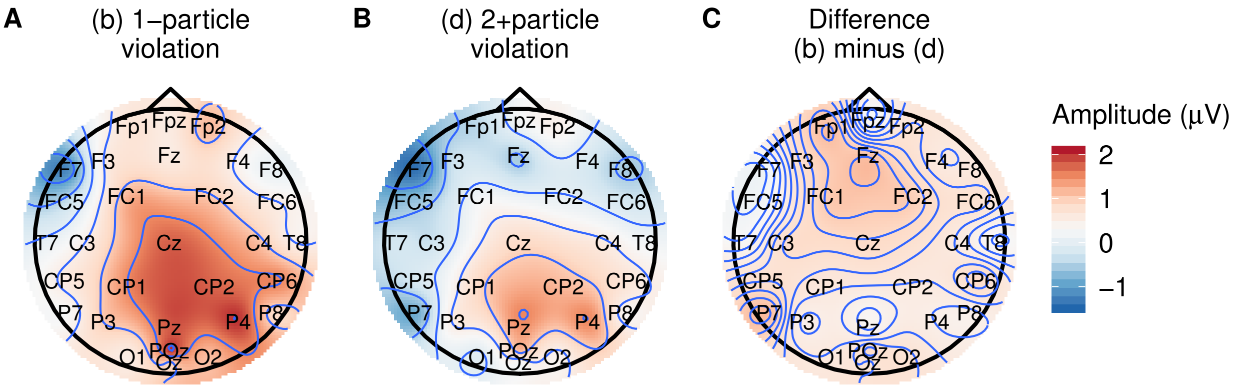


Figure 12: Topography of the late positivity. A and B: The peak of the late positivity is seen posteriorly and is more widespread for the 1-particle condition (b) than for the 2+ condition (d). **C:** The difference in amplitude of the between 1-particle and 2+particle violation conditions in the window 600-900 ms.

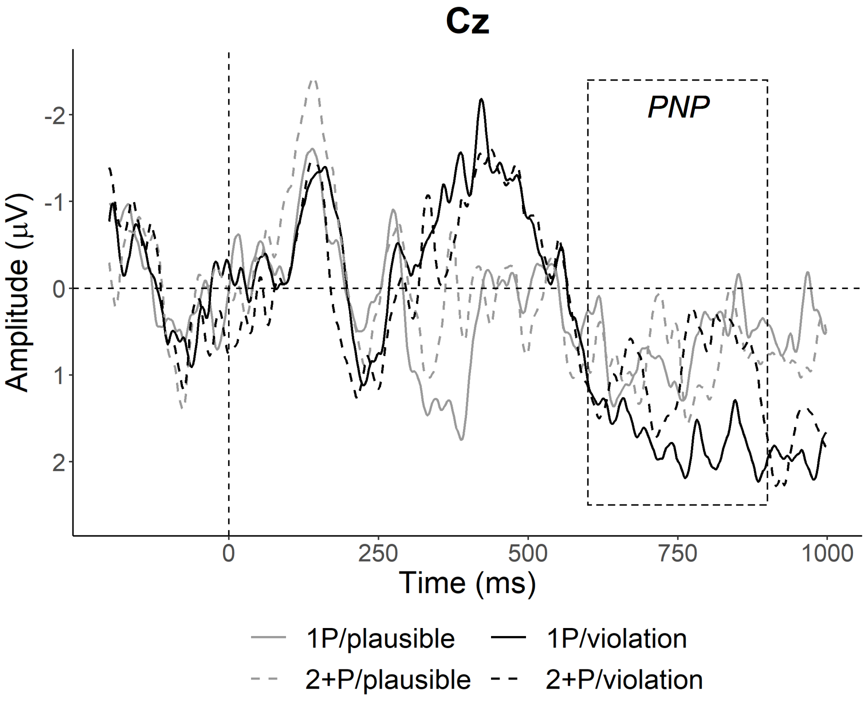


Figure 13: Results of the exploratory analysis. A late positivity peaking at around 750 ms is seen in the violation condition and is larger in the 1-particle condition. No such effect is seen in the plausible condition.

5.4 Discussion

We compared sentence pairs where either a single verb particle or a small set of verb particles were predictable from the context. We hypothesised that commitment to a lexical prediction

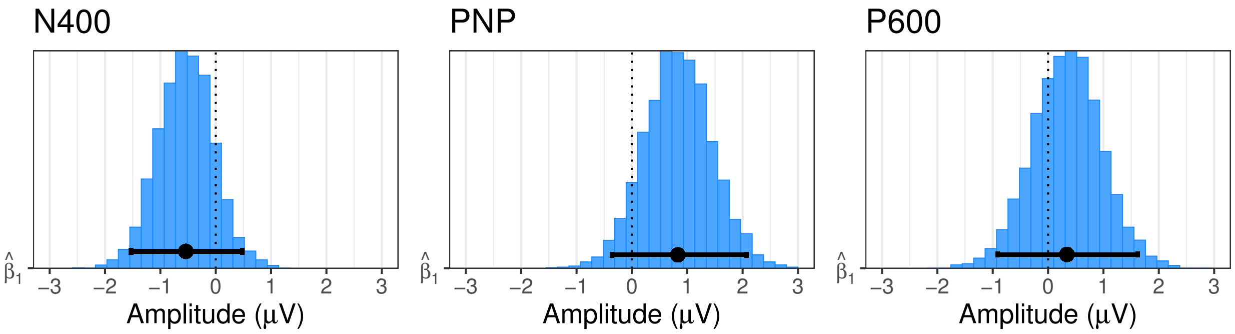


Figure 14: Model posteriors for Experiment 3. The distribution represents the estimated change in amplitude associated with the 1-particle condition relative to the grand mean (dotted line). The point and errorbar reflect the posterior mean and 95% credible interval. The posterior of the 2+particle condition would be the mirror image of the displayed posterior, on the other side of zero.

would only be made when a single verb particle was possible. Our pre-registered prediction was that violating this prediction would result in greater surprise and a larger N400. Instead, we found some suggestion that the violation elicited a larger frontal post-N400 negativity (PNP) relative to the condition where presumably no commitment was made, although the Bayes factor was moderately in favour of the null hypothesis.

The lack of an N400 difference between the two violation conditions is consistent with hypotheses that correlate the N400 with the congruency of a word in its context, regardless of the strength of that context (Kutas & Federmeier, 2011; Van Petten & Luka, 2012). In the current experiment, the violation particles were equally incongruent with the sentence context, despite one of the contexts being stronger. The presence of at least one strong lexical competitor in the weaker context condition did not appear to affect N400 amplitude. We therefore find no evidence disagreeing with current accounts of the N400.

The difference in the PNP is consistent with accounts associating the component with attempts at reanalysis, revision, and repair of a sentence parse (DeLong et al., 2014a; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). Unexpected words have consistently been found to elicit larger PNPs in high constraint versus low constraint sentences, suggesting it may represent an index of disconfirmed expectations (Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). However, several of these studies have linked the PNP to unexpected but still plausible words, such as reading *The children went outside to look* where *play* would be more expected (Federmeier et al., 2007). As far as we know, a constraint-related difference in the PNP has not previously been elicited by unexpected words that are also implausible in the context. Implausible words appear to exclusively influence the P600 (DeLong et al., 2014a; Kuperberg & Wlotko, 2019;

Thornhill & Van Petten, 2012).

Our violation particles were carefully selected to be completely implausible in the given context and, accordingly, elicited a P600 which did not appear to be affected by constraint. The P600 was elicited despite the fact that the violation of the sentence was semantic and not syntactic. The influence of semantics on the P600 was discussed in *Chapter 3*, and its presence here offers a clue as to how our readers dealt with the implausibility. First, the N400 indicates that readers had indeed developed a probabilistic representation of the sentence with which the implausible particle was incompatible. The P600 suggests that the semantic violation may have triggered syntactic reanalysis of the particle. Since verb particles are identical to prepositions, readers may have successfully revised the particle as a preposition, reflected by the PNP; for example, **__make__ [a story] __on__* could be revised as ***make** [a story [**on** the war in Iraq]]. . . .* The revision appeared to be more difficult when a lexical prediciton had been possible, and this may be why the PNP was larger in the 1-particle condition. Thus, even though the particle was semantically implausible, the revision attempt may still have been successful, making the current results consistent with a "successful update" account of the PNP (DeLong et al., 2014a; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012).

Why should stronger constraint make revision more difficult? We hypothesised that being able to make a specific lexical prediction in the 1-particle condition allowed greater commitment to a particular sentence representation. This commitment may be reflected by deeper processing of the sentence, including spreading activation to related concepts and targeted pre-activation of downstream information. If the prediction is then violated, all of this processing then has to be revised, repaired, or discarded, resulting in the PNP.

Two accounts not compatible with the current results are that, first, when the violation was encountered, all plausible particles were reactivated in order to revise the sentence. Reactivating more particles in the 2+particle condition (d) could reasonably be presumed to require more effort than reactivating a single particle in the 1-particle condition (b), and thus one would expect a larger component in the 2+particle condition (d). This was the opposite of what we observed. The second is the proposal of Piai et al. (2013) that verbs that take particles are maintained in working memory to facilitate their retrieval once the particle is encountered. If that were the case, there should have been no differences in ERP amplitude observable at the particle site, since the verbs should have been equally retrievable. On the other hand, the statistical analysis was inconclusive about whether the difference in PNP represented a difference from the null hypothesis.

5.4.0.1 P300

A further component elicited by our stimuli was the P300, which has been linked to task-based expectations (S. M. Garnsey, 1993; Kutas & Hillyard, 1980a). The component has been further divided into a frontally distributed P3a with a slightly earlier peak (200-300 ms) associated with the novelty of a stimulus, and a posteriorly distributed P3b with a peak depending on task complexity (but approximately 500 ms) and reflecting the relevance of a stimulus to the task at hand (Coulson et al., 1998; Osterhout, McKinnon, Bersick, & Corey, 1996; Van Petten & Luka, 2012). Early research considered the P300 as having the same neural generative process as positivities in the later part of the ERP (500+); however, subsequent research has supported the two as being distinct components (reviewed in Osterhout et al, 1996). This is still the subject of debate, however (Coulson et al., 1998; Osterhout, 1999; Sassenhagen & Fiebach, 2019).

Our experiment was not designed to test the P300/P600 distinction, but did elicit a positivity around 300 ms which may be worthy of brief discussion. There was a P3a in the anterior region which is perhaps slightly larger for 1-particle condition violations, but generally comparable for all conditions. That the P3a was comparable between conditions is consistent with the "novelty" account of the P3a, which would predict that only novel stimuli would change the component's amplitude. Since particle stimuli represented two thirds of the sentence stimuli, particles were not novel to the participants. Although the violation particle could be considered novel, detecting the semantic violation obviously recruited later processes that did not influence the P3a.

There was also a large P3b for both plausible conditions in the posterior region, slightly larger in the 1-particle condition. The P3b in the plausible conditions is difficult to reconcile with a "disconfirmed expectations" account, since the cloze test indicated that a particle was expected at this sentence region in preference to some other constituent. However, any conclusion about this account could only be made by comparing the P3bs between the plausible conditions and the violation conditions. This is not possible in the current experiment, because the cloze probabilities of the plausible conditions are not matched and because the semantic nature of the implausible conditions elicited a large N400 in the same time window. The P3b was therefore not statistically analysed and we remain agnostic as to whether they represent the same underlying cognitive process.

5.4.1 Limitations

While the sample size was relatively large for an ERP study (50 participants) and the results consistent with previous research, the study design meant that each participant saw only 11 target trials per condition, only 10 of which were included in the statistical analysis. This is not usually considered to be sufficient for an appropriate level of signal-to-noise reduction (Luck, 2005a), but was unavoidable due to the need to present the two grammatical in addition to the two ungrammatical conditions. This was necessary as we needed to establish that the verb particles were able to elicit the target ERP components. The ability of verb particles to elicit meaning-related ERP components was not a given, since some accounts of particle verbs analyse them as function words (Dehé et al., 2002) and some accounts claim function words do not elicit N400s (Brown et al., 1999; Frank et al., 2013; although cf. Van Petten & Kutas, 1991). In addition, creating stimuli that matched all the study requirements was difficult and limited the total number of experimental items. The use of mean amplitude and large components may have made some amends for the small number of trials (Luck, 2005a; Luck & Gaspelin, 2016), as well as having only made a statistical comparison of a main effect between two conditions. Nonetheless, the results should be interpreted cautiously. To address this limitation, we replicated the study with a higher powered design in the next chapter.

5.5 Conclusions

We tentatively propose that in the 1-particle condition (b), a lexical prediction was triggered and a richer mental representation of the sentence built before the particle was seen. This representation then had to be revised or discarded once the violating particle was encountered and we presume that the PNP reflects this cost. This may suggest that German native speakers make long-distance lexical predictions if constraint is not just high, but also strongly favours a single lexical item. However, statistical evidence was inconclusive.

5.6 Appendix

5.6.1 Prior distributions and predictive check for the statistical models

5.6.2 Posterior predictive checks for all three models

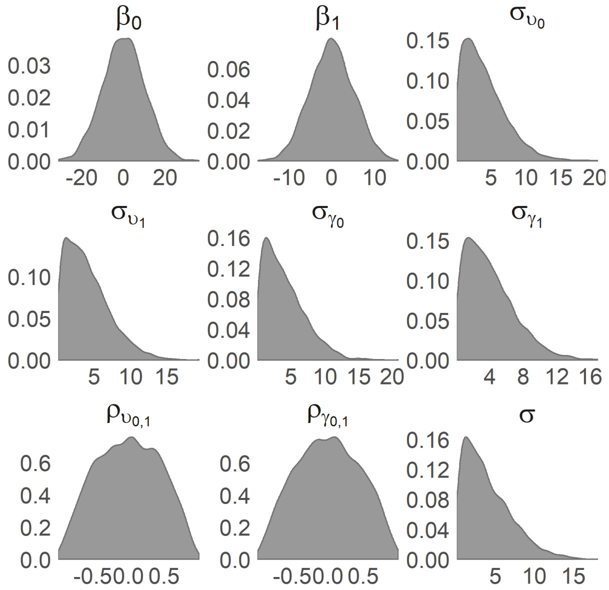


Figure 15: Prior distributions for the model parameters.

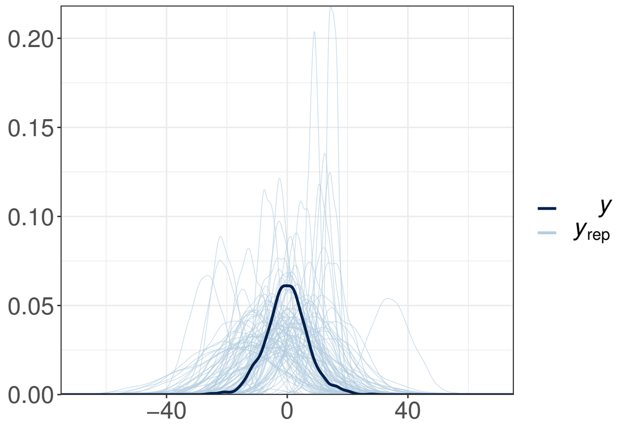


Figure 16: Prior predictive check. Simulation (light blue lines) of the observed data (dark blue line) using only the model priors suggests the priors generally appear to capture the shape of the observed data.

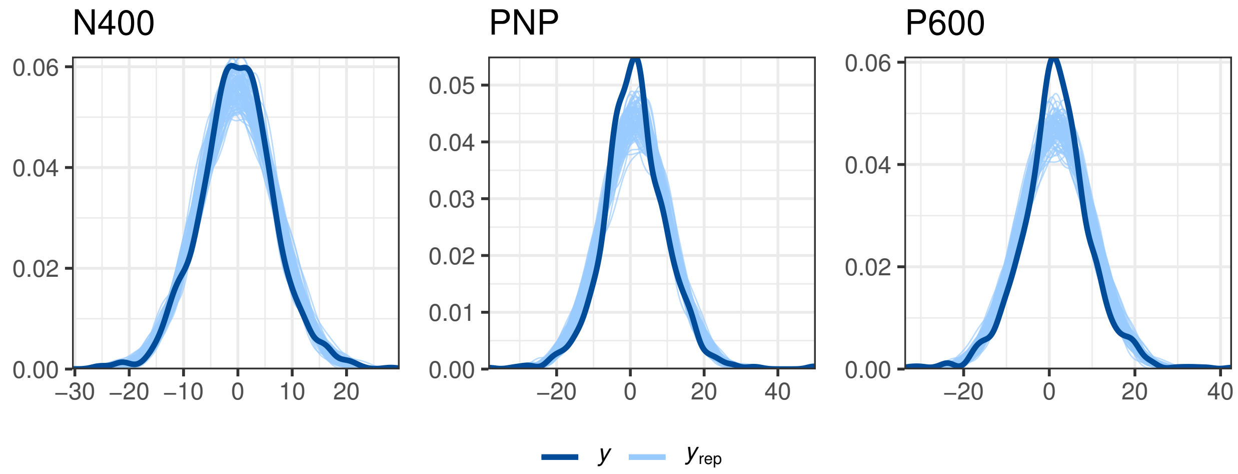


Figure 17: Posterior predictive checks for the models of the N400, PNP, and P600. The dark blue line represents the observed data and the light blue lines, data predicted by the model.

Chapter 6

6 The post-N400 frontal positivity as a possible index of prediction

6.1 Abstract

In the previous chapter, it was proposed that in German sentences where a particle appears at some distance downstream from its verb, e.g. *he* ***hung*** *the phone* ***up***, readers predicted the identity of the particle, but only if the cloze probability of the target particle was sufficiently high. Possible evidence came from the post-N400 positivity (PNP), which was observed at violations of the expected particle in contexts where only one particle had been plausible, but not at violations of expected particles in contexts where a small set of two or more particles had been plausible. This was interpreted as a re-analysis cost for revising a more strongly committed sentence representation in the 1-particle condition enabled by prediction of the particle. Interestingly, whereas previous research has suggested the PNP is not sensitive to implausible words (DeLong et al., 2014a; Federmeier et al., 2007; Kuperberg & Wlotko, 2019), the violation particles in Experiment 3 had been designed to be completely implausible. In light of this, we reasoned that participants were perhaps revising the implausible violation particle as a different word class, e.g. *he* ***hung*** *the phone [****out*** *near the door].* With this possibility in mind, the results of Experiment 3 were highly consistent with previous work showing that the PNP is larger at unexpected but plausible words in high vs. low contextual constraint contexts (DeLong et al., 2014a; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). However, statistical evidence for the PNP effect in Experiment 3 was weak. The current chapter therefore presents a pre-registered replication attempt with a larger sample size and a larger number of target trials (Experiment 4). Surprisingly, the pre-registered analysis of Experiment 4 provided inconclusive evidence that the PNP effect was actually in the reverse direction: more positive amplitude in the PNP window was observed for violations in the 2+particle condition where we assumed no lexical prediction for the particle had been made. However, exploratory analyses provided moderate evidence that the difference between the 1- and 2+particle conditions was actually consistent with zero. Further exploratory analyses suggested that the PNP was, in fact, elicited by the violation particles relative to expected particles. This latter finding raises the possibility either that participants were successfully revising the particle or that the PNP may indeed be sensitive to implausible words. In either case, the 1-particle/2+particle constraint difference may not

have been sufficiently large to produce the same constraint effects observed in previous PNP research. Evidence for long-distance lexical predictions was therefore inconclusive.

6.2 Introduction

The experiment detailed in the previous chapter suggested that commitment to a lexical prediction resulted in a larger reanalysis cost when that prediction was violated, reflected by a larger ERP component called the post-N400 positivity (PNP). This finding was consistent with previous literature suggesting the differences in the PNP are positively correlated with contextual constraint; the stronger the constraint, the larger the PNP when an unexpected word appears. However, ERP experiments are notorious for their high signal-to-noise ratio (Luck, 2005b). This means that in order to detect differences between conditions, a large number of participants must be recorded and each participant must be exposed to a sufficient number of target trials. If power is low, large effect sizes may be detected but are more likely to be Type M errors leading to overestimates of the true effect size (Gelman & Carlin, 2014). Larger effect sizes are also more likely be significant and significant results are in turn more likely to be published. The pressure to publish significant, “novel” results has led to the literature being flooded with overestimated effect sizes with a high likelihood of being simply false (Ioannidis, 2005; Luck & Gaspelin, 2016). One way to combat false positives and inflated effect size estimates is to directly replicate an experiment.

In this chapter, the results of a replication of Experiment 3 with double the participants and more than double the items are presented. Using the data from the original experiment as a guide, the replication study design, predictions, and planned analysis were pre-registered with the Open Science Framework (OSF; Foster & Deardorff, 2017) at https://osf.io/m96cq/.

A successful replication would strengthen evidence that lexical predictions can be generated in long-distance verb-particle dependencies, likely at much longer distances than previously indicated by research with verb-noun constructions (e.g. DeLong et al., 2005). Furthermore, it would demonstrate that lexical predictions incur a higher recovery cost when disconfirmed, suggesting that committing to a prediction early in the sentence may allow deeper processing of other words appearing before the predicted word. It would also support the frontally distributed post-N400 positivity (PNP) as being an index of the cost of recovery from violated predictions. Finally, the finding that a lexical prediction is avoided even when there are very few plausible sentence continuations would provide further support for the rarity of lexical predictions, presumably due to the demonstrably higher cost of recovery (Luke & Christianson, 2016). An unsuccessful replication could suggest several alternative hypotheses, depending on the magnitude and direction of new observed effect.

6.3 Experiment 4

6.3.1 Methods

6.3.1.1 Participants

Recruitment, inclusion, and exclusion criteria were identical to those Experiment 3. Participants who had taken part in Experiment 3 were not permitted to take part in Experiment 4. In total, 115 participants were recruited, 4 of whom were excluded as they did not meet the inclusion criteria. A further 7 were excluded due to technical problems with the EEG recording, and 5 were excluded due to preprocessing issues that had not been resolved at the time of writing. This left a total of 100 subjects (24 male), with a mean age of 24 years (range = 18 to 35 years, SD = 4 years).

6.3.1.2 Power analysis

A power analysis was carried out by using the data from Experiment 3 to simulate new 'experiments'. However, the effect size of Experiment 3 is likely larger or smaller than the true effect size and should not be relied to accurately estimate the power of a second experiment (Albers & Lakens, 2018). We therefore computed power using a range of plausible effect sizes by calculating the average number of 'significant' results for each effect size. The simulated experiments assumed the planned sample size of 100 participants and 27 target items per participant. The process for simulating the data was as follows: First, a linear mixed effects model (LMM) was fit to the original data in `R` (R Core Team, 2018) using the package `lme4` (Bates, Mächler, Bolker, & Walker, 2015):

$$\hat{y} = \beta_0 + \upsilon_0 + \gamma_0 + \beta_1 X + \epsilon$$

where β_0 was the intercept, υ_0 and γ_0 the by-subject and by-item intercept adjustments, β_1 the slope, and ϵ the error term. A frequentist approach to the power analysis was used even though the planned analysis was Bayesian, simply because at the time the analysis was conducted, the computational power to conduct such a power analysis in a Bayesian framework was not available. Unfortunately, this also meant that models with full random effects structures could not be fit due to frequent convergence failures. We therefore made only by-subject (γ) and by-item (υ) adjustments for correlated varying intercepts in the power analysis. No adjustments for varying slopes were added. True power is therefore likely to be much more conservative than the estimates generated by the presented analysis and as such, the analysis is interpreted cautiously.

Next, a vector of plausible β_1 values was generated between $0.1\mu V$ and $2.50\mu V$ (the β_1

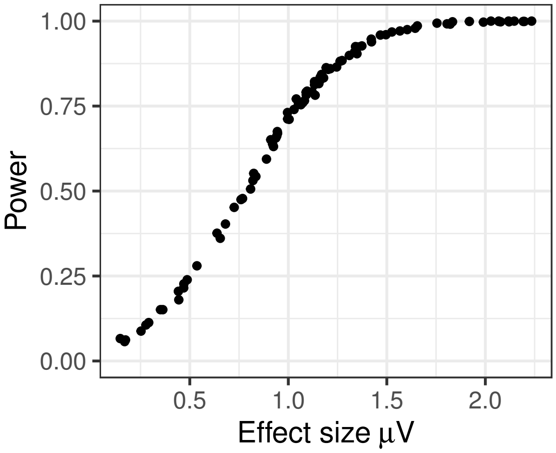


Figure 18: Estimated power for a range of effect sizes of the late positivity. Plausible effect sizes were randomly sampled using parameters estimated from the original experiment. Note that since data for the power analysis were simulated without full random effects structure, the power estimates should be interpreted as very optimistic.

yielded by a Bayesian analysis of the original data was $0.83\mu V$). Then, for each coefficient in this vector, 1000 datasets were simulated using the simulate function from the R core package `stats`. The `simulate()` function used the β_0, υ_0, γ_0, and ϵ estimates of the LMM applied to the original data, while the slope, β_1, was one of the values in the vector of plausible coefficients. Each of the 1000 simulated datasets was fitted with the LMM specified above. For each of the 1000 model fits, the t-value for the slope estimate of β_1 was extracted. The mean number of significant t-values at an alpha of 0.05 (t > 2.00) over the 1000 simulations was calculated. This mean was then assigned to the relevant β_1 coefficient as its 'power'. This process was repeated for each values in the vector of plausible β_1 coefficients. The power of each of the plausible β_1 coefficients is plotted in Figure 18.

The coefficient of the late positivity from the Bayesian model fitted to the original data, obtained from 50 participants and 11 trials per participant, was $0.83\mu V$. Differences in the late positivity elicited by experiments with similar designs range between approximately 1 and 5 μV (DeLong et al., 2014a; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). As can be seen in Figure 18, power to detect an effect size of $0.83\mu V$ is at most approximately 70%, although as mentioned above, this is likely to be overly optimistic.

6.3.1.3 Materials

The materials used were 43 of the sentences used in Experiment 3, plus 11 new sentences. As for Experiment 3, sentences were constructed that constrained the set of plausible particles to either just one (condition a/b), or to a small set of two or more particles (condition c/d). For the replication, an additional 20 sentence pairs were cloze tested with 30 more native German speakers (mean age 24 years, SD 5 years, range 18-38 years). From these 20 items, 11 final items were selected. One of these replaced the lowest-ranked of the 44 items from the original cloze test, and the remaining 10 additional items were added to give a new total of 54 items. However, as noted in *Chapter 5*, 4 items had to be excluded, making the total number of items included in the replication analysis 50. Cloze probabilities for the best completion (BC) particles in the final set of 50 items are summarised in Table 17, as well as the difference in cloze between the best and next-best completion.

	Target particle		**Difference between 1st- and 2nd-BC**	
Condition	**Mean**	**95% CrI**	**Mean**	**95% CrI**
1-particle	0.89	0.73, 1.00	0.72	0.64, 0.84
2+particle	0.53	0.52, 0.55	0.28	0.14, 0.38

Table 17: Cloze probability summary statistics for the plausible conditions. 1st- and 2nd-best completions (BC) refer to the highest and second-highest cloze particles at the target site.

Experiment 4 deviated slightly from Experiment 3 in that the plausible conditions (a,c) were *not* presented to participants. Instead, grammatical filler sentences matched for length preserved the proportion of grammatical to ungrammatical sentences in the paradigm. The reason for doing this was to double the number of violation trials seen by each participant by showing them one of the two violation conditions from each item (either b or d) rather than one of the four total conditions (a, b, c, or d). It was not felt necessary to again compare the ERPs of the violation conditions to a precisely matched plausible condition as it had already been demonstrated in Experiment 3 that the violation particles did indeed elicit the expected ERP components (N400 and late positivity) in relation to their plausible counterparts (see *Chapter 5*). As such, for the replication study, we reasoned that doubling the number of target trials was more important than having perfectly matched plausible conditions. In addition to the target items, unmatched filler sentences were also added such that the final proportion of grammatical to ungrammatical sentences remained 3-to-1.

An example experimental item is presented in 14. As discussed above, only the violation conditions (b) and (d) from the original stimuli were used. The additional condition (e) was a grammatical filler with a plausible particle, matched for structure and length, as well as for the identity of the pre-critical, critical (particle), and spillover regions. Each participant saw either (b) *or* (d), *and* (e) from each item:

(14) (b) 1 plausible particle/violation:

Der ordentliche Professor* ***fuhr* *mit seinem Vortrag trotz*
The orderly professor **carried** with his lecture despite
regelmäßiger Störungen immer ordnungsgemaß ***wahr****, da er für seine*
regular interruptions always properly **true**, as he for his
Unaufgeregtheit bekannt war.
unflappability known was.

The orderly professor carried true with his lecture despite regular interruptions always as directed, as he was known for being unflappable.

(d) 2+ plausible particles/violation:

Der ordentliche Buchhalter* ***fuhr* *seinen zuverläßssigen Computer bei*
The orderly accountant **turned** his reliable computer at
der Arbeit immer ordnungsgemaß ***wahr****, da er für seine korrekte*
work always properly **true**, as he for his correct
Arbeitsweise bekannt war.
work practices known was.

The orderly accountant turned true his reliable computer at work always as directed, as he was known for his faultless work practices.

(e) Matched plausible filler:

Der pingelige Feldwebel ***gab*** *jeden Befehl im Radio pünktlich und*
The fastidious sergeant **gave** every order in the Radio punctually and
immer ordnungsgemaß ***durch****, da er für seine korrekte Arbeitsweise*
always properly **through**, as he for his correct work practices
bekannt war.
known was.

The fastidious sergeant gave every order over the radio punctually and always as directed, as he was known for his faultless work practices.

In addition to the 54 target items and 54 matched fillers presented to participants, 108 more general filler sentences were randomly interspersed. These were 62 of the fillers from the original experiment plus 46 new fillers. Each participant therefore saw a total of 216 sentences during the testing session. After each target or filler sentence, participants answered a yes/no question as for the previous experiment.

6.3.1.4 Procedure

The procedure was identical to that of Experiment 3.

6.3.1.5 EEG recording and preprocessing

EEG recording and preprocessing followed the same pipeline as for Experiment 3.

6.3.1.6 Pre-registered analysis

The model specification for the N400 and the PNP in Experiment 4 were identical to that in Experiment 3: only conditions (b) and (d) were compared. The dependent variables were changed, however. Instead of single electrodes, the dependent variables were two regions of electrodes selected based on the topographical presentation of the N400 and PNP in Experiment 3. The new dependent variable for the N400 was therefore mean amplitude of electrodes Cz, Pz, CP1, and CP2 in the 250-500 ms time window, and for the PNP, mean amplitude of electrodes Fz, FC1, and FC2 in the 600-900 ms window. The prior specification was as for the previous experiment:

$$\begin{gathered}
\beta_0 \sim Normal(0, 10) \\
\beta_1 \sim Normal(0, 5) \\
\upsilon_{0,1} \sim Normal(0, \sigma_\upsilon) \\
\gamma_{0,1} \sim Normal(0, \sigma_\gamma) \\
\sigma_\upsilon, \sigma_\gamma \sim Normal_+(0, 5) \\
\rho_\upsilon, \rho_\gamma \sim LKJ(2) \\
\sigma \sim Normal_+(0, 5)
\end{gathered}$$

6.3.2 Results

6.3.2.1 Accuracy and reaction times

Mean accuracy and reaction times to the yes/no questions following each stimulus are presented in Table 18. Overall, accuracy and reaction times were comparable between all conditions.

	Accuracy		**Response time**	
Condition	**Mean (%)**	**95% CI (%)**	**Mean (ms)**	**95% CI (ms)**
1P violation	87	85, 88	2,230	2177, 2282
2+P violation	86	84, 87	2,292	2238, 2345
Plausible	90	89, 90	2,322	2286, 2358

Table 18: Mean accuracy and response times.

6.3.2.2 EEG data exclusion

Of the total data collected from 100 subjects, 3.33% were excluded due to artefact. Out of 5400 target trials, 0.57% were excluded due to question response times over 10 seconds (indicating a technical problem) and 0.28% were not recorded due to experimenter error.

6.3.2.3 Pre-registered N400 analysis

Figure 19 plots the ERPs and visually confirms that Experiment 4 successfully elicited the expected N400 and P600 components in the violation conditions relative to the plausible filler. ERPs at a broader range of electrodes are plotted in Appendix 6.6, Figure 24. For convenience, the results of the statistical analyses of the N400, PNP, and P600 components are combined in Table 19 and Figure 21. Each analysis is reported separately, beginning with the N400.

The N400 appeared to be equally large for both the violation conditions in the replication experiment. This statistical analysis was less conclusive, however. Table 19 shows the model estimates and Bayes factors (BF) with only anecdotal evidence in favour of the null hypothesis (according to the scale of Jeffreys, 1939). The posterior for the N400 model can be visualised in Figure 21.

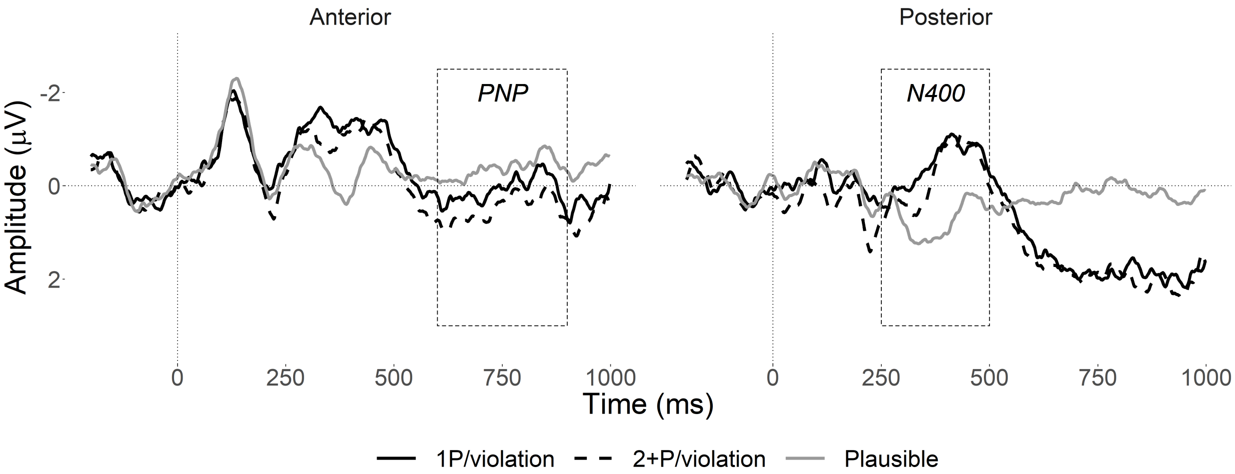


Figure 19: ERPs elicited by verb particles. The pre-registered regions of interest analysed in the replication experiment were anterior (Fz, FC1, FC2) and posterior (Cz, CP1, CP2, Pz) regions.

6.3.2.4 Pre-registered analysis of the post-N400 positivity (PNP)

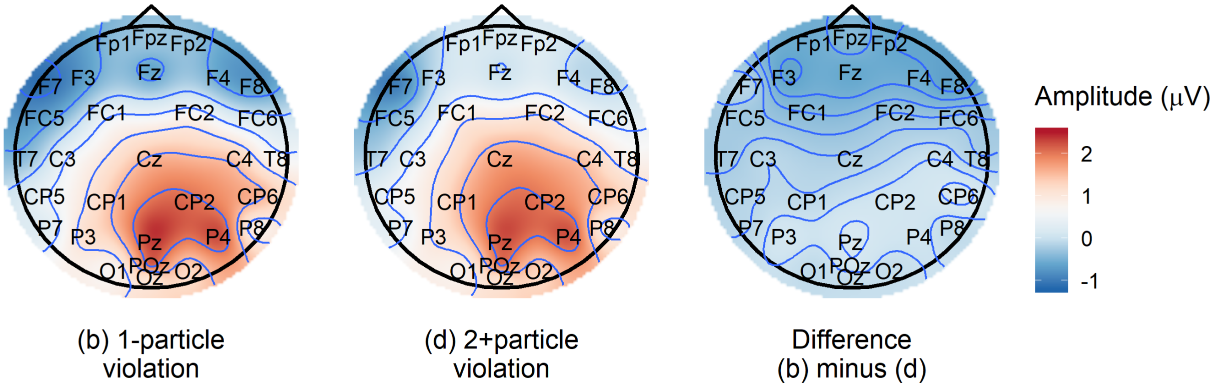


Figure 20: Topography of the late positivity 600-900 ms. The first two topographical plots show mean ERP amplitude of the plausible fillers subtracted from the respective violation conditions. Red thus indicates a larger amplitude ERP in the 1-particle condition. In the third plot, mean amplitude in the 1-particle violation condition is subtracted from the 2+particle violation condition. Blue indicates higher amplitude in the 2+particle condition.

The plot of the ERP in Figure 19 suggests there was a PNP in both violation conditions relative to the plausible fillers, but that it was larger in the condition where two or more particles had been plausible than the condition where only one had been plausible. This can also be seen in the topographical plot in 20 (right panel), where the faint blue region frontally suggests a slightly larger PNP for violations in the condition where two or more particles had been plausible. Statistical evidence for the pre-registered analysis in Table 19 was inconclusive, with the Bayes factor indicating no evidence in favour of either the null or the alternative model. The model posterior in Figure 21 is consistent with a positive effect, although zero falls within the 95% credible interval, meaning it is also plausible that the effect was zero.

6.3.2.5 Exploratory analysis of the P600

As for the original experiment, the P600 was analysed using the same model and priors as for the other components. Predictions about the P600 were not pre-registered as the component was not predicted to be modulated by semantic constraint. The original experiment did not suggest that there was any statistical difference in the P600, although there was a slight visual difference, and it was therefore analysed again for comparison. The dependent variable for the P600 was mean amplitude in the region Cz, Pz, CP1, and CP2 in the 600-900 ms time window. Additionally, a Bayes factor analysis was conducted to assess the evidence for a difference in amplitude against zero using the same range of priors reported in Experiment 3, see Table 19.

			BF_{10}:		
Predictor	**Estimate (μV)**	**95% CrI**	**Informative**	**Planned**	**Diffuse**
N400					
Intercept	-0.28	-0.74, 0.19	-	-	-
2+ particles	-0.22	-0.66, 0.22	0.33	0.07	0.03
PNP					
Intercept	0.29	-0.28, 0.86	-	-	-
2+ particles	-0.42	-0.96, 0.11	1.75	1.00	0.22
P600					
Intercept	1.82	1.26, 2.37	-	-	-
2+ particles	-0.01	-0.51, 0.49	0.26	0.05	0.03

Table 19: Model estimates and Bayes factors for the replication experiment. Bayes factors are reported for the planned β_1 prior, $N(0, 5)$, as well as more informative, $N(0, 1)$, and more diffuse $N(0, 10)$ priors.

As can be seen visually in Figure 19, from the model estimates in Table 19, and from the model posterior in Figure 21, there was no suggestion of a constraint-related difference in the P600. The Bayes factor for the P600 was just under 4:1 ("moderately", Jeffreys, 1939) in favour of the null hypothesis.

6.3.2.6 Exploratory assessment of replication

In order to determine the overall success of the replication, a region of practical equivalence was used (L. S. Freedman, Lowe, & Macaskill, 1984; ROPE, Kruschke, 2011; Spiegelhalter, Freedman, & Parmar, 1994). The ROPE was determined using the 95% credible intervals of the effect coefficients estimated for the three ERP component analyses in Experiment 3. According to the ROPE approach, if the effect coefficients estimated from the analyses of Experiment 4 and their 95% credible intervals fall within their respective ROPE, they would be considered to have been successfully replicated. If not, success or failure of the replication would be determined based on the direction and magnitude of the mismatch. The decision schema can be visualised in Figure 22A. Since single electrodes were analysed in Experiment 3, the ROPE comparison was also based on analysis of Experiment 4 using single electrodes instead of the pre-registered regions.

As can be seen in Figure 22B, both the N400 and P600 results were successfully replicated (black intervals). As concerns the PNP, the sign of the posterior mean reversed between Experiments 3 and 4, but was still within the range of values predicted by Experiment 3. However, according to the decision schema in Figure 22A, the success of the PNP replication was inconclusive.

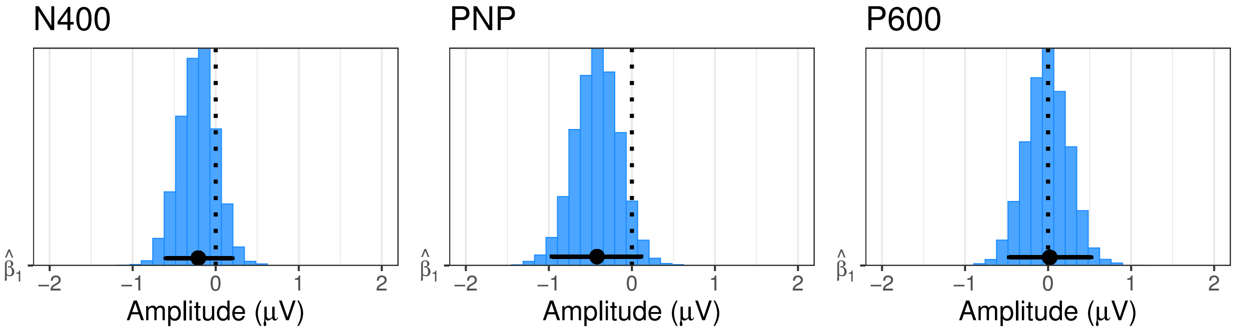


Figure 21: Model posteriors for Experiment 4. The posterior distributions reflect estimated amplitude in the 1-particle condition relative to the grand mean (dotted line). The point and errorbar reflect the posterior mean and 95% credible interval. The posterior of the 2+particle condition would be the mirror image of the displayed posterior, on the other side of zero.

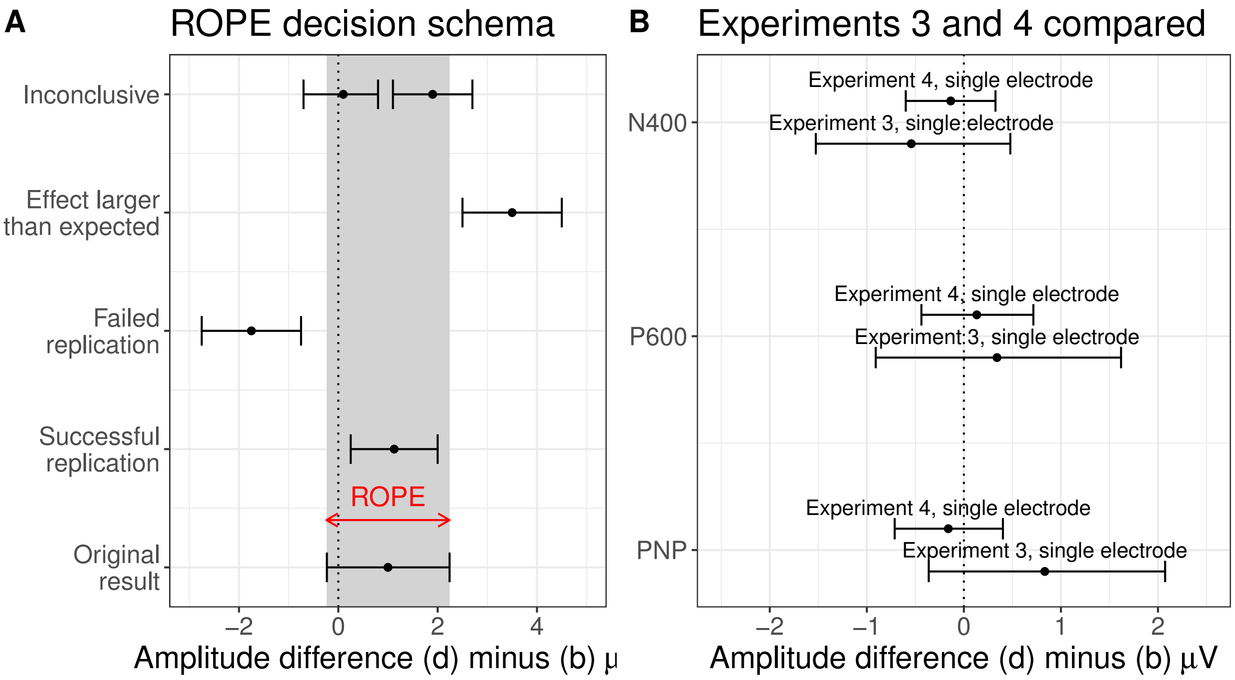


Figure 22: Decision schema and results of the ROPE assessment. A. A range of hypothetical results and respective decisions based on a ROPE. **B.** Comparison of results from Experiments 3 and 4.

6.3.2.7 Interim discussion and exploratory analysis of the PNP

What does the inconclusive replication mean for the PNP effect? Was there an effect of the 1-particle/2+particle manipulation or not? One way to assess this would be to conduct a Bayes factor analysis of the PNP effect in Experiment 4 using the posterior and 95% credible interval estimates from Experiment 3 as a prior. This informative prior is justified based on the fact that we have knowledge from an almost identical experiment which is consistent with evidence from conceptually similar studies (Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). This analysis produces a Bayes factor 7:1 ("moderately"; Jeffreys, 1939) in favour of the null hypothesis that there was no PNP difference between the 1-particle/2+particle violation conditions.

However, while the results of Experiment 3 were consistent with a number prior studies directly testing the PNP, these are arguably only a handful of studies that potentially reflect publication bias (Ioannidis, 2005). In fact, only around 30% of published ERP studies on predictive processing elicit a PNP at all (Van Petten & Luka, 2012). Furthermore, all of the studies directly testing the PNP have done so in verb-noun dependencies and usually in English (with the exception of Hebrew in Ness & Meltzer-Asscher, 2018). The PNP has not been directly investigated in ERP studies of verb-particle dependencies, but ERP plots from these studies suggest that the ERP pattern in the anterior scalp region following the N400 is quite variable. In Dutch, no difference in ERPs is visibly elicited either by semantically odd but still plausible or by implausible particles relative to well-formed verb-particle constructions (Piai et al., 2013). In German, a small PNP (mid-frontally) or even a negative component (pre-frontally) can be seen for implausible vs. well-formed verb-particle constructions (Czypionka et al., 2019). The prior for the direction of the PNP effect in verb-particle constructions is therefore not especially strong. Why should verb-particle constructions produce different results to verb-noun dependencies? As already mentioned, a particle is perhaps more readily revisable as a different word class than is a noun, or at least this possibility may be entertained for longer. For example, the particle in 15a potentially offers several different (albeit awkward) avenues for reanalysis compared to the noun in 15b (adapted from Kuperberg & Wlotko, 2019):

(15) a. * After many unsuccessful job applications, she finally gave on...

b. * The lifeguards received a report of sharks right near the beach. Hence, they cautioned the drawer...

For example, 15a could possibly continue as *. . . the condition of secrecy to the charity* (although arguably, 15b could also continue in a similarly awkward manner, e.g. *. . . of the bath*). An argument against this kind of revision in the particle verb constructions in Experiments 3 and 4 is that a comma was presented with the particle, which in German strictly indicates a clause boundary and means that the particle could not have been a preposition. On the other hand, if attempts at revision are being made, those attempts could conceivably also include assuming the comma was displayed in error.

Combining the data from Experiment 3 and 4 offers a solution to this puzzle. Using the combined data from both experiments with 150 subjects and 51 target items, we applied the same LMM described in the pre-registered PNP analysis with the same dependent variable (mean amplitude in the 600-900 ms window across electrodes Fz, FC1, and FC2). An additional predictor was added to reflect from which experiment the data had come (β_2). Due to convergence issues, the priors had to be made more informative than the pre-registered priors and were as follows:

$$\begin{gathered}
\beta_0 \sim Normal(0, 5) \\
\beta_{1,2} \sim Normal(0, 1) \\
\upsilon_{0,1,2} \sim Normal(0, \sigma_\upsilon) \\
\gamma_{0,1,2} \sim Normal(0, \sigma_\gamma) \\
\sigma_\upsilon, \sigma_\gamma \sim Normal_+(0, 0.5) \\
\rho_\upsilon, \rho_\gamma \sim LKJ(2) \\
\sigma \sim Normal_+(0, 5)
\end{gathered}$$

The posterior estimated by this model (Figure 23, right panel) did not suggest any difference between the 1-particle/2+particle violation conditions, $\hat{\beta} = -0.16\mu V$, $95\% CrI = [-0.62, 0.28]$, and a Bayes factor was just under 4:1 (“moderately”, Jeffreys, 1939) in favour of the null hypothesis. The posterior for the “experiment” predictor was not consistent with a difference between the two experiments, $\hat{\beta} = -0.06\mu V$, $95\% CrI = [-0.85, 0.74]$.

6.3.2.8 Exploratory analysis of expected vs. implausible particles

Having established that there was no constraint-related PNP difference elicited by the violation particles, the last remaining question was whether a PNP was elicited in the violation conditions at all. The basis for this question is that previous studies have suggested that the PNP is *only* elicited by plausible (but unexpected) vs. expected words and not by implausible vs. expected words, regardless of whether constraint is manipulated (DeLong et al., 2014a; Kuperberg & Wlotko, 2019).

The presence of a PNP at violation versus expected particles in Experiments 3 and 4 can only be cautiously analysed, since the target particles in the “expected” conditions are not matched with the violation particles for identity or cloze probability. This introduces length, frequency, and other semantic association differences. Nonetheless, an exploratory analysis

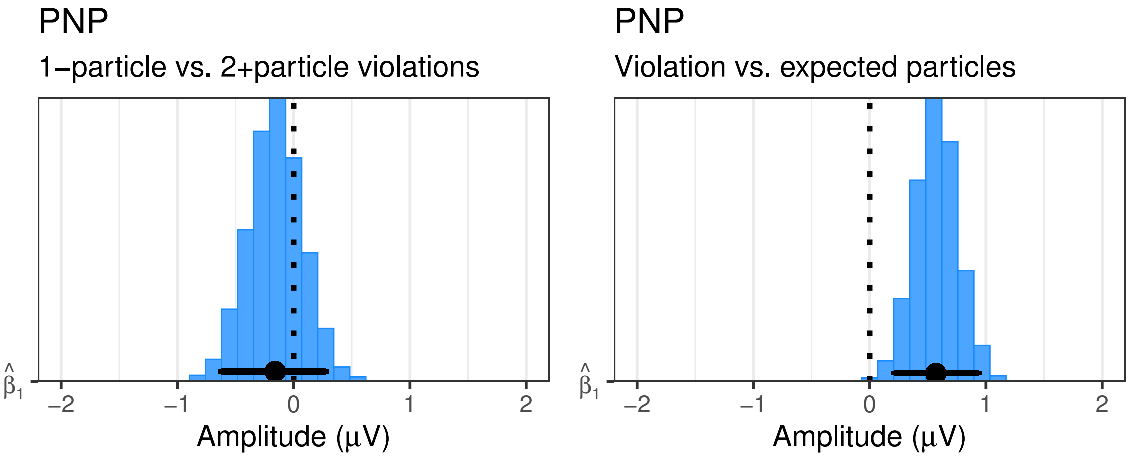


Figure 23: Model posteriors for analyses of the combined data from Experiments 3 and 4. LEFT: Posterior distribution of the difference in amplitude of violations in the 1-particle condition relative to the grand mean (dotted line). The point and errorbar show the posterior mean and 95% credible interval. RIGHT: Posterior distribution of the difference in amplitude of violation particles relative to the grand mean (dotted line). The posterior for expected particles would be the mirror image on the other side of the grand mean.

comparing the two conditions is presented here as a launching point for future investigations.

For this analysis, mean amplitude in anterior region in the 600-900 ms window again formed the dependent variable and the 1-particle/2+particle conditions were collapsed given that the results above suggested they did not differ. The combined data of Experiments 3 and 4 were analysed. An LMM with full variance-covariance matrices estimated for the random effects of subject and item was fit with the predictors “expectedness” and “experiment number” (contrast coding was: expected -0.5/violation 0.5; Experiment 3: -0.5/Experiment 4: 0.5). As for the model of the combined data above, informative priors were used to avoid convergence issues (listed above). The posterior of this model was consistent with a more positive PNP for violation vs. expected particles, $\hat{\beta} = 0.57\mu V$, $95\% CrI = [0.22, 0.92]$ (see Figure 23, right panel). Given the unmatched target regions used in this analysis, a Bayes factor was not computed in order to avoid over-interpretation of the result.

6.4 Discussion

An attempt was made to replicate the result of Experiment 3 which found that the frontal post-N400 positivity (PNP) was elicited at violation particles in sentences which had only one plausible particle continuation versus sentences which had more plausible particle continuations. Statistical analysis of the effect in Experiment 3 was weakly suggestive of a null effect, although the visual pattern was consistent with a number of studies suggesting a positive correlation between the amplitude of the PNP and contextual constraint (Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Thornhill & Van Petten, 2012). Our tentative conclusion had been that commitment to a specific lexical item possible in the 1-particle sentences made revision of the sentence representation more difficult than when commitment was less likely to have been made in the 2+particle sentences. In an attempt to replicate this result in a larger sample with more experimental items, a PNP was elicited in both the 1-particle and 2+particle violation conditions relative to a plausible filler, but a ROPE analysis suggested the success of the replication attempt was "inconclusive". A Bayes factor analysis of Experiments 3 and 4 combined was strongly in favour of the null hypothesis that there was no PNP difference between the 1- and 2+particle conditions.

Since constraint has previously been found not to modify the PNP when the target word is implausible, the most straightforward interpretation of the results is that the violation particles were not being revised as e.g. prepositions, and that successful revision of the mental sentence representation was therefore not possible. This would be highly consistent with previous studies suggesting that the PNP and its corresponding constraint effect are only elicited by unexpected words if they are still plausible (DeLong et al., 2014a; Kuperberg & Wlotko, 2019). On the other hand, the exploratory finding that a PNP was elicited in both violation conditions relative to an expected particle contradicts this interpretation. While the unmatched target regions involved in this analysis mean that its results should not be over-interpreted, it does tentatively suggest either that the implausible particles were being revised after all, or that the PNP can be elicited by implausible words. Assuming that either of these possibilities is true, the larger PNP at violation vs. expected particles has relevance to the research question about whether long-distance lexical predictions were being generated.

In both the 1- and 2+particle conditions, the probabilistic environment associated with the mental sentence representation would have been quite strong, even if uncertainty in the 2+particle condition discouraged specific lexical predictions. The resources required to revise or suppress the representations may therefore not have been sufficiently different between the 1- and 2+particle conditions to be detected by the scalp EEG. Thus, while we cannot tell whether specific lexical predictions were made in the 1-particle condition, we can at least infer that both the 1- and 2+particle conditions resulted in strong probabilistic environments which were comparably difficult to update following implausible input. While not overly surprising, this inference does support a graded effect of constraint on predictive processing rather than a "prediction or nothing" hypothesis.

In terms of why the current experiments found a possible PNP difference between implausible and expected words where previous experiments have not, one possible explanation is, of course, power. Experiments 3 and 4 combined included 150 participants who saw 10-25 target trials per condition. Previous studies have tested fewer participants (25-39), but also more trials per participant (28-70, Federmeier et al. (2007); Thornhill & Van Petten (2012); DeLong et al. (2014a); Ness & Meltzer-Asscher (2018), with the exception of one which had 20, Kuperberg & Wlotko (2019)]. The large-sample study reported here is therefore arguably more powerful and may suggest that previous experiments have falsely failed to reject the null hypothesis that there is no PNP elicited by implausible vs. expected words (i.e. low power has resulted in a type II error).

A second possible explanation is that the current experiment differs in the linguistic manipulation tested. Previous studies specifically testing the PNP have utilised verb-noun combinations, where the noun is either unexpected but still plausible in the given context, or completely implausible. As discussed above, revision as a different syntactic element is not possible with a noun (or is at least more difficult than with a particle), which may mean there are different reanalysis processes involved between the current and previous experiments. This may be supported by the presence of a "semantic" P600 in both Experiment 3 and 4, which suggests that syntactic revision of the particle was triggered by the implausible meaning (Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2017; A. Kim & Osterhout, 2005; Kuperberg, 2007; Kuperberg et al., 2003). On the other hand, the presence of a comma after the particle in both Experiments 3 and 4 strongly discouraged this type of revision. Interestingly, a comma was also present in Czypionka et al. (2019) where illegal particles elicited larger anterior ERPs in the post-N400 than licensed particles (albeit in the opposite direction). A comma was not present in Piai et al. (2013) where no such ERP difference was observed. The difference in findings between different grammatical constructions raises the question as to whether the PNP may be additionally sensitive to different reanalysis strategies.

6.5 Conclusions

We were unable to replicate the finding from Experiment 3 that a larger post-N400 positivity (PNP) was elicited by committed lexical predictions in comparison to situations where commitment to a lexical item was less likely. Previous research on the PNP suggests that this was due to the implausible particle used as the target word, since the PNP indexes the cost of successfully updating the mental sentence representation following violated predictions with unexpected but plausible input. However, in contrast with previous PNP research, there was some suggestion that a PNP was elicited by implausible words relative to expected words.

This raised the possibility that the implausible particles did indeed elicit a PNP, but that the 1-particle vs. 2+particle manipulation was not sufficiently strong for a processing difference to be detected. While we were unable to determine whether participants were generating long-distance lexical predictions, the findings do suggest that a constraint-related PNP effect may be observed at implausible words if a more dramatic constraint manipulation were used. Finally, the contrasting findings with respect to previous research suggest that the PNP and its relationship with the cost of disconfirmed predictions has not yet been fully characterised.

6.6 Appendix

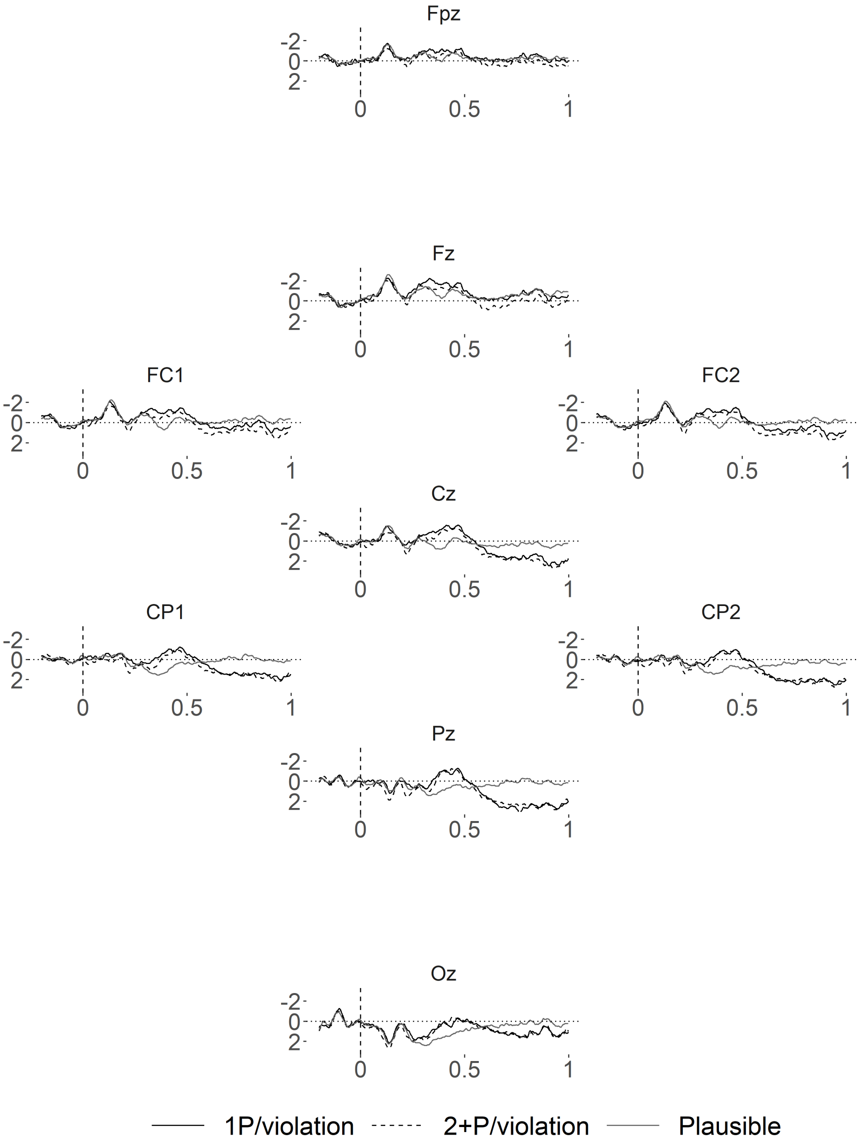


Figure 24: Grand average ERPs at a range of electrodes. The right- and left-most lateral electrodes have been excluded for space considerations.

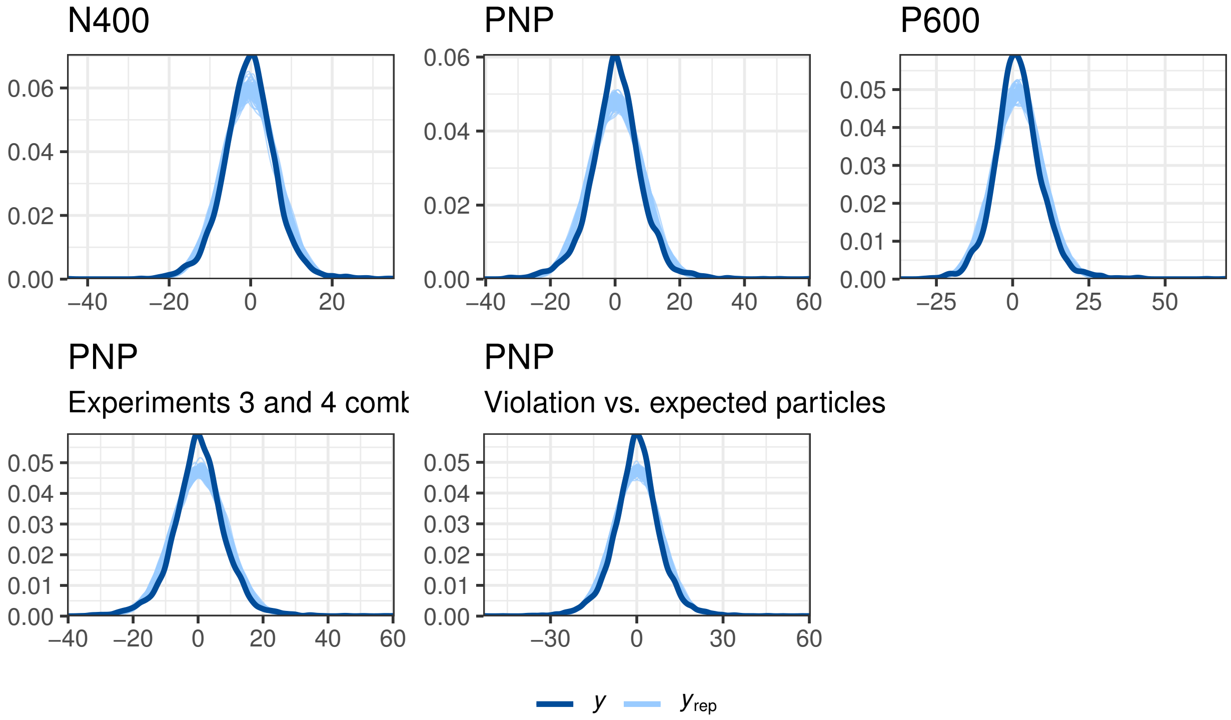


Figure 25: Posterior predictive checks for the models of the N400, PNP, and P600. The dark blue line represents the observed data and the light blue lines, data predicted by the model.

Chapter 7

7 Conclusions

In this thesis, I have argued that studying the presence of lexical predictions over distance contributes to the distinguishing of predictive processing from lexical priming and adds to knowledge about the depth and specificity of predictive processing. Four experiments tested the hypothesis that when encountering a verb in a context where a particle is necessary, native German speakers will preactivate and, in some cases, predict the lexical entries of probable particles in advance of the particle being seen.

Experiment 1 was a self-paced reading study testing the existence of particle preactivation by manipulating the predictability of a verb particle and varying the distance at which it was separated from its verb. The goal was to show that differences in predictability of the particle would result in stronger or weaker preactivation and that stronger preactivation would be more resistant to temporal activation decay. The results of Experiment 1 did not provide any evidence of preactivation, but a subsequent, more sensitive eye tracking experiment (Experiment 2) suggested that the effects of preactivation were apparent in measures of earlier cognitive processing than those to which self-paced reading is sensitive. Experiment 2 used the same experimental materials and found evidence consistent with a facilitatory effect of higher particle predictability in several reading time measures, but evidence was strongest in measures associated with word predictability (Kliegl et al., 2004; Rayner, 1998; Staub, 2015). This was interpreted as evidence in support of particle preactivation in advance of the particle being seen. This is consistent with a variety of models that include activation of plausible upcoming lexical information, context-permitting (Gibson, 1998; Kuperberg, 2016; Lewis & Vasishth, 2005), but contrasts with a previous hypothesis specific to verb-particle constructions in which only the possibility of an upcoming verb particle is entertained, rather than plausible particles preactivated (Piai et al., 2013).

With respect to the question of temporal decay, the lack of evidence for a distance effect on reading times in Experiments 1 and 2 was consistent with models suggesting that decay is not a significant factor in predicting reading times and that previously observed distance effects are better explained by interference from intervening information (Lewandowsky et al., 2009; Vasishth et al., 2019). Furthermore, the lack of evidence for a distance effect was at odds with the surprisal account of sentence processing under which reading times should generally become faster with increasing distance (Levy, 2008). The results therefore suggest that simply increasing the pressure to resolve a predicted syntactic dependency by increasing the distance between dependent elements is not sufficient to speed up reading if the intervening information does not provide additional contextual constraint, as was the case in Experiments 1 and 2. Such a possibility has previously been proposed and the results of Experiment 2 add evidence in support of this proposition (Grodner & Gibson, 2005). In sum, the reading time results were consistent with the lexical preactivation of verb particles before they were seen. To extend these results, Experiments 3 and 4 addressed the question of whether specific lexical predictions about the particle's identity were generated using event-related potentials (ERPs).

In Experiment 3, verb particle predictability was again manipulated, but this time strictly controlled the range of plausible particles to either one or a small set of two or more (similar to the medium and high contextual constraint manipulations used in prediction research). We hypothesised that having only one plausible particle option would encourage readers to commit to a specific particle, whereas having at least one strong competitor for the particle position would discourage lexical predictions. In support of this, we saw a post-N400 positivity (PNP) to violations of the expected particles in the condition where there commitment was more likely (1 plausible particle) relative to the condition where a commitment may not have been made (2+ plausible particles). This suggested that the target particle had been lexically predicted in the 1-particle condition allowing commitment to and deeper processing of a particular sentence representation which was more difficult to recover from when violated. However, statistical evidence was inconclusive. The result was consistent with previous research finding that the PNP may index the cost of recovery from disconfirmed predictions in high vs. low constraint conditions, although the PNP is not thought to be elicited by implausible words (DeLong et al., 2014a; Delong et al., 2011; Federmeier et al., 2007; Kuperberg & Wlotko, 2019; Ness & Meltzer-Asscher, 2018; Thornhill & Van Petten, 2012; Van Petten & Luka, 2012). Given the inconclusiveness of the statistical evidence for the PNP, an attempt to replicate Experiment 3 was made using a larger sample with more than double the number of target trials per participant.

Experiment 4 failed to replicate the PNP finding from Experiment 3, although effects for the other major components, the N400 and P600, were successfully replicated. The data from both experiments combined provided some evidence that a PNP was elicited in both particle violation conditions relative to expected particles, but that the amplitude of the PNP did not differ in line with the 1-particle/2+particle manipulation. This contrasted with previous research suggesting that the PNP is not elicited by implausible vs. expected words (DeLong et al., 2014a; Kuperberg & Wlotko, 2019). This suggested either than the violation particles were not as implausible as thought, or that the PNP may indeed be elicited by implausible words. In either of these cases, the lack of a constraint effect similar to previous research may have been due to the 1-particle/2+particle distinction not being large enough to elicit observable differences. This in turn has its own interesting implications, such as that particles were predicted in both 1- and 2+particle conditions, or that lexical prediction is a graded rather than an all-or-nothing phenomenon.

In conclusion, the experiments presented here did not provide evidence either for or against specific lexical predictions in German verb-particle constructions, but did provide evidence consistent with the preactivation of lexical information about plausible particles in advance of their being seen. This preactivation appeared to survive long distances, arguing against a transient lexical priming effect. Future avenues of research include a more fine-grained definition of how and with what resources preactivated information is maintained, how contextual information guides preactivation towards a specific sentence representation, and whether lexical prediction is best characterised as a strong probabilistic mental sentence representation or as a separate phenomenon triggered by a sufficiently strong probabilistic representation. A further exciting avenue for research on predictive processing is characterisation of the PNP and its relation to reanalysis processes under varying levels of contextual constraint. Implications for current models of sentence processing may be the inclusion of long-distance lexical preactivation, or the possibility that preactivation of lexical material can facilitate reading times and ERP amplitude without commitment to a specific lexical item.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518.

Bader, M. (2012). Complex center-embedded relative clauses in German. *Goethe-Universität Frankfurt*.

Bader, M., Bayer, J., & Häussler, J. (2003). Explorations of centerembedding and missing VPs. *Poster presented at the 16th CUNY conference on sentence processing, MIT, cambridge, MA.[2729 march 2003]*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355. https://doi.org/10.1016/0013-4694(85)90008-2

Besson, M., & Macar, F. (1987). An Event-Related Potential Analysis of Incongruity in Music and Other Non-Linguistic Contexts. *Psychophysiology*, *24*(1), 14–25. https://doi.org/10.1111/j.1469-8986.1987.tb01853.x

Betancort, M., Carreiras, M., & Sturt, P. (2009). The processing of subject and object relative clauses in Spanish: An eye-tracking study. *The Quarterly Journal of Experimental Psychology*, *62*(10), 1915–1929. https://doi.org/10.1080/17470210902866672

Bock, K. (1986a). Meaning, Sound, and Syntax: Lexical Priming in Sentence Production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 575–586.

Bock, K. (1986b). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Booij, G. (2002). Separable complex verbs in Dutch: A case of periphrastic word formation.

In N. Dehé, R. Jackendoff, A. McIntyre, & S. Urban (Eds.), *Verb-particle explorations* (pp. 21–41). Walter de Gruyter.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. https://doi.org/10.1016/J.BRAINRESREV.2008.05.003

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1). https://doi.org/10.16910/jemr.2.1.1

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., . . . Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, *2*(4), 597–620. https://doi.org/10.1007/s11168-004-7431-3

Brehm, L., & Goldrick, M. (2017). Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(10), 1537–1556. https://doi.org/10.1037/xlm0000390

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*. https://doi.org/10.1016/j.jml.2016.10.002

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34

Brown, C., Hagoort, P., & Keurs, M. ter. (1999). Electrophysiological Signatures of Visual Lexical Processing: Open-and Closed-Class Words. *Journal of Cognitive Neuroscience*, *11*(3), 261–281. https://doi.org/10.1162/089892999563382

Buerkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Cappelle, B., Shtyrov, Y., & Pulvermüller, F. (2010). Heating up or cooling up the brain?

MEG evidence that phrasal verbs are lexical units. *Brain and Language*, *115*(3), 189–201. https://doi.org/10.1016/j.bandl.2010.09.004

Chang, F. (2002). *Symbolically speaking: A connectionist model of sentence production* (Vol. 26). https://doi.org/10.1016/S0364-0213(02)00079-4

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chomsky, N. (1970). Remarks on nominalization. In R. A. Jacobs & P. S. Rosenbaum (Eds.), *Readings in English transformational grammar*. Waltham, MA: Ginn.

Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, *0*(0), 1–26. https://doi.org/10.1080/23273798.2018.1427878

Christiansen, M. H., & Chater, N. (1999). Toward a Connectionist Model of Recursion in Human Linguistic Performance. *Cognitive Science*, *23*(2), 157–205. https://doi.org/10.1207/s15516709cog2302_2

Chwilla, D. J., & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research*, *25*(3), 589–606. https://doi.org/10.1016/j.cogbrainres.2005.08.011

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). *Visualizing and Measuring the Geometry of BERT*. Retrieved from http://arxiv.org/abs/1906.02715

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21–58. https://doi.org/10.1080/016909698386582

Czypionka, A., Golcher, F., Błaszczak, J., & Eulitz, C. (2019). When verbs have bugs: Lexical and syntactic processing costs of split particle verbs in sentence comprehension. *Language, Cognition and Neuroscience*, *34*(3), 326–350. https://doi.org/10.1080/23273798.2018.1539756

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, *41*(1), 60–74. https://doi.org/10.1111/1469-8986.00120

Dehé, N., Jackendoff, R., McIntyre, A., & Urban, S. (2002). *Verb-Particle Explorations.*

Berlin, New York: Walter de Gruyter.

DeLong, K. A., Quante, L., & Kutas, M. (2014a). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*(1). https://doi.org/10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Troyer, M., & Kutas, M. (2014b). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Linguistics and Language Compass*, *8*(12). https://doi.org/10.1111/lnc3.12093

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Delong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

Drummond, A. (2016). *Ibex: Software for psycholinguistic experiments.* Retrieved from https://github.com/addrummond/ibex

Ehrenhofer, L., Lau, E., & Colin Phillips. (2019). A possible cure for "N400 blindness" to role reversal anomalies in sentence comprehension. *Submitted.*

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225. https://doi.org/10.1007/BF00114844

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *In press.* https://doi.org/10.31234/osf.io/w2ckt

Falk, S., & Öhl, P. (2010). *Syntactic Characteristics of Particle Verbs : Empirical Evidence for Complex Predicate Processing in German.*

Farmer, T. A., Yan, S., Bicknell, K., & Tanenhaus, M. K. (2015). Form-to-expectation matching effects on first-pass eye movement measures during reading. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 958–976. https://doi.org/10.1037/xhp0000054

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior.*

Retrieved from http://books.google.com?id=VMhJDwAAQBAJ

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., Ochoa-Dewald, E. D., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, (1146), 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Ferreira, F., & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. *Current Directions in Psychological Science*, *27*(6), 443–448. https://doi.org/10.1177/0963721418794491

Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 555–568. https://doi.org/10.1037/0278-7393.16.4.555

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*(6), 725–745. https://doi.org/10.1016/0749-596X(91)90034-H

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2001). Syntactic Working Memory and the Establishment of Filler-Gap Dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research*, *30*(3), 321–338. https://doi.org/10.1023/A:1010447102554

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, *105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 878–883. Sofia, Bulgaria: Association for Computational Linguistics.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-Linguistic Differences in

Processing Double-Embedded Relative Clauses: Working-Memory Constraints or Language Statistics? *Cognitive Science*, *40*(3), 554–578. https://doi.org/10.1111/cogs.12247

Fraser, B. (1976). *The verb-particle combination in English.* Academic Press New York.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing. Psychological, computational and theoretical perspectives* (pp. 129–189). Cambridge: Cambridge University Press.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping Rules for Clinical Trials Incorporating Clinical Opinion. *Biometrics*, *40*(3), 575–586. https://doi.org/10.2307/2530902

Frost, R., Deutsch, A., Gilboa, O., Tannenbaum, M., & Marslen-Wilson, W. (2000). Morphological priming: Dissociation of phonological, semantic, and morphological factors. *Memory & Cognition*, *28*(8), 1277–1288. https://doi.org/10.3758/BF03211828

Futrell, R., & Levy, R. (2016). *Noisy-context surprisal as a human sentence processing cost model. 1*(Section 2), 688–698.

Garnsey, S. M. (1993). *Event-related brain potentials in the study of language: An introduction: Language and Cognitive Processes: Vol 8, No 4.* Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01690969308407581

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language*, *37*(1), 58–93. https://doi.org/https://doi.org/10.1006/jmla.1997.2512

Gaston, P., Lau, E., & Phillips, C. (2019). Syntactic category does not inhibit lexical competition. *ERA*. https://doi.org/10.7939/r3-1t0d-5833

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gernsbacher, M. A. (1991). Cognitive processes and mechanisms in language comprehension: The structure building framework. In *The psychology of learning and motivation* (pp. 217–264). New York, NY: Academic Press.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

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

Gibson, E., & Thomas, J. (1999). Memory Limitations and Structural Forgetting: The Perception of Complex Ungrammatical Sentences as Grammatical. *Language and Cognitive Processes*, *14*(3), 225–248. https://doi.org/10.1080/016909699386293

Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155. https://doi.org/10.1080/01690965.2010.536656

Gimenes, M., Rigalleau, F., & Gaonac'h, D. (2009). The effect of noun phrase type on working memory saturation during sentence comprehension. *European Journal of Cognitive Psychology*, *21*(7), 980–1000. https://doi.org/10.1080/09541440802469523

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96–106. https://doi.org/10.1080/17470215808416261

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic Complexity in Ambiguity Resolution. *Journal of Memory and Language*, *46*(2), 267–295. https://doi.org/10.1006/jmla.2001.2808

Hagoort, P., Baggio, G., & Willems, R. M. (2009). *Semantic unification.* Retrieved from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_64579

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Hagoort, P., Wassenaar, M., & Brown, C. (2003). Syntax-related ERP-effects in Dutch. *Cognitive Brain Research*, *16*(1), 38–50. https://doi.org/10.1016/S0926-6410(02)00208-2

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*, 1–8. https://doi.org/10.3115/1073336.

1073357

Hale, J. (2006). Uncertainty About the Rest of the Sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Häussler, J., & Bader, M. (2015). An interference account of the missing-VP effect. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00766

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology*, *30*(1), 47–61. https://doi.org/10.1111/j.1469-8986.1993.tb03204.x

Howes, D., & Osgood, C. E. (1954). On the Combination of Associative Probabilities in Linguistic Contexts. *The American Journal of Psychology*, *67*(2), 241–258. https://doi.org/10.2307/1418626

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3–27. https://doi.org/10.1016/S0010-0277(03)00124-0

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PloS One*, *9*(7), e100986. https://doi.org/10.1371/journal.pone.0100986

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *0*, 1–12. https://doi.org/10.1080/23273798.2016.1242761

Jaeger, T. F., Fedorenko, E., Hofmeister, P., & Gibson, E. (2008). Expectation-based syntactic processing: Anti-locality outside of head-final languages. *Oral Presentation at CUNY.*

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

Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science*, *20*(2), 137–194. https://doi.org/10.1207/s15516709cog2002_1

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kliegl, R., Grabner, E., Rolfs, M., & 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. https://doi.org/10.1080/09541440340000213

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the Mind During Reading: The Influence of Past, Present, and Future Words on Fixation Durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35. https://doi.org/10.1037/0096-3445.135.1.12

Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. https://doi.org/10.1080/23273798.2018.1524500

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645. https://doi.org/10.1023/A:1026528912821

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

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https:

//doi.org/10.1177/1745691611406925

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 326.

Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5). https://doi.org/10.1080/23273798.2015.1130233

Kuperberg, G., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1). https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G., & Wlotko, E. (2019). A Tale of Two Positivities (and the N400): Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*. https://doi.org/10.1101/404780

Kuperberg, G., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7

Kush, D., Dillon, B., Eik, R., & Staub, A. (2019). Processing of Norwegian complex verbs: Evidence for early decomposition. *Memory & Cognition*, *47*(2), 335–350. https://doi.org/10.3758/s13421-018-0870-0

Kutas, M. (2018). *Electrifying Psycholinguistics: A Historical Perspective on Cognitive Electrophysiology, Expectancy, and Language*. Presented at the CUNY Conference on Human Sentence Processing 2018, University of California, Davis.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980a). Reading between the lines: Event-related brain potentials during natural sentence processing. *Brain and Language*, *11*(2), 354–373. https://doi.org/10.1016/0093-934X(80)90133-9

Kutas, M., & Hillyard, S. A. (1980b). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.

7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Retrieved from http://books.google.com?id=r6nOYYtxzUoC

Lau, E., Holcomb, P. J., & Kuperberg, G. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3). https://doi.org/10.1162/jocn_a_00328

Lau, E., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra: Psychology*, *2*(1).

Lau, E., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)Constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lee, M., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. https://doi.org/10.1017/CBO9781139087759

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., & Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00229

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222. https://doi.org/10.1016/j.jml.2012.02.005

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, *69*(4), 461–495. https://doi.org/10.1016/j.jml.2012.10.005

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, *13*(3), 120–126. https://doi.org/10.1016/j.tics.2008.12.003

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/

s15516709cog0000_25

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*(1), e99–e108. https://doi.org/10.1353/lan.2019.0015

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, *40*(6). https://doi.org/10.1111/cogs.12274

Logačev, P., & Vasishth, S. (2013). Em2: A package for computing reading time measures for psycholinguistics. (Version 0.9). Retrieved from https://github.com/cran/em2

Luck, S. J. (2005a). Ten Simple Rules for Designing and Interpreting ERP Experiments. In T. C. Handy (Ed.), *Event-related Potentials: A Methods Handbook* (pp. 17–32). MIT press.

Luck, S. J. (2005b). *The Event-Related Potential Technique in Cognitive Neuroscience*. Iowa: MIT Press.

Luck, S. J., & Gaspelin, N. (2016). How to Get Statistically Significant Effects in Any ERP Experiment (and Why You Shouldn't). *Psychophysiology*, *4424*. https://doi.org/10.1111/psyp.12639

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Lupyan, G., & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, *24*(4), 279–284. https://doi.org/10.1177/0963721415570732

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54. https://doi.org/10.1037/0033-295X.109.1.35

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Association for computational linguistics (ACL) system demonstrations*, 55–60. Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

Martens, S. (2013). TüNDRA: A Web Application for Treebank Search and Visualization. *Proceedings of The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, 133–144. Retrieved from http://bultreebank.org/TLT12/TLT12Proceedings.pdf

Mathot, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(4), 913–934. https://doi.org/10.1037/a0022964

Mayer, E., Dillon, B., & Staub, A. (2019). *The (non-)influence of even's likelihood-based presupposition on lexical predictability effects.* Presented at the Psycholinguistics in Iceland and Prediction.

McClelland, J. L., & O'Regan, J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 634–644. https://doi.org/10.1037/0096-1523.7.3.634

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*(6), 578–586. https://doi.org/10.3758/BF03203972

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.* Retrieved from http://arxiv.org/abs/1902.01007

McDonald, S. A., & Shillcock, R. C. (2003). Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(6), 1155–1172. https://doi.org/10.1037/0278-7393.18.6.1155

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Metzner, P., von der Malsburg, T., Vasishth, S., & Rösler, F. (2017). The Importance of

Reading Naturally: Evidence From Combined Recordings of Eye Movements and Electric Brain Potentials. *Cognitive Science*, *41*(S6), 1232–1263. https://doi.org/10.1111/cogs.12384

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 69–89). Hillsdale, N.J.: Erlbaum.

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813.

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

Nakatani, K., & Gibson, E. (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, *46*(1), 63–87. https://doi.org/10.1515/LING.2008.003

Ness, T., & Meltzer-Asscher, A. (2017). Working Memory in the Processing of Long-Distance Dependencies: Interference and Filler Maintenance. *Journal of Psycholinguistic Research*, *46*(6), 1353–1365. https://doi.org/10.1007/s10936-017-9499-6

Ness, T., & Meltzer-Asscher, A. (2018). Lexical inhibition due to failed prediction: Behavioral evidence and ERP correlates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(8), 1269–1285. https://doi.org/10.1037/xlm0000525

Ness, T., & Meltzer-Asscher, A. (2019). When is the verb a potential gap site? The influence of filler maintenance on the active search for a gap. *Language, Cognition and Neuroscience*, *34*(7), 936–948. https://doi.org/10.1080/23273798.2019.1591471

Nicenboim, B. (2018). *Eeguana: A package for manipulating EEG data in R.* Retrieved from https://github.com/bnicenboim/eeguana

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

Nicenboim, B., Logacev, P., Gattei, C., & Vasishth, S. (2015). When high - capacity readers slow down and low - capacity readers speed up : Working memory and locality effects. *Frontiers in Psychology*, 1–22. https://doi.org/10.3389/fpsyg.2016.00280

Nicenboim, B., Vasishth, S., & Rösler, F. (2019). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *PsyArXiv*.

Nieuwland, M. S., Barr, D., Bartolozzi, F., Busch-Moreno, S., Donaldson, D., Ferguson, H. J., . . . Von Grebmer Zu Wolfsthurn, S. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*. Retrieved from http://dx.doi.org/10.1101/267815

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., . . . Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Osterhout, L. (1999). A Superficial Resemblance Does Not Necessarily Mean You Are Part of the Family: Counterarguments to Coulson, King and Kutas (1998) in the P600/SPS-P300 Debate. *Language and Cognitive Processes*, *14*(1), 1–14. https://doi.org/10.1080/016909699386356

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786–803. https://doi.org/10.1037/0278-7393.20.4.786

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the Language Specificity of the Brain Response to Syntactic Anomalies: Is the Syntactic Positive Shift a Member of the P300 Family? *Journal of Cognitive Neuroscience*, *8*(6), 507–526. https://doi.org/10.1162/jocn.1996.8.6.507

Otten, M., & Van Berkum, J. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Otten, M., Nieuwland, M. S., & Van Berkum, J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(1), 89. https://doi.org/10.1186/1471-2202-8-89

Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, *52*(11), 1456–1469. https://doi.org/10.1111/psyp.12515

Phillips, C., Kazanina, N., & Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research*, *22*(3), 407–428. https://doi.org/10.1016/j.cogbrainres.2004.09.012

Piai, V., Meyer, L., Schreuder, R., & 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. https://doi.org/10.1016/j.bandl.2013.09.015

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940–961. https://doi.org/10.1037/0278-7393.24.4.940

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693. https://doi.org/10.1038/s41562-018-0406-4

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology (2006)*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K. (2014). The gaze-contingent moving window in reading: Development and review. *Visual Cognition*, *22*(3-4), 242–258. https://doi.org/10.1080/13506285.2013.879084

Rayner, K., & 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. https://doi.org/10.3758/BF03197692

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

Roark, B., & 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. https://doi.org/10.3115/1699510.1699553

Rohde, D. (2003). Linger: A flexible platform for language processing experiments (Version 2.94). Retrieved from https://tedlab.mit.edu/~dr/Linger/

Rugg, M. D., & Barrett, S. E. (1987). Event-related potentials and the interaction between orthographic and phonological information in a rhyme-judgment task. *Brain and Language*, *32* (2), 336–361. https://doi.org/10.1016/0093-934X(87)90132-5

Rumelhart, D. E., & McClelland, J. L. (1986a). On learning the past tenses of English. In D. E. Rumelhart, J. L. McClelland, & P. R. Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition1* (Vol. 2, pp. 216–217). MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986b). *Parallel Distributed Processing.*

Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates : Evidence against the expectation-based account. *Frontiers in Psychology*, 1–21. https://doi.org/10.3389/fpsyg.2016.00403

Sassenhagen, J., & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage*, *200*, 425–436. https://doi.org/10.1016/j.neuroimage.2019.06.048

Schad, D., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow: A tutorial for cognitive science. *arXiv.* https://doi.org/None

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96* (4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Singer, M., & Ferreira, F. (1983). Inferring consequences in story comprehension. *Journal of Verbal Learning and Verbal Behavior*, *22* (4), 437–448. https://doi.org/10.1016/S0022-5371(83)90282-7

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33* (33).

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128* (3). https://doi.org/10.1016/j.cognition.2013.02.013

Smolka, E., Gondan, M., & Rösler, F. (2015). Take a Stand on Understanding: Electrophysiological Evidence for Stem Access in German Complex Verbs. *Frontiers in Human*

Neuroscience, *9*(62). https://doi.org/10.3389/fnhum.2015.00062

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *157*(3), 357–387. https://doi.org/10.2307/2983527

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002

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. https://doi.org/10.1111/lnc3.12151

Staub, A., & Goddard, K. (2019). The role of preview validity in predictability and frequency effects on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 110–127. https://doi.org/10.1037/xlm0000561

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. *Expert Systems*, *17*(1), 41–56. https://doi.org/10.1111/1468-0394.00126

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Bulletin*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Team, R. C. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language*, *47*(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

Twomey, K. E., Chang, F., & Ambridge, B. (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology*, *73*(0), 41–71. https://doi.org/http://dx.doi.org/10.1016/j.cogpsych.2014.05.001

Ueno, M., & Garnsey, S. M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Language and Cognitive Processes*, *23*(5), 646–688. https://doi.org/10.1080/01690960701653501

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Towards a New Functional Anatomy of Language*, *92*(12), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Van Berkum, J., Brown, C., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Van Berkum, J., Hagoort, P., & Brown, C. (1999). Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671. https://doi.org/10.1162/089892999563724

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166. https://doi.org/10.1016/j.jml.2006.03.007

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263. https://doi.org/10.1016/j.jml.2011.05.002

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, *19*(1), 95–112. https://doi.org/10.3758/BF03198500

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center* *embeddings*. Routledge.

Vasishth, S., & Drenhaus, H. (2011). Locality in German. *Dialogue and Discourse*, *2*(1), 59–82.

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

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese Relative Clauses: Evidence for the Subject-Relative Advantage. *PLOS ONE*, *8*(10), e77006. https://doi.org/10.1371/journal.pone.0077006

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). *Computational models of retrieval processes in sentence processing* [Preprint]. https://doi.org/10.31234/osf.io/e4jds

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567. https://doi.org/10.1080/01690960903310587

von der Malsburg, T., & 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. https://doi.org/10.1016/j.jml.2016.10.003

von der Malsburg, T., Poppels, T., & Levy, R. (2019). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 US and 2017 UK election. *Psychological Science*. https://doi.org/https://psyarxiv.com/n5ywr/

Wagers, M. W., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112. https://doi.org/10.1016/S0010-0277(02)00087-2

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, *103*(4), 761–772. https://doi.org/10.1037/0033-295X.103.4.761

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender

Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wicha, N. Y. Y., Moreno, E. M., Kutas, M., Jolla, L., & Related, E. (2003). *Special Section Expecting Gender : An Event Related Brain Potential Study on the Role of Grammatical Gender in Comprehending a Line Drawing Within*. 483–508.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). *What do RNN Language Models Learn about Filler-Gap Dependencies?* Retrieved from http://arxiv.org/abs/1809.00042

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. https://doi.org/10.1080/23273798.2014.995679

Xiang, M., Dillon, B., Wagers, M. W., Liu, F., & Guo, T. (2014). Processing covert dependencies: An SAT study on Mandarin wh-in-situ questions. *Journal of East Asian Linguistics*, *23*(2), 207–232. https://doi.org/10.1007/s10831-013-9115-1

Zeller, J. (2001). *Particle verbs and local domains*. Amsterdam, Philadelphia: J. Benjamins.