Journal of Memory and Language 132 (2023) 104435

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Lexical prediction does not rationally adapt to prediction error: ERP evidence from pre-nominal articles

Elise van Wonderen [a,b,1], Mante S. Nieuwland [a,c,1]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[b] Radboud University, Nijmegen, The Netherlands
[c] Donders Institute for Cognition, Brain and Behaviour, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Predictive validity
Cue validity
Discourse comprehension
N400
Rational adaptation
Expectation adaptation
Bayesian optimality
Pre-nominal prediction effect
Gender mismatch
Belief updating

ABSTRACT

People sometimes predict upcoming words during language comprehension, but debate remains on when and to what extent such predictions indeed occur. The rational adaptation hypothesis holds that predictions develop with expected utility: people predict more strongly when predictions are frequently confirmed (low prediction error) rather than disconfirmed. However, supporting evidence is mixed thus far and has only involved measuring responses to supposedly predicted nouns, not to preceding articles that may also be predicted. The current, large-sample (N = 200) ERP study on written discourse comprehension in Dutch therefore employs the well-known 'pre-nominal prediction effect': enhanced N400-like ERPs for articles that are unexpected given a likely upcoming noun's gender (i.e., the neuter gender article 'het' when people expect the common gender noun phrase 'de krant', *the newspaper*) compared to expected articles. We investigated whether the pre-nominal prediction effect is larger when most of the presented stories contain predictable article-noun combinations (75% predictable, 25% unpredictable) compared to when most stories contain unpredictable combinations (25% predictable, 75% unpredictable). Our results show the pre-nominal prediction effect in both contexts, with little evidence to suggest that this effect depended on the percentage of predictable combinations. Moreover, the little evidence suggesting such a dependence was primarily observed for unexpected, neuter-gender articles ('het'), which is inconsistent with the rational adaptation hypothesis. In line with recent demonstrations (Nieuwland, 2021a,b), our results suggest that linguistic prediction is less 'rational' or Bayes optimal than is often suggested.

Introduction

People sometimes casually and implicitly predict upcoming words based on the meaning of a story or conversation. Such lexical predictions, along with predictions at other levels of representation, play a prominent role in current theories of language comprehension (e.g., Kuperberg & Jaeger, 2016; Kutas, DeLong & Smith, 2011; Pickering & Gambi, 2018). According to the 'rational adaptation hypothesis' (Kuperberg & Jaeger, 2016), lexical predictions come with costs and benefits, and people manage this trade-off by adapting predictions to their expected utility: predictions strengthen when frequently confirmed and weaken when frequently disconfirmed. In this article, we briefly review theoretical background and empirical evidence, and then we present an ERP experiment on Dutch mini-story comprehension to investigate whether probability of confirmed/disconfirmed predictions impacts an established ERP signature of lexical prediction.

Lexical prediction and rational adaptation

While there seems to be a general consensus that people can and sometimes do predict upcoming words during comprehension, there is ongoing debate as to when or under which circumstances such predictions occur. Strong proponents argue that prediction is an integral mechanism to propel incremental language comprehension and that people continuously predict and update predictions at all levels of representation (Altmann and Mirković, 2009; Dell & Chang, 2014; Pickering & Garrod, 2007, 2013). In contrast, skeptics view prediction not as a fundamental, organizing principle of language comprehension, but as essentially a by-product of successful comprehension (e.g., Huettig, 2015; Huettig & Mani, 2016; Martin, 2016, 2020). Ongoing debate also concerns whether people actively generate hypotheses about specific upcoming words (for discussion, see Baggio, 2018; Kutas et al., 2011; Van Berkum, 2009; Van Petten & Luka, 2012), or merely passively pre-

E-mail addresses: e.vanwonderen@uva.nl (E. van Wonderen), mante.nieuwland@mpi.nl (M.S. Nieuwland).
[1] Shared first-authorship.

https://doi.org/10.1016/j.jml.2023.104435
Received 13 April 2023; Received in revised form 23 May 2023; Accepted 26 May 2023
Available online 10 June 2023

activate semantic content as it naturally emerges from a context representation. Lexical predictions are sometimes viewed as rather exceptional and limited to highly constraining laboratory circumstances that might not be representative of natural language processing (e.g., Huettig, 2015; Huettig & Guerra, 2019; Luke & Christianson, 2016; Nieuwland, 2019).

In an attempt to find middle ground, Kuperberg and Jaeger (2016) suggested that the variable nature of lexical prediction can be understood by considering its potential costs and benefits. Correct predictions are thought to benefit word recognition, which helps dealing with distorted or incomplete input (e.g., Norris, McQueen & Cutler, 2016; Pickering & Garrod, 2007; Strauß, Kotz & Obleser, 2013) and may increase processing efficiency by reducing the processing load to deviations from predicted input (prediction error; e.g., Clark, 2013; Friston, 2005; but see Kwisthout & van Rooij, 2020). However, predictions may also harbor costs. Incorrect predictions incur a processing cost compared to when no predictions are generated (e.g., Van Petten & Luka, 2012), possibly because they require inhibition of a predicted word (e.g., Ness & Meltzer-Asscher, 2021; but see also Cevoli, Watkins & Rastle, 2022; Frisson, Harvey & Staub, 2017; Luke & Christianson, 2016). Moreover, maybe the generation of lexical predictions itself incurs a "metabolic cost" (Kuperberg & Jaeger, 2016). These supposed costs could help to explain why lexical predictions are only observed in highly constraining contexts and when people have sufficient time and processing resources available (e.g., Huettig & Janse, 2016; Ito, Corley, & Pickering, 2018; Ito, Corley, Pickering, Martin & Nieuwland, 2016). In other words, the variable nature of lexical prediction may thus reflect people's engagement in prediction to balance its costs and benefits, by adapting predictions to the probability of prediction error. People may engage less in prediction when predictions are frequently disconfirmed rather than confirmed (Clark, 2013; Kuperberg & Jaeger, 2016).

In this view, dubbed the 'rational adaptation hypothesis' of prediction (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019; Kuperberg & Jaeger, 2016), lexical prediction is one example of an adaptive, probabilistic process based on rational Bayesian principles, whereby people capitalize on statistical regularities in their environment to guide behavior. Rational adaptation is a general theoretical framework that describes how individuals adapt rationally to a utility function given constraints imposed by their cognitive architecture (limited resources) and the local task environment (e.g., Anderson, 1991; Howes, Lewis, & Vera, 2009; Lieder & Griffiths, 2020). Rational adaptation has been used to explain a wide range of phenomena in domains such as speech perception (e.g., Kleinschmidt & Jaeger, 2015; Norris, McQueen & Cutler, 2003), syntactic processing[2] (e.g., Fine, Jaeger, Farmer & Qian, 2013; Myslín & Levy, 2016), semantic processing (e.g., Delaney-Busch et al., 2019; Gibson, Bergen, Piantadosi, 2013) and pragmatics (e.g., Degen, Hawkins, Graf, Kreiss & Goodman, 2020; Goodman & Frank, 2016; Roettger & Franke, 2019).

The rational adaptation hypothesis of prediction assumes a probabilistic, predictive language comprehension system that continuously represents a prior probability distribution of the next word (see also Levy, 2008). A wide distribution with similarly low probability for many different words corresponds to a weak prediction, whereas a strongly peaked distribution with high probability of a small set of words corresponds to a strong prediction. The occurrence of the next word then changes the prior distribution to a posterior distribution in a process called *belief updating*, and the difference between the prior and posterior distribution is called *prediction error* (or Bayesian surprise). The posterior distribution then acts as the prior distribution for the next upcoming word.

Crucially, the rational adaptation hypothesis asserts that people inform their predictions with the reliability of the prior distribution as estimated from previous encounters. This history of reliability is sometimes referred to as *predictive cue validity*, which expresses the extent to which a stimulus is a valid cue to predict a target word. For example, when strong predictions are always disconfirmed, reliability of the prior distribution is low, and people could weaken their predictions accordingly (see also Yan, Kuperberg and Jaeger, 2017). Under the assumption that predictions are costly to generate, it is considered 'rational' in this situation not to generate any predictions that always turn out futile. Reversely, when predictions are always confirmed, reliability of the prior is high, and people may strengthen their predictions accordingly.

In sum, the rational adaptation hypothesis holds that people adapt to statistical regularities in the environment, thereby exerting strategic, rational control over their predictive behavior. Prediction strength may therefore change in any given experiment, as part of a continuous learning process. This hypothesis has previously been tested in two ways. The simplest and most common *block*-approach tests whether predictions are stronger in trial blocks with a high proportion of prediction-confirming stimuli than in blocks with a low proportion (e.g., Lau, Holcomb, & Kuperberg, 2013; Brothers, Swaab, & Traxler, 2017). While this approach inherently assumes that people learn and adapt to relative proportions of occurrence from sequences of stimuli, it is unable to detect changes over time that may constitute evidence for or against the adaptation hypothesis. Absence of a proportion block-effect is certainly inconsistent with adaptation, but presence of such an effect merely demonstrates adaptation, not necessarily rational adaptation, because the proportion effect could arise from an underlying pattern that is itself inconsistent with rational adaptation. Instead, the more fine-grained *trial*-approach takes into account estimated prediction success at each point in time (e.g., Delaney-Busch et al., 2019). This explicitly addresses the development of predictions over time and stipulates how people learn relative proportions over sequences of stimuli on a trial-by-trial basis.

The below section reviews the experimental evidence from the block- and trial-approach for rational adaptation.

Rational adaptation of lexical prediction: mixed evidence

Several studies have investigated adaptation of prediction using word pairs. Lau et al. (2013) focused on N400 effects from associative word priming, observing a reduced N400 when targets follow associatively related primes (e.g., salt-pepper) compared to unrelated primes. This effect was stronger in trial blocks with 50% related trials than in blocks with only 10% related trials, suggesting a relatedness proportion effect in semantic priming which has often been reported in behavioral measures (e.g., Hutchison, 2007; see also Brown, Hagoort & Chwilla, 2000; Holcomb, 1988). Such a pattern demonstrates adaptation albeit not necessarily rational adaptation. Subsequent re-analysis with a by-trial approach by Delaney-Busch et al. (2019) suggested that these N400 patterns arose from rational adaptation. They used a rational adaptor model that weighted the probability of a target word with the prime (from association norms) and without the prime (from corpus-based lexical frequency) by the estimated probability of a related trial at a given moment in time. Increase in the proportion of related trials presumably strengthened the expectation of a related trial in the model, effectively increasing the effect of forward association strength and decreasing the effect of lexical frequency. Consistent with rational adaptation, model output was a statistically significant predictor of

[2] People read infrequent (a priori unexpected) syntactic structures such as reduced relative clauses more quickly when they are often repeated, accompanied with slower reading of frequent structures. Such 'syntactic adaptation' may reflect an increased prediction of infrequent structures to minimize associated surprise (e.g., Fine & Jaeger, 2016; Fine, Jaeger, Farmer, & Qian, 2013) along with decreased prediction of frequent structures. However, recent findings suggest that syntactic adaptation is largely driven by a more general task adaptation whereby readers speed up in the experiment, which benefits infrequent structures most (Prasad & Linzen, 2020; see also Nieuwland, 2021b). In other words, syntactic adaptation may not genuinely reflect changes in prediction.

N400 amplitude. However, further re-analysis by Nieuwland (2021b) revealed that the main result was primarily caused by the lexical frequency of unrelated targets,[3] which is inconsistent with rational adaptation of predictions. Moreover, Nieuwland showed that the associative priming effect gradually strengthened in the two blocks in a similar way. In other words, the proportion of related trials did not have any meaningful impact. These findings are not only inconsistent with the stronger claim of rational adaptation but also with the weaker claim regarding a relatedness proportion effect on semantic priming.

Also using a word-word priming paradigm, Ness and Meltzer-Asscher (2021) examined the effects of repeated prediction disconfirmation on the costs associated with disconfirmation. Participants made speeded congruency judgments on trials that were prediction-confirming (high constraint prime, high cloze target), prediction-disconfirming (high constraint prime, low cloze target) or did not involve prediction (low constraint prime, low cloze target). Proportion of prediction confirmation was manipulated between participants in the filler materials. Using a rational adaptor model, Ness and Meltzer-Asscher computed an 'inhibition index' that weighted prediction error (the prime word constraint minus target word cloze) by the estimate of the likelihood of encountering the expected word at each given trial, and reported that the inhibition index was a statistically significant predictor of reaction time. However, re-analyses by Nieuwland (2021a) revealed problems with the original analyses and showed that responses changed during the experiment in a way that did not support rational adaptation, and that, like the results of Lau et al., also did not support the relatedness proportion effect.

Several studies focused on prediction effects during sentence comprehension. For example, Brothers et al. (2017, Experiment 2) compared self-paced reading times to predictable and unpredictable words and the immediately following word, and manipulated the proportion of predictable/unpredictable words between participants (87.5, 50 or 12.5% predictable words). The facilitatory effect of predictability on reading times was smaller with a lower proportion of predictable words and even absent when most sentences were unpredictable. The authors concluded that readers shift away from predictive processing when they notice that many predictions are disconfirmed, although changes during the experiment were not investigated. The pattern of presumed adaptation therefore remained unknown. However, Brothers et al. did try to determine whether participants adapted only 'locally', namely to the predictability of the immediately preceding trial. In a post-hoc analysis, they did not find a statistically significant effect of the previous trial. From this null-result, they concluded that people adapted to the 'global statistical regularities of the environment', although they did not further specify 'global' (minimally 2 previous trials).

In an ERP experiment on spoken sentence comprehension, Brothers, Dave, Hoversten, Traxler and Swaab (2019) manipulated the proportion of prediction confirmation across two different speakers: one reliable speaker who produced more predictable than unpredictable sentence-final words (80/20%), and one unreliable speaker producing the reverse (20/80%). Both speakers yielded a predictability effect on the N400, but this effect was larger for reliable speakers. Moreover, the obtained difference was due to an effect of speaker reliability on the ERPs elicited by predictable words (predictable words from the reliable speaker elicited smaller N400s than those from the unreliable speaker), whereas speaker reliability did not appear to impact responses to unpredictable words, or only in an early time window.[4] These patterns suggested that participants kept track of speaker reliability and engaged more in predictive processing for reliable than unreliable speakers (see also Grodner & Sedivy, 2011; Kamide, 2012).

In a follow-up study by Dave, Brothers, Hoversten, Traxler and Swaab (2021), younger adults showed the same pattern observed by Brothers et al. (2019), but older adults did not show the reduced N400 for predictable words spoken by the reliable speaker (i.e., no N400 effect of speaker reliability). Both younger and older adults showed a positive correlation between cognitive control (measured with the Stroop interference effect) and the speaker reliability effect. Dave et al. concluded therefore that rational adaptation could take place via non-linguistic, domain-general executive resources such as cognitive control (for related discussion, see Ryskin, Levy & Fedorenko, 2020; Shain, Blank, Van Schijndel, Schuler & Fedorenko, 2020). As one of several potential explanations, Dave et al. argued that whereas younger and older adults both generated lexical-semantic predictions rapidly and automatically, younger adults may have been better at inhibiting these predictions in the context of an unreliable speaker than older adults. In this 'inhibition failure' account, the inhibition of predictions is presumably too cognitively effortful for older adults to employ ahead of time. As in Brothers et al. (2017, 2019), changes during the experiment were not investigated.

Unlike the work by Brothers and Dave and colleagues, Zhang et al. (2019) did not find the ratio of predictable/unpredictable items to impact prediction-related N400s. They manipulated predictability between experiments/participants (target words were predictable in Experiment 1 and unpredictable in Experiment 2) and manipulated the fillers between blocks/within participants (predictable fillers in block 1, unpredictable/semantically anomalous fillers in block 2). Thus, block 1 and 2 contained respectively 100% and 50% predictable trials in Experiment 1, and 50% vs. 0% in Experiment 2. Whereas predictable words elicited smaller N400s than unpredictable words, there was no evidence for any effect of block, which is at odds with a proportion effect.

To sum up, we consider extant evidence for rational adaptation of lexical prediction to be mixed at best. Furthermore, we emphasize that the reviewed studies all investigated prediction by analyzing responses to the (un)predictable words themselves, which cannot yield strong evidence for lexical prediction to begin with (e.g., Kutas et al., 2011; Pickering & Gambi, 2018). It remains unclear whether, or to what extent, such effects reflect prediction of a specific noun or merely part of its meaning (e.g., Kutas & Hillyard, 1984). Moreover, noun-elicited responses may in part reflect factors other than noun predictability, such as sentence plausibility and semantic similarity to words in the context (see also Fleur, Flecken, Rommers, & Nieuwland, 2020; Nieuwland et al., 2020b). The current study therefore tested for prediction-related effects on pre-nominal words (words that occur *before* the presumably predicted noun, e.g., DeLong, Urbach, & Kutas, 2005; Nieuwland, Arkhipova, & Rodríguez-Gómez, 2020; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2003, 2004).

Pre-nominal prediction effects on ERPs

ERP responses to pre-nominal articles or adjectives can depend on their match with a likely upcoming noun, for example, in terms of grammatical gender (e.g., Van Berkum et al., 2005; Wicha et al., 2003a, b). Because these pre-nominal words are grammatically correct and do not differ in meaning (e.g., 'el/la', 'groot/grote'), the observed ERP effect can be ascribed to the grammatical relationship between the presented pre-nominal word and the predicted - but not yet presented - noun. Different pre-nominal prediction manipulations appear to elicit different types and strengths of effects (for recent in-depth overviews, see Fleur et al., 2020; Nieuwland et al., 2020b; Nieuwland et al., 2020a). Effects associated with the English indefinite articles 'a/an' have proven controversial (Ito, Martin, & Nieuwland, 2017a,b). First reported by DeLong et al. (2005), these articles elicit an N400-like effect when

[3] The Delaney-Busch et al. model computed surprisal from trial history, lexical frequency and association strength but the latter was zero for unrelated words.

[4] However, ERP responses in an early, 100–200 ms time window, did show a difference between unpredictable words spoken by reliable and unreliable speakers.

mismatching the initial sound of a likely upcoming noun (e.g., 'an' when the expected noun is 'kite'; see also Bañón & Martin, 2019). A large-scale replication study with pre-registered analyses by Nieuwland et al. (2018) also yielded a small negative effect, but concluded that this effect is likely too small to observe without large sample sizes. As discussed in Nieuwland et al. (2018, 2020b), tiny effect sizes from large-scale studies do not constitute evidence against lexical prediction per se, but they do raise doubt on whether people routinely or consistently use pre-nominal words to inform their predictions. One potential reason for small effect sizes is that 'a/an' articles are only diagnostic of the next word, which need not be a noun (e.g., 'an old kite'). Unexpected 'a/an' articles thus do not actually refute the upcoming noun altogether.

Effects on gender-marked adjectives have also been difficult to replicate. In a canonical study on spoken discourse comprehension, Van Berkum and colleagues (2005) reported an early positive ERP effect. However, Otten, Nieuwland and Van Berkum (2007) reported a negative effect in a similar study. A high-powered, pre-registered replication study by Nieuwland et al. (2020a) also reported a negative effect, although it was very small.

In contrast, ERP effects from gender-marked *articles* have been relatively robust and consistent. Gender-mismatching articles elicit a negativity in the N400 time window in different languages (e.g., Foucart, Martin, Moreno, & Costa, 2014; Ito, Gambi, Pickering, Fuellenbach, & Husband, 2020; Kochari & Flecken, 2019; Martin, Branzi, & Bar, 2018; Molinaro, Giannelli, Caffarra, & Martin, 2017; Nicenboim, Vasishth, & Rösler, 2020; Wicha et al., 2003a,b; Otten & Van Berkum, 2009), whereas only one study reported a positive ERP effect (Wicha et al., 2004). This consistency is further corroborated by recent work on Dutch mini-story comprehension by Fleur et al. (2020). Fleur and colleagues reported an enhanced N400-like response for pre-nominal articles ('de/het') whose gender mismatched that of likely upcoming definite nouns, compared to matching articles. This pre-nominal prediction effect was highly similar in size in two identical experiments (N = 48 and N = 80 for Experiment 1 and 2, respectively). The current study used this effect to investigate rational adaptation of lexical prediction.

The current study

Our participants read Dutch, two-sentence mini-stories containing definite article-noun combinations as targets, see example stories (1–2) below (and see the complete set of materials on our OSF project page). At the position before the target article, each mini-story strongly suggested a specific article-noun combination as the best continuation (as established in a cloze task, see Methods). The mini-stories contained either the expected target combination (e.g., 'de krant', common gender, the newspaper) or an unexpected combination with a different-gender noun (e.g., 'het dagblad', neuter gender, the daily newspaper) that was somewhat plausible or at least not incoherent with the context.

(1) Ik lees thuis graag het nieuws op papier. Daarom lees ik thuis dus iedere ochtend **de$_{COM}$ krant$_{COM}$/*het$_{NEUT}$ dagblad$_{NEUT}$** tijdens het ontbijt.

*At home I prefer reading the news on paper. That's why every morning at home I read **the$_{COM}$ newspaper$_{COM}$/*the$_{NEUT}$ daily newspaper$_{NEUT}$** at breakfast.*

(2) De politieagenten hadden de verdachte opgepakt. Hij moest direct mee naar **het$_{NEUT}$ bureau$_{NEUT}$/*de$_{COM}$ cel$_{COM}$** van de politie.

*The police officers had arrested the suspect. He immediately had to come to **the$_{NEUT}$ station$_{NEUT}$/*the$_{COM}$ cell$_{COM}$** of the police.*

Our main dependent measure was the pre-nominal prediction effect: the difference in ERPs elicited by the expected and unexpected articles. Notably, this effect is observed (Fleur et al., 2020) despite 'de' and 'het' not being perfectly diagnostic cues to noun gender, because 'de' is also used for plural nouns of either gender ('het boek/de boeken', the book/books), whereas 'het' is also used for diminutive nouns of either gender ('de kast/het kastje', the closet/small closet) and for other grammatical functions (as an expletive pronoun like the English 'it' or as provisional subject).

We tested the adaptation-hypothesis by varying the ratio of predictable/unpredictable stimuli in a between-participants manipulation. Each participant read 60 expected and 60 unexpected target trials from the same set of 120 items, and either read 120 filler mini-stories with *predictable* article-noun combinations (yielding a 75/25% ratio of predictable/unpredictable stimuli in the entire list) or 120 filler mini-stories with *unpredictable* article-noun combinations (a 25/75% ratio of predictable/unpredictable stimuli). The filler mini-stories had been constructed in the same way as the target trials, see example stories 3–4.

(3) Margriets opa houdt erg van zoet en feestelijk eten. Op haar verjaardag eet hij een groot stuk van **de$_{COM}$ taart$_{COM}$/*het$_{NEUT}$ gehaktbrood$_{NEUT}$** met een vorkje.

*Margaret's grandfather really likes sweet and festive food. On her birthday he eats a big piece of **the$_{COM}$ cake$_{COM}$/*the$_{NEUT}$ meatloaf$_{NEUT}$** with a little fork.*

(4) Floris hoorde iemand uit Amerika Nederlands praten. Dat hoorde hij aan **het$_{NEUT}$ accent$_{NEUT}$/*de$_{COM}$ uitspraak$_{COM}$** van de man.

*Floris heard someone from America speak Dutch. He heard that from **the$_{NEUT}$ accent$_{NEUT}$/*the$_{COM}$ pronunciation$_{COM}$** of the man.*

According to the rational adaptation hypothesis (Kuperberg & Jaeger, 2016), prediction strength depends not only on the discourse-contextual constraints but also on the likelihood of prediction error established from the context. Therefore, the pre-nominal prediction effect may be absent or reduced when the presented materials contain mostly unpredictable items (see also the unpublished commentary by Yan et al., 2017[5]). The alternative hypothesis is that predictive processing proceeds without regard for potential prediction error (Nieuwland, 2021a,b), such that prediction strength is primarily a function of discourse-contextual constraints. The pre-nominal prediction effect may then be similar when the presented materials contain mostly predictable or unpredictable items.

We estimated the pre-nominal prediction effect with Bayesian linear mixed-effects regression, and compared the estimate of this effect when all fillers were either predictable or unpredictable. We also examined whether this effect changed during the experiment, and did so dependent on the nature of the fillers.

Although of secondary interest, we also examined ERP responses to the specific form of the article ('de/het') and to expected and unexpected

[5] Yan et al. (2017) offered rational adaptation as a post-hoc explanation for why the large-scale replication study by Nieuwland et al. (2018) observed a smaller effect on 'a/an' articles than the original study by DeLong et al. (2005). Because DeLong et al.'s methods description was incomplete (Nieuwland, 2018), Nieuwland et al. had not included fillers and exposed participants to a greater proportion (50%) of prediction disconfirming sentences compared to DeLong et al. According to Yan et al. (2017), this caused a smaller effect in Nieuwland et al. due to weaker predictions of expected articles and stronger predictions of the unexpected articles (analogous to claims regarding syntactic adaptation). However, they did not test this hypothesis with the available data. Our own re-analysis of Nieuwland et al. yielded clear evidence for increasingly negative article-elicited N400s throughout the experiment ($t = 2.9$, $p < .005$), but not for an accompanying change in the article expectedness effect ($t = 0.2$, $p = 0.83$), hence no support for rational adaptation of prediction.

nouns. In Fleur et al. (2020), 'het' elicited larger expectedness effects than 'de', and unexpected nouns elicited enhanced N400s compared to expected nouns but this effect was rather small (probably because of an early-onset enhanced positive ERP response).

Methods

Participants

Participants were recruited from the participant database of the Max Planck Institute for Psycholinguistics in Nijmegen. We tested 204 participants (age range 19–39 years) to arrive at our final sample size of 200 (details follow below in the section on Sampling plan). All participants were Dutch native speakers, right-handed, with normal or corrected-to-normal vision and no history of language impairment. After receiving information about the experimental procedures, participants gave informed written consent to take part in the experiment. Participants were paid for their participation (18€).

Participant data were excluded from statistical analyses based on pre-registered criteria (https://osf.io/wh5ke) about the number of artifact-free trials (fewer than 40 out of 60 trials for either the unexpected or expected trials, or fewer than 45 out of 60 trials on average) and the accuracy with which they answered the comprehension questions (less than 80% correct). We excluded and replaced 4 participants.

Materials and design

We prepared two sets of mini-stories: one critical set to measure the pre-nominal prediction effect, and one filler set to manipulate the proportion of (un)expected items. The final set of materials contained 240 sentences in total (120 critical sentences and 120 fillers), which were selected from a larger set of 284 items (160 items from Fleur et al., 2020, and 124 newly created items; because these new items were created in the same way as in Fleur et al., we here describe the creation of these item sets as a single procedure).

The two-sentence mini-stories were created such that they presumably led people to expect a specific, singular definite noun phrase. To establish whether the newly created stories were indeed sufficiently constraining towards these noun phrases, we administered a cloze test in the form of an online questionnaire. All mini-stories were truncated before the target article and appeared in a different randomized order for each participant. We recruited 20 participants (age range: 19–34) from the participant database of the Max Planck Institute for Psycholinguistics in Nijmegen who received financial compensation for filling out the questionnaire (6€). Participants were first given two examples of a mini-story that matched the structure of the test items, but together with a possible ending. They were instructed to complete the incomplete mini-stories with whatever word or words first came to mind. In addition, they were told to avoid repeating words over multiple stories and to not think too long about their answer.

From the obtained responses, we counted how often the expected article or noun was used. For the noun cloze specifically, we also counted responses towards the target noun when the response had both semantic *and* lexical overlap with the target noun (e.g., *trouwpak* 'wedding suit' for *pak* 'suit' or *postbezorger* for *postbode* 'mailman'), when the response was a misspelling or differently-spelled version of the target noun, and when the response was a plural or diminutive noun phrase involving the target base noun (for the selected items there were maximally five out of twenty responses that contained a diminutive or plural form). We calculated cloze probability as the percentage of responses containing the target article or target noun. We then selected 240 items for which both the article and noun cloze were above 75%, and for which an unexpected article or unexpected noun was never higher than 15%.

The average cloze value was 98.6% for expected articles ($SD = 2.24$, range = 95–100) and 96.0% for the expected nouns ($SD = 5.65$, range = 75–100). For the fillers these values were 88.7% ($SD = 6.03$, range = 75–95), and 90.1% ($SD = 7.43$, range = 75–100), respectively. Gender was fully balanced for the critical stimuli (this was not the case in Fleur et al., 2020) but not for the filler sentences, such that in the full item set there were 128 target nouns of common gender ('de'-words) and 112 of neuter gender ('het'-words). On average, the target article was the 8th word in the second sentence ($SD = 2$, range = 4–14), and was immediately followed by the target noun.

As in Fleur et al. (2020), we created the unexpected condition by replacing the target article-noun combination with an unexpected, different-gender article-noun combination. We selected unexpected nouns that we considered relevant and somewhat plausible or at least non-anomalous given the story context. For both the critical sentences and the fillers the unexpected article and noun had a cloze probability of less than 2% on average, and never more than 15% (article: $M = 2\%$, $SD = 4.09$, range = 0–15; noun: $M = 0\%$, $SD = 1.87$, range = 0–15). In both the critical sentences and filler sentences, the expected and unexpected noun was never the sentence-final word and was followed by between 2 and 5 additional words. As in Fleur et al. (2020), we added these sentence-final words to avoid the situation wherein people generate expectations about sentence-endings and pay additional attention to those endings. For a given item, these additional words were identical for the expected and unexpected condition.

We created two critical stimulus lists for which half of the critical sentences contained an unexpected noun phrase and the other half contained an expected noun phrase. None of the sentence contexts was repeated within the same list. These lists were combined with two different lists of filler items, containing either only filler stories with unexpected noun phrases or with expected noun phrases, such that the percentage of disconfirmed predictions was either 75% or 25% (60 critical items plus 120 fillers, out of 240 items in total).

The lists were randomized in such a way that the critical sentences were roughly equally spread across the experiment (i.e., a maximum of 4 subsequent trials with the same article-form and a maximum of 8 subsequent trials from the same condition). Different trial order lists were used such that items appeared in different positions in different lists.

To encourage participants to pay attention to the meaning of the stories, they were asked to answer a yes/no comprehension question on 72 trials (i.e., on 30% of all trials). These questions were spread across the entire experiment, and were separated from each other by maximally eight trials.

Procedure

Participants were seated before a monitor in a soundproof, electrically shielded room. Participants could start each trial by pressing a key on the keyboard. Each trial started with a fixation cross displayed at the centre of the screen, followed by the first sentence of a story shown in its entirety. Participants could press a key to start the second sentence, which was presented one word at a time at the centre of the screen. All words were presented for 300 ms followed by a 300 ms blank screen. If the story was followed by a comprehension question, participants were required to respond yes or no before the next trial started.

To familiarize the participants with the procedure, a brief practice session with five trials preceded the actual experiment, which were selected from sentence pairs that had not been selected as experimental items and that corresponded to the high or low prediction error context in the actual experiment. The experiment was divided in six blocks of forty items each, each block lasting approximately eight minutes, with brief breaks in between. In total, the experiment lasted about one hour, excluding the time needed to prepare the EEG cap.

EEG data recording and pre-processing

We recorded a continuous EEG signal from 27 active scalp electrodes mounted in an elastic cap (ActiCap), placed according to the 10–20

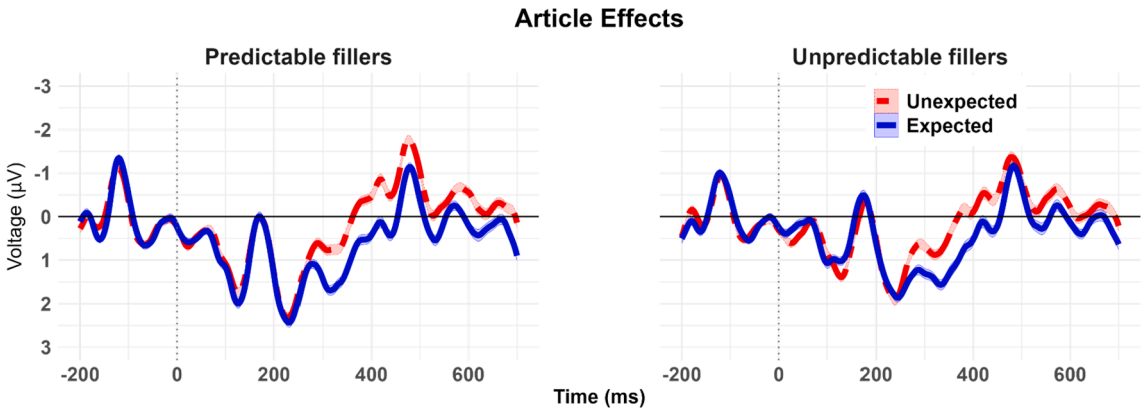

Fig. 1. Article effects when the fillers were predictable or unpredictable. The graph shows the grand-average ERPs in the pre-registered parietal-occipital ROI, elicited by unexpected articles (dashed red lines) and expected articles (solid blue lines) in a context with 75% (left) or 25% (right) predictable story-endings. Shaded areas show the within-subject standard error of the condition mean (Cousineau, 2005; Morey, 2008; calculated with the 'Rmisc' package in R). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

convention and each referenced online to the left mastoid. An additional reference electrode was placed at the right mastoid. Furthermore, we recorded voltage at 4 EOG electrodes (above and under the left eye for the vertical dimension, next to the left and right eye for the horizontal dimension). The signal was amplified using BrainAmps amplifiers and recorded with Brain Vision Recorder (Brain Products, München) at 500 Hz, with a band-pass filter at 0.016–150 Hz (time constant 10 s).

We performed offline data pre-processing with BrainVision Analyzer according to our pre-registration, which was identical to pre-processing Fleur et al. (2020). First, we visually screened the data for bad channels (due to drifting, spiking, excessive line noise) and interpolated bad channels through spline interpolation. We then filtered the continuous data with a 0.1–100 Hz (24 dB/octave roll-off) band-pass filter, re-referenced all channels to the average of the left and right mastoid. We then epoched the data into segments from −500 to 1000 ms relative to target article or noun onset, removed artifact-containing segments (i.e., containing large movement-related artifacts, large bursts of muscle activity, or amplifier blocking) after visual inspection, and performed an ICA-based correction for blinks, eye movements, and steady muscle activity. After this, we applied a 30 Hz low-pass filter (24 dB), and a baseline correction by subtracting the average value within the −200 to 0 ms time window for each trial and channel. Finally, we automatically rejected segments with values that exceeded ± 75 μV at any channel. On average, we retained 58 expected and unexpected article/noun segments for each participant in the predictable/unpredictable fillers condition.

Spatiotemporal regions-of-interests

We used a spatiotemporal region-of-interest (ROI) approach, wherein our main dependent measure (N400 amplitude) for the article analysis was the average voltage across six parietal-occipital channels (P3, Pz, P4, O1, Oz, O2) in the 300–500 ms time window after article onset. This ROI was based on the pre-nominal prediction effect for expectedly definite articles observed by Fleur et al. (2020). For the current noun analysis, we used an N400 ROI that measured average voltage in the 300–500 ms time window at central-parietal channels (Cz, CP1, CP2, P3, Pz, P4), and an additional ROI measuring activity in the 500–700 ms time window at frontal-central channels (F3, Fz, F4, FCz, FC1, FC2), identical to Fleur et al. (2020).

Sampling plan

We performed a power analysis simulation with the SIMR package (Green & MacLeod, 2016) to estimate the required sample size. This power analysis was inspired by peer review comments on a previous draft and was performed when half of the data had already been collected.[6] The analysis took as a starting point the gender-mismatch effect observed in the first 60 subjects in the unpredictable fillers condition (unexpected minus expected, −0.53 μV). We then computed the required sample to achieve 80% power to detect an interaction effect if the effect in the predictable fillers condition was twice that of the unpredictable fillers condition (e.g., Brothers et al., 2019), in a between-subject manipulation. Our analysis demonstrated a total sample of 200 participants was sufficient.

Bayesian mixed-effects model analysis

We performed Bayesian linear mixed-effects analyses in R (R Core Team, 2018) with the 'brms' package (Bürkner, 2017, 2018). Our predictor of interest was the two-level factor 'expectedness' (expected vs. unexpected) and its interaction with the two-level factor 'fillers' (predictable vs. unpredictable), both coded with a deviation contrast. For the article analysis, we included an additional factor 'article-form' (de/het) and associated interactions, to account for potential effects associated with the specific articles, which was important given the lexical differences between 'de' and 'het' ('de' is more frequent than 'het', and may thus elicit smaller N400s overall; Kutas & Federmeier, 2011). The models used a maximal random effects structure as recommended by Barr, Levy, Scheepers and Tily (2013), including random intercepts for participants and items, by-participant random slopes for 'expectedness', 'article-form' and their interaction, a by-item random slope for 'expectedness' and 'fillers' and their interaction (no by-item random slope is included for article-form because for a given item the two levels of 'expectedness' are inseparable from 'article-form'), and all associated random correlations. For the main analyses, we only used brms' default priors. In addition, we performed analyses with informative priors to compute Bayes Factors, including sensitivity analyses. All models were fit using four chains, with 10,000 iterations each (2000 warm-up).

For the analyses and plots, we used Rmarkdown (Xie, Dervieux & Riederer, 2020) and the following packages for R (R Core Team, 2021): "brms" (Bürkner, 2017, 2018), "cowplot" (Wilke, 2020), "dplyr" (Wickham, François, Henry & Müller, 2021), "emmeans" (Lenth, 2021), "ggplot2" (Wickham, 2016), "patchwork" (Pedersen, 2020), "Rmisc" (Hope, 2013), "tidyverse" (Wickham et al., 2019).

[6] Our initial sampling plan and pre-registration from before we collected any data can be found in a previous version of this manuscript on OSF.

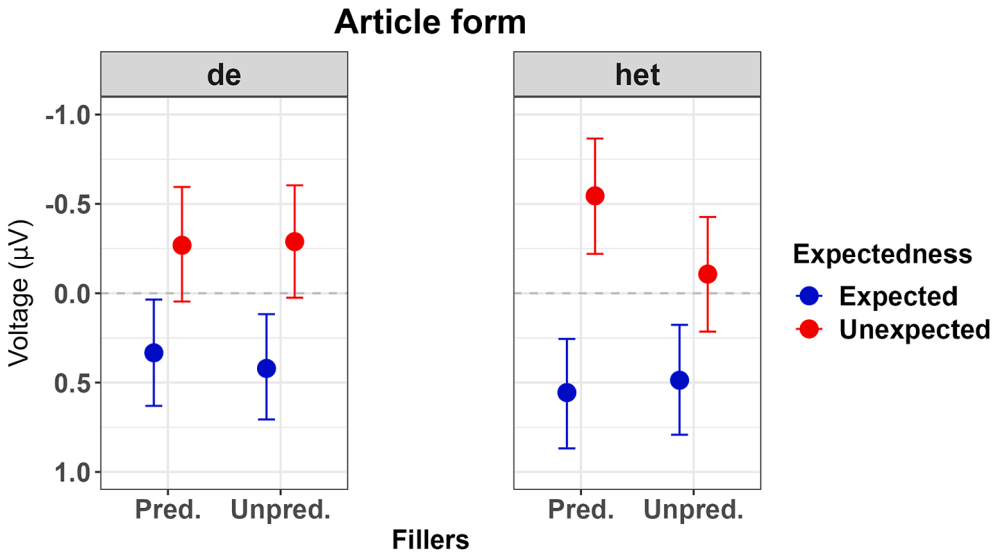

Fig. 2. Effects of expected and unexpected 'de' and 'het' articles when the fillers were predictable or unpredictable. Dots depict the mean voltage estimate at the 300–500 ms ROI and vertical bars indicate the highest density interval of the posterior distribution.

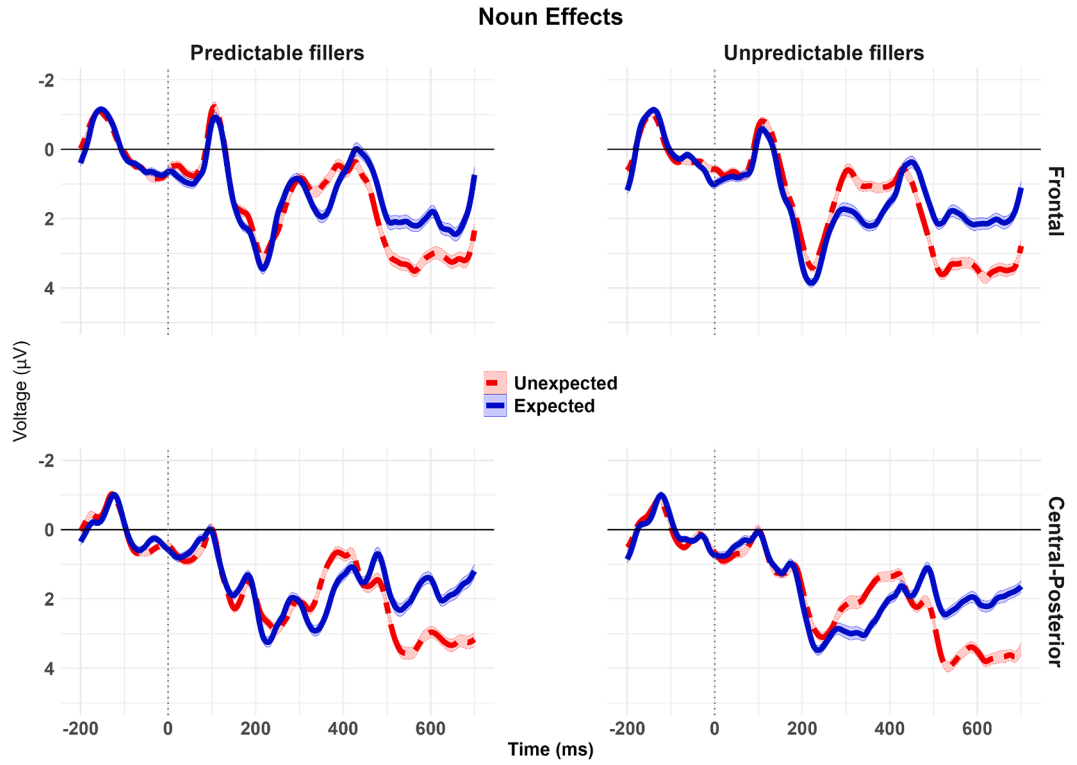

Fig. 3. Noun effects when fillers were predictable or unpredictable. The graph shows the grand-average ERPs in the pre-registered frontal and central-posterior ROIs (top and bottom panels, respectively), elicited by unexpected nouns (dashed red lines) and expected nouns (solid blue lines) in a context with 75% (left) or 25% (right) predictable story-endings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Results

Articles

Visual inspection of the ERP signatures showed that unexpected articles elicited more negative voltage than expected articles both when fillers were predictable and unpredictable (Fig. 1; see also Appendix Figs. A1 and A2 for the ERP waveforms at all recorded channels and Fig. A5 for the associated scalp distribution plots). Analysis in the 300–500 time window at our spatial ROI confirmed this effect of expectedness in the entire dataset (b = −0.75 μV, SE = 0.10, 95 % CI [−0.94, −0.56]), as well as separately with predictable fillers (b = −0.85 μV, [−1.10, −0.59]) and unpredictable fillers (b = −0.65 μV, [−0.92, −0.40]). Although the expectancy effect was numerically larger with predictable fillers than with unpredictable fillers, the analysis did not reveal sufficient evidence to support this interaction (b = 0.20 μV, SE = 0.18, [−0.15, 0.55]). Interestingly, however, this interactive pattern was almost entirely caused by responses to unexpected articles. Unexpected articles elicited more positive voltage with predictable fillers than with unpredictable fillers (b = 0.21 μV, [−0.14, 0.55]),

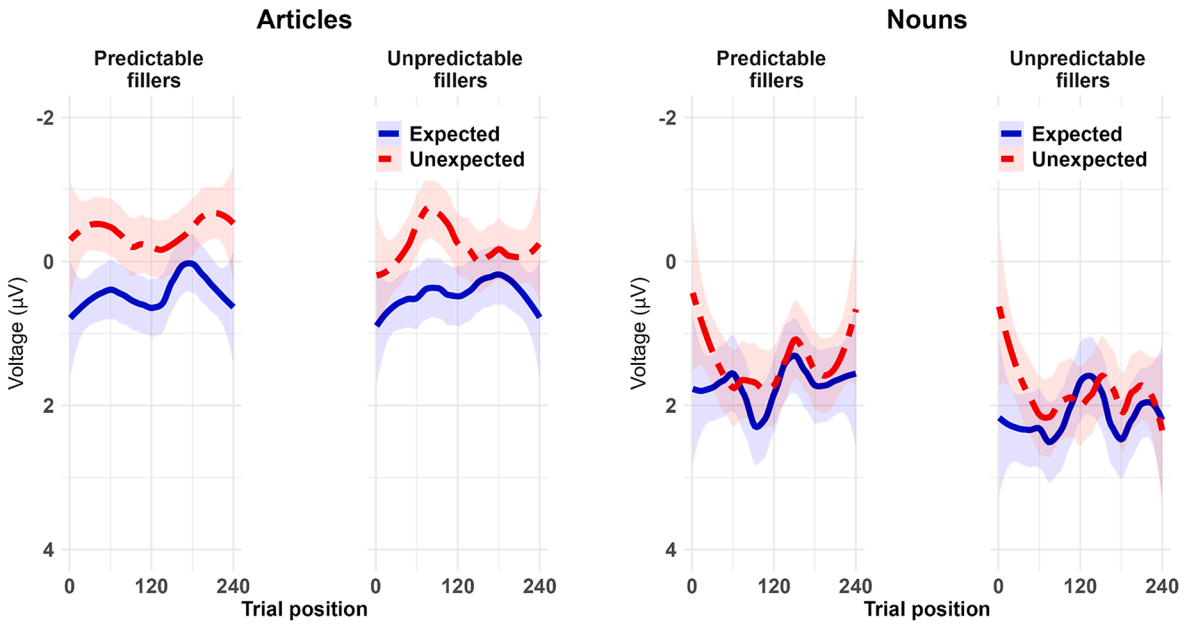

Fig. 4. Expectedness effects over time. Article-elicited (left panels) and noun-elicited (right panels) ERP effects of expectedness in the N400 ROI (300–500 ms time window) as a function of trial position when fillers were predictable or unpredictable. Depicted lines are the local regression ('loess') curves for the raw N400 responses. Shaded areas show the 95% confidence intervals.

whereas for expected articles this effect of fillers was close to zero (b = 0.01 μV, [−0.31, 0.32]). This pattern is inconsistent with the rational adaptation hypothesis, which stipulates a change in processing of only predictable words (e.g., Delaney-Busch et al., 2019; Brothers et al., 2017, 2019; Nieuwland, 2021b).

In the 500–700 ms time window, we also observed effects of expectedness, although these were generally smaller than in the 300–500 ms time window (with predictable fillers, b = −0.52 μV, [−0.79, −0.24]; with unpredictable fillers, b = −0.37 μV, [−0.65, −0.10]).

We also analyzed the three-way interaction between expectedness, fillers and article form. Although neither time window yielded strong support for this interaction (300–500 ms time window, b = 0.61 μV, SE = 0.36, [−0.10, 1.32]; 500–700 ms time window, b = 0.70 μV, SE = 0.42, [−0.13, 1.54]), the observed patterns in both time windows were consistent with those reported by Fleur et al. (2019). For brevity, we here report follow-up comparisons only for the 300–500 ms time window. Fig. 2 depicts the mean voltage estimate and highest density interval of the posterior distributions per condition. Both 'de' and 'het' articles elicited expectedness effects (unexpected minus expected yielding a negative voltage difference; 'de': b = −0.65 μV, [−0.95, −0.37]; 'het': (b = −0.85 μV, [−1.14, −0.56]) with predictable and unpredictable fillers, but for 'het' this effect was almost twice as large with predictable fillers (b = −1.10 μV, [−1.49, −0.72]) than with unpredictable fillers (b = −0.60 μV, [−0.98, −0.21]), whereas for 'de' the effect was slightly smaller with predictable fillers (b = −0.60 μV, [−0.99, −0.22]) than with unpredictable fillers (b = −0.70 μV, [−1.09, −0.33]).

Fig. 2 also suggests that whether fillers were predictable or unpredictable mostly impacted the responses to unexpected 'het' articles. This pattern, too, is not in line with the theoretical claim that rational adaptation only impacts effects on predictable words (e.g., Delaney-Busch et al., 2019; Brothers et al., 2017; 2019; Nieuwland, 2021b).

Nouns

As shown in Fig. 3, unexpected nouns with predictable and unpredictable fillers elicited enhanced N400 activity compared to expected nouns (see also Figs. A3 and A4 for ERPs at all channels and Fig. A5 for scalp distribution plots). As in Fleur et al. (2020), this N400 effect was somewhat short-lived for a typical N400 effect, visibly most prominent in the 300–400 ms time window, and immediately followed by enhanced post-N400 positivity (PNP) that already started in the 300–500 ms time window and extended into later time windows (visible up to and including 1000 ms after noun onset). Statistical analysis confirmed that unexpected nouns elicited enhanced negativity compared to expected nouns in the N400 ROI, b = −0.32, SE = 0.16, 95 % HDI [−0.63, 0.00], although this effect was small. There was no evidence to suggest this effect depended on the fillers (interaction term, b = −0.01 μV, SE = 0.24, [−0.47, 0.46] and associated estimates were highly similar for predictable fillers (b = −0.32 μV, [−0.70, 0.06]) and unpredictable fillers (b = −0.31 μV, [−0.72, 0.07]).

In contrast, the expectedness effect in the PNP ROI (frontal ROI in the 500–700 ms time window after noun onset) was large and strong, with unexpected nouns eliciting more positive voltage than expected nouns (b = 1.29 μV, SE = 0.18, [0.94, 1.63]). There was little evidence to support an interaction with the fillers (interaction term, b = 0.35 μV, SE = 0.29, [−0.22, 0.92]), although the estimate was slightly larger with unpredictable fillers, b = 1.46 μV, [1.02, 1.91], than with predictable fillers, b = 1.11 μV, [0.66, 1.57]. Of note, the obtained PNP effect was, numerically at least, even stronger at the central-posterior ROI (b = 1.51 μV, SE = 0.17, [1.18, 1.83]), which makes our results similar to those of Fleur et al. (2020) and possibly distinct from reports of a frontal PNP (e.g., Van Petten & Luka, 2012).

Bayes factor analyses

We computed Bayes Factors (BF) to quantify the evidential strength for/against an interaction between the effects of article expectedness and filler predictability. We ran a total of 6 Bayesian mixed-effects models with the same random effects structure as our main analysis. Each model used the same prior for the effect of expectancy (unexpected minus expected, mean = −0.78 μV, SD = 0.16) based on Fleur et al. (2020), in addition to the default priors as provided by the brms package for all other model estimates. Crucially, the models used a different prior

for the critical interaction, with a mean prior corresponding to either a 0 or 50% reduction of the expectancy effect with predictable fillers compared to unpredictable fillers (i.e., a mean effect of 0 or 0.39 μV,[7] respectively) and a standard deviation of 0.20 (corresponding to roughly half that of the mean prior, as recommended by Dienes, 2014, because it renders the plausibility of a negative voltage difference to be negligible), or a double or quadruple of the standard deviation (i.e., 0.40 or 0.80) corresponding to weaker prior beliefs. We calculated BFs with the Savage–Dickey method (e.g., Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), namely as the ratio between the posterior and prior distribution at an effect size of 0 μV. Following previous research (Nieuwland et al., 2020a), we interpreted the strength of the obtained evidence according to the convention of Jeffreys (1961).

All six analyses favored the null hypothesis, meaning that the posterior probability at the effect size of 0 was greater than the prior probability. However, the values ranged did not exceed 3 so only constituted anecdotal support (prior mean 0.39 μV: BF 1.14, 1.44 and 2.55 for prior SD 0.20, 0.40 and 0.80 respectively; prior mean 0 μV: BF 1.08, 1.48 and 2.65 for prior SD 0.20, 0.40 and 0.80 respectively).

Did prediction effects change during the experiment?

We investigated how the pre-nominal prediction effect in the 300–500 ms time window changed during the experiments (Fig. 4), using a Bayesian mixed-effects model analysis which added the (z-transformed) continuous predictor 'trial position' and associated interactions to the previously reported analysis.[8] This analysis did not yield evidence for a three-way interaction between fillers, expectedness and trial position (articles, b = 0.06 μV, SE = 0.18, [−0.30, 0.42]; nouns, b = 0.08 μV, SE = 0.22, [−0.35, 0.52]), nor for a two-way interaction between expectedness and trial position (articles, b = 0.10 μV, SE = 0.09, [−0.07, 0.29]; nouns, b = 0.18 μV, SE = 0.12, [−0.05, 0.40]).

Discussion

We investigated whether lexical prediction during discourse comprehension is impacted by the overall rate of prediction success. As a dependent variable, we employed the pre-nominal prediction ERP effect on gender-marked articles (e.g., Fleur et al., 2020). According to the rational adaptation hypothesis of prediction (e.g., Kuperberg & Jaeger, 2016; Yan et al., 2017), this effect should be stronger when the ratio of predictable and unpredictable stimuli is high compared to low, specifically due to smaller N400 responses to predictable stimuli (corresponding to strengthening predictions, see Brothers et al., 2017, 2019; Lau et al., 2013).

In our large-sample ERP study, we observed a strong pre-nominal prediction effect (unexpected minus expected, −0.75 μV), replicating the pattern found by Fleur et al. (2020). Crucially, we failed to find evidence for the rational adaptation hypothesis of prediction. Our results did not support a modulation of the pre-nominal prediction effect by the predictable/unpredictable fillers, and Bayes Factor evidence anecdotally favored the null hypothesis. Notably, the prediction effect was indeed numerically larger when fillers were predictable (−0.85 μV) as compared to unpredictable (−0.65 μV), a pattern that on its own is consistent with rational adaptation. Crucially, however, the underlying pattern causing this difference was inconsistent with rational adaptation: if predictability of the fillers had any effect to begin with, it was primarily on responses to *unexpected*, neuter gender 'het' articles. This is inconsistent with the rational adaptation hypothesis, which predicts modulation of N400 responses to *expected* 'de' and 'het' articles.

Subsequent nouns also replicated the patterns reported by Fleur et al., namely eliciting an N400 effect of expectancy that was smaller in size than for the articles (−0.32 μV), followed by a much stronger PNP effect (1.29 μV). Neither effect showed evidence of rational adaptation.

We conclude that our participants either did not rationally adapt their predictions to the (estimated) probability of prediction success, they estimated prediction success inaccurately, or they did not estimate it at all. In the below, we summarize our findings in further detail, we consider critiques and potential caveats of our study, and then we discuss potential theoretical implications.

The pre-nominal prediction effect and rational adaptation

Our study was the first to investigate rational adaptation of prediction using the pre-nominal prediction effect. Articles with unexpected gender elicited enhanced negativity at a pre-registered ROI compared to articles with expected gender, and this effect was rather strong (cf. Nicenboim et al., 2020) and replicated the pattern observed by Fleur et al. (2020). Based on the results of Fleur et al. (2020), we take the pre-nominal prediction effect to primarily reflect processing of a mismatch between an expected and an encountered article.[9]

Most importantly, this effect overall proved resistant to influences from the filler materials, that is, potential influences associated with the frequency of prediction success. Although the impact of the fillers on the article-expectancy manipulation was not zero, the estimate for the interaction term was only one standard error away from zero, and our Bayes Factor analyses (anecdotally but consistently) supported the null hypothesis. However, the inconsistency of our results with rational adaptation of prediction was most apparent when we examined the effects of the fillers for common and neuter gender articles separately.

Namely, the fillers had little impact on the expectancy effect elicited by the common gender article 'de', and whatever effect it did have was in the opposite direction of what the rational adaptation hypothesis predicts. For the neuter gender article 'het', the expectancy effect was indeed numerically greater with predictable fillers than with unpredictable fillers, as the rational adaptation hypothesis predicts. However, even this result came from an underlying pattern that was inconsistent with the rational adaptation hypothesis: the impact of fillers was almost zero for expected 'het' and only visible for unexpected 'het'.

[7] This value is only slightly smaller than the corresponding 50% reduction of the currently observed effect with predictable fillers, which would be 0.425 μV instead of 0.39 μV, and which would increase the BF support for the null hypothesis when used as prior. However, we decided against basing a prior on the dataset to be analyzed.

[8] This resulted in the following model for the articles: Voltage ~ Fillers*Expectedness*TrialPosition*ArticleForm + (1 + Expectedness*TrialPosition*ArticleForm | subject) + (1 + Fillers*Expectedness*TrialPosition | item).

[9] Fleur et al. found distinct ERP effects associated with mismatching article form and gender. Definite, gender-mismatching articles elicited enhanced parietal-occipital negativity in the 300–400 ms time only when they matched the expected definiteness (e.g., 'het' when people expected 'de', but not when people expected the indefinite 'een'). This suggest people do not merely use article-gender to revise their noun prediction (e.g., Rabovsky, 2020; Szewczyk, & Wodniecka, 2020; Van Berkum et al., 2005), but suggests a sensitivity to the mismatch between predicted and encountered article form (which in Dutch is determined by gender and definiteness). However, enhanced parietal negativity in the 500–700 ms time window was observed for gender-mismatching articles regardless of expected definiteness. This pattern was inconsistent with article form prediction, consistent with a process of noun prediction revision. The latter was also supported by exploratory analyses. ERP responses to expectedly definite, gender-mismatching articles correlated with next-word entropy, that is, the extent to which unexpected articles suggested one specific alternative noun. For example, 'de' may rule out the expected 'boek' (book), but suggest a plausible alternative like 'roman' (novel). Furthermore, N400 responses to nouns that were relatively unexpected before the article (e.g., 'roman' in the previous example) correlated with predictability of the noun after the mismatching article (while controlling for other relevant factors, e.g., plausibility, semantic relatedness; see Nieuwland et al., 2020b).

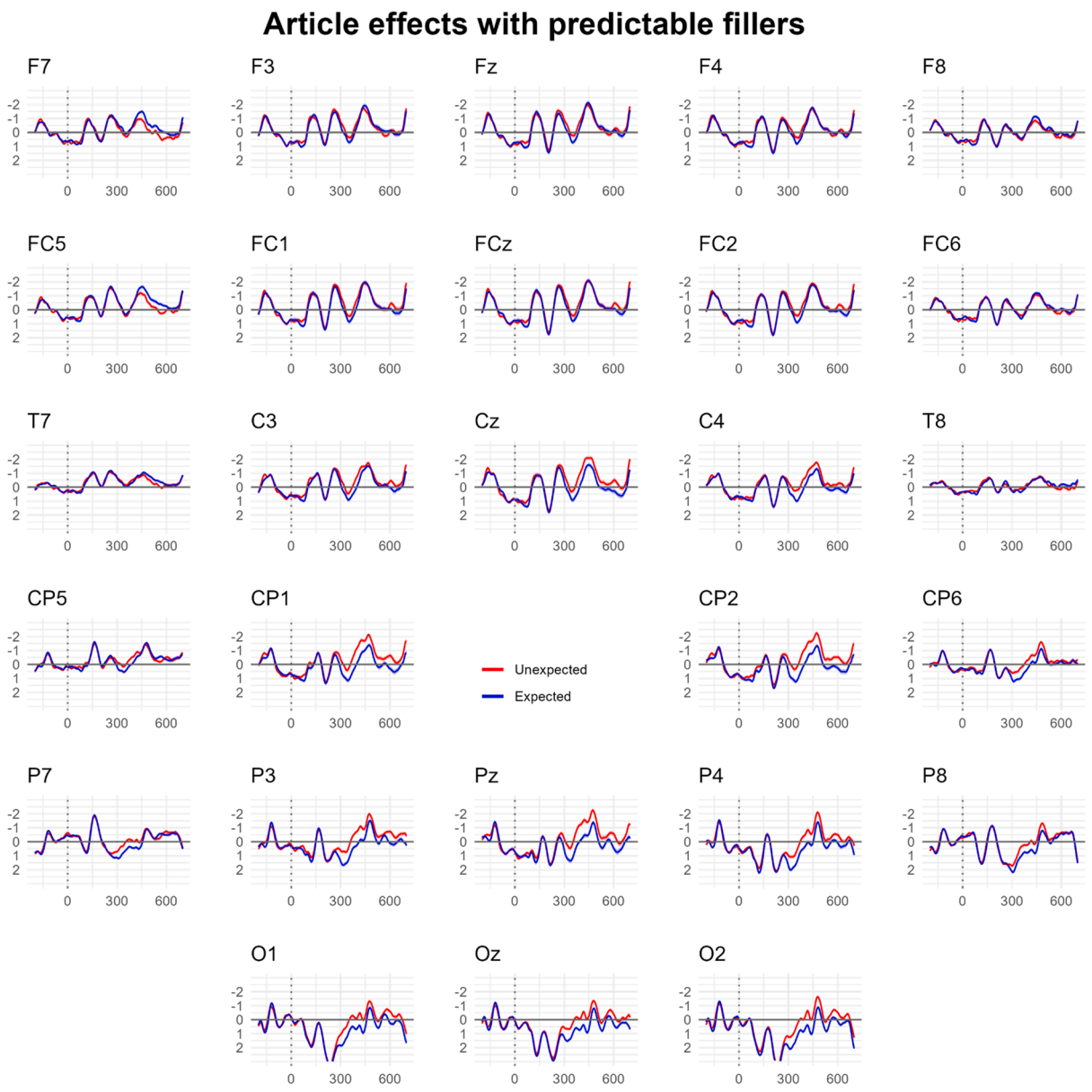

Fig. A1. Article effects with predictable fillers. The graphs show the grand-average ERPs of expected and unexpected articles for each recorded channel.

Before we proceed, we acknowledge that our sample size was not tailored to reliably detect three-way interaction effects, and the resulting estimates lacked the required precision to support strong claims. Therefore, if one applies the criterion of whether the credible interval includes zero as a measure of 'statistical significance', the three-way interaction between fillers, expectedness and gender is not statistically significant, but neither is the two-way interaction between fillers and expectedness. Nevertheless, we consider the observed patterns of sufficient importance to attempt explanation, also because predictive processing differences between 'de' and 'het' have been reported previously (e.g., Brouwer, Sprenger & Unsworth, 2017; Fleur et al., 2020; Loerts, Wieling & Schmid, 2013).

Although the exact nature of this interactive pattern remains to be established, an explanation could be sought in the differences between 'de' and 'het' in their lexical frequency and acquisition (see also Nieuwland et al., 2020a, for related discussion concerning gender-marking on adjective-suffix inflection). Corpus counts show that 'de' is about 2 or 3 times more frequent than 'het', depending on how they are counted (e.g., Van Berkum, 1996). Possibly related to this difference, and to the different alternate uses of 'de' (for plural nouns) and 'het' (e.g., for diminutives), developmental studies show that children and L2 speakers learn to use 'de' before 'het' and are more likely to overgeneralize 'de' than 'het' (e.g., Wijnen & Verrips, 1998), suggesting that 'de' is typical while 'het' is atypical (or 'marked').

We first need to explain why the pre-nominal prediction effect might be smaller for 'de' than for 'het' to begin with. Although perhaps a far-fetched hypothesis, comprehenders may have reduced sensitivity to incorrect use of 'de' due to exposure to overgeneralizations (by L2 speakers, children or by themselves in childhood). Alternatively, unexpected 'het' is particularly disruptive because it is a relatively low-frequent article where participants expected a relatively high-frequent article. Another possibility is that participants considered unexpected 'de' as heralding a plural version of the expected noun (i.e., they predicted 'het boek/the book' but took 'de' as referring to multiple books) and found this easier to accommodate than a diminutive interpretation for the unexpected 'het'. Of note, this would have to hold despite the absence of diminutive or plural nouns in that story-position in the experiment (for related discussion, see Rabovsky, 2020).

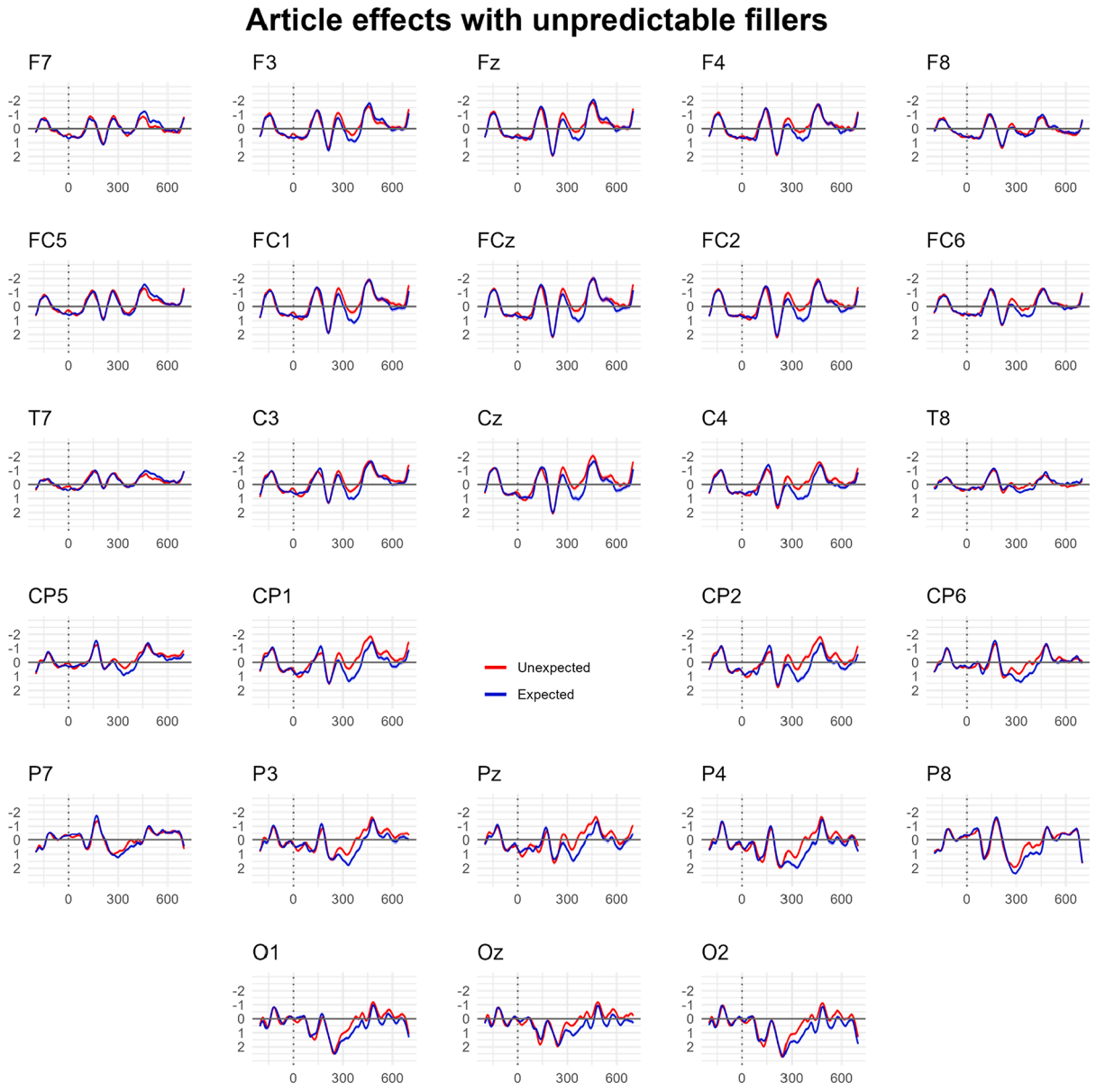

Fig. A2. Article effects with unpredictable fillers. The graphs show the grand-average ERPs of expected and unexpected articles for each recorded channel.

For each of these alternatives, the question then becomes why the observed effects may depend on the fillers. Although the ratio of 'de/het' articles was not perfectly balanced (128/112) in the current study, the target articles were certainly more balanced than in their natural occurrence. Thus, one potential explanation is that with unpredictable fillers, participants became relatively less sensitive to unexpected 'het', because repetition impacts a marked article more than an unmarked one. It is also possible that participants became better at accommodating the diminutive interpretation for unexpected 'het'. Unfortunately, we do not have conclusive evidence for or against these alternatives, and we therefore refrain from conclusions regarding the differences between 'de' and 'het'.

Noun-elicited ERPs

Although nouns elicited a reliable expectancy N400 effect, this effect was smaller than is typically reported (e.g., DeLong et al., 2005; Ito et al., 2016). Visual inspection suggested that before the end of the ROI time window, enhanced N400s for unexpected nouns were cut short by an enhanced positive going ERP, which counteracted the average negativity in the ROI. This pattern is reminiscent of Fleur et al. (2020; Experiment 2), who observed a small N400 effect that did not reach the standard criterion of statistical significance despite a large sample of 80 participants. As in Fleur et al. (2020), the enhanced positive ERP continued until at least the end of the epoch at 1000 ms. We consider this ERP effect an example of the post-N400 positivity observed for unexpected but at least somewhat plausible nouns in highly constraining sentences (taken to indicate processing costs of disconfirmed predictions, possibly the inhibition of a strongly anticipated word, e.g., Van Petten & Luka, 2012), even though its scalp distribution was not as anterior as is typically observed.

Neither the noun-elicited N400s nor the subsequent positive-going ERPs yielded evidence for rational adaptation. For the N400s, the estimate for the interaction between fillers and noun expectancy was close to zero. For the subsequent positive-going ERPs, the estimate was numerically greater with unpredictable fillers than with predictable fillers, a direction that is inconsistent with the rational adaptation hypothesis.

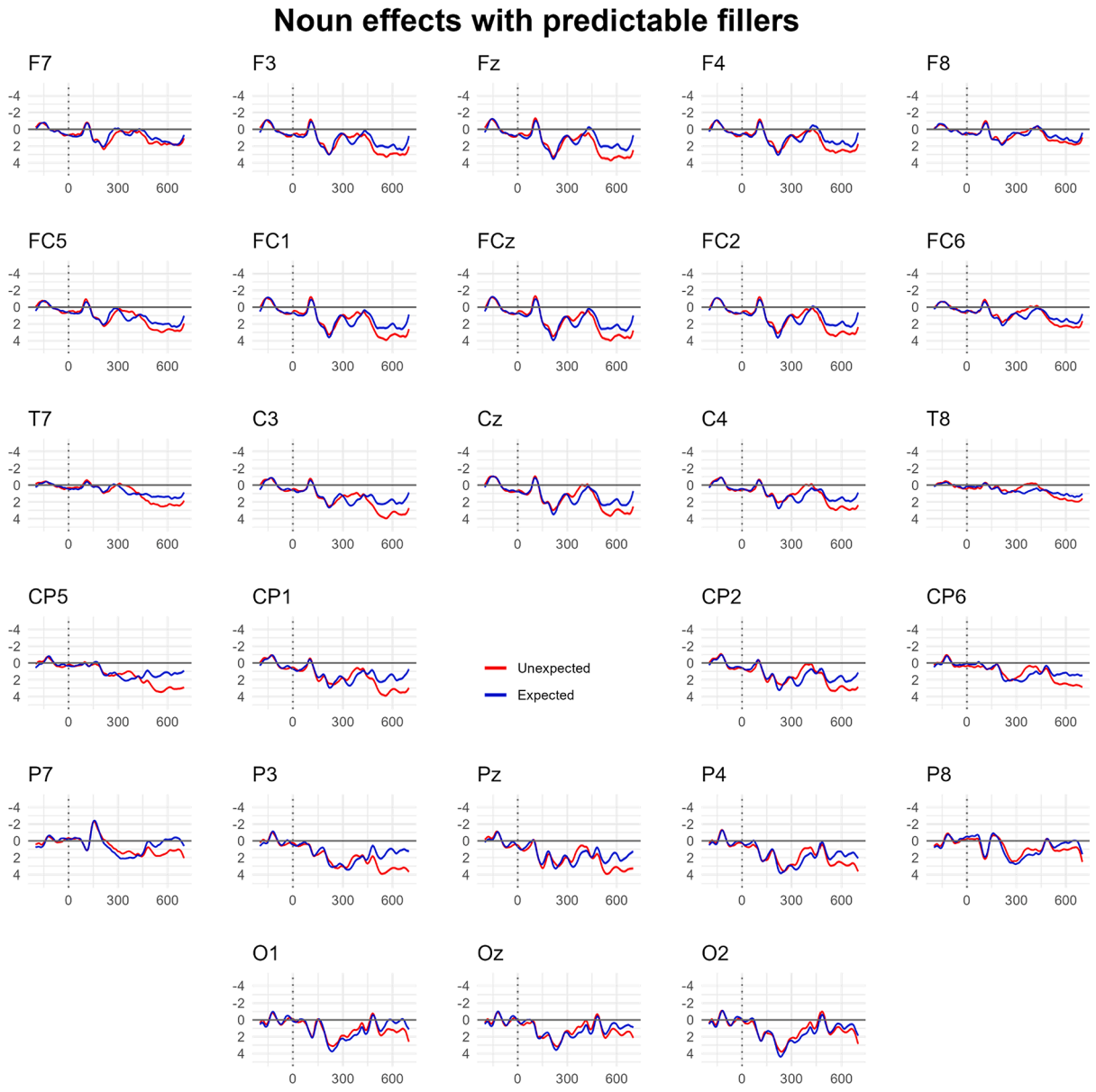

Fig. A3. Noun effects with predictable fillers. The graphs show the grand-average ERPs of expected and unexpected nouns for each recorded channel.

Concerns and caveats

Three concerns were leveled at our conclusions during peer review. The first concern was that we did not find evidence for rational adaptation of predictions because our ratio of predictable/unpredictable stimuli (75/25%) was not as extreme as in other studies (Brothers et al., 2017: 82.5/17.5%; Brothers et al., 2019; Dave, et al., 2021: 80/20%). Indeed, our results could have differed had we used a more extreme ratio. However, we do not consider it plausible that a small ratio change would suddenly yield support for rational adaptation, given that it would not change the underlying patterns.

The second concern was that our noun N400s indicate a 'failed manipulation control' because they yielded a relatively small prediction effect compared to other studies (e.g., Kutas & Federmeier, 2011). While an absent or very small noun effect may be surprising, we do not take it to indicate absence of prediction, chiefly because the article effects already demonstrate prediction on their own. Our primary interest was the article effects and we did not match expected and unexpected nouns on lexical and sentence-level variables that may modulate noun N400 amplitude. In addition, noun N400s may also be modulated by change in sentence constraint rendered by the articles (e.g., Fleur et al., 2020), that is, reduced N400s may be found for previously unexpected nouns that became predictable upon encountering a gender-mismatching article. Moreover, we think that the most important reason for obtaining a small expectancy effect on noun N400s was the positivity which started well within the N400 analysis time window (see also Fleur et al., 2020).

The third concern was that our manipulation was not strong enough to elicit adaptation because our target words were embedded in a two-sentence story context, and were not therefore in the focus of attention. This concern is twofold: our target articles and nouns may not have attracted enough attention to engage adaptation, in contrast to single-sentence final words (e.g., Brothers et al., 2017, 2019), and our contexts may have contained other articles and nouns that themselves either did or did not violate predictions, therefore potentially reducing prediction error on the target articles. This second point can also be raised against studies with single sentences (e.g., Brothers et al., 2017, 2019), although our contexts were longer than those of previous studies and therefore likely contained more article-noun combinations.

A simple response is that this argument predicts the strongest rational adaptation in word-pair studies (e.g., Delaney-Busch et al.,

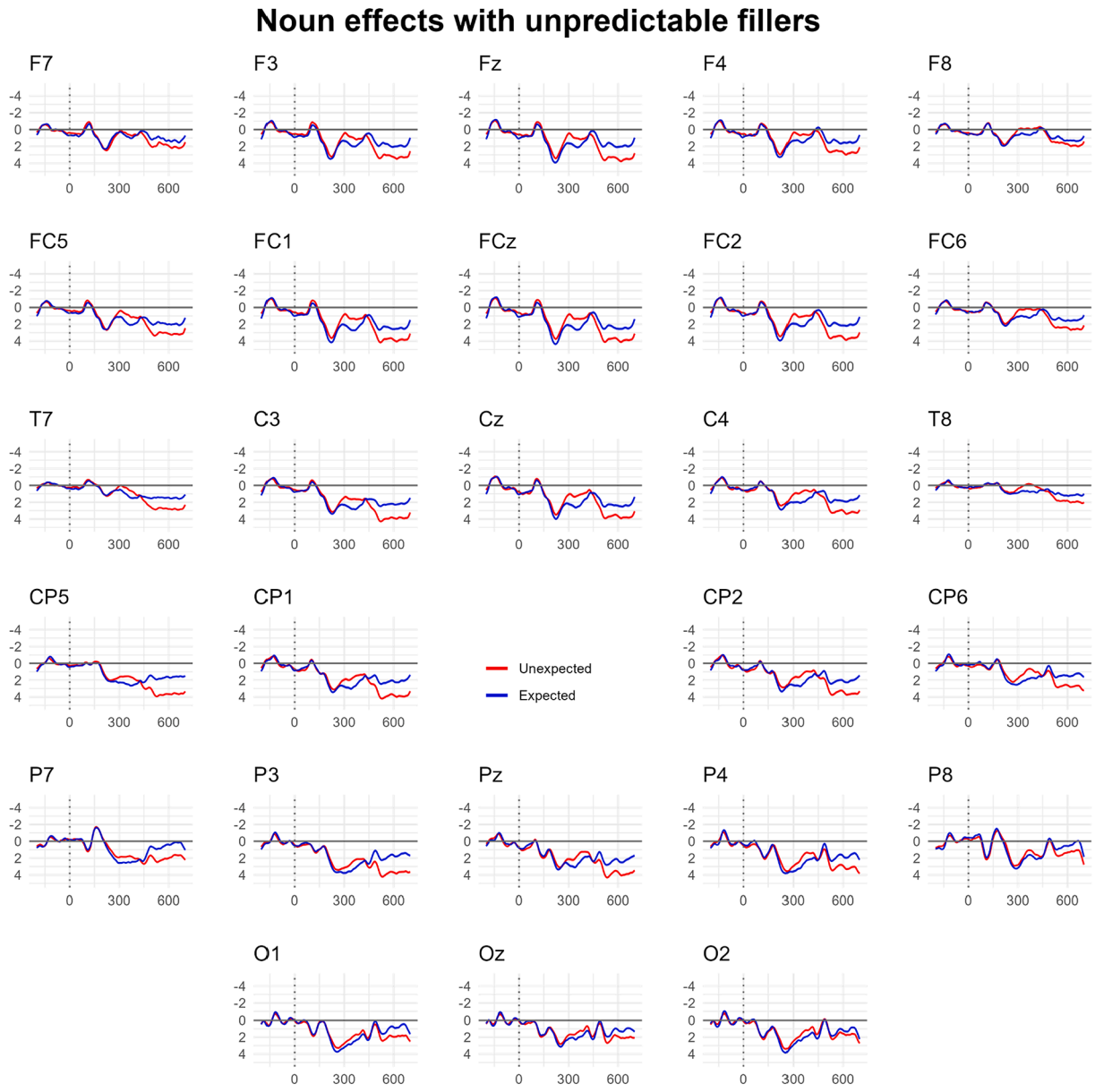

Fig. A4. Noun effects with unpredictable fillers. The graphs show the grand-average ERPs of expected and unexpected nouns for each recorded channel.

2019; Lau et al., 2013; Ness & Meltzer-Asscher, 2021), which does not seem to pan out (Nieuwland, 2021a,b). However, this concern deserves a more in-depth consideration because it brings up a more fundamental theoretical question: If people rationally adapt their predictions, how do they know what to adapt to?

In predictive accounts of language comprehension, listeners predict upcoming information all the time and at all levels of representation to varying degrees (e.g., Altmann and Mirković, 2009; Dell & Chang, 2014; Pickering & Garrod, 2007, 2013). In such accounts, prediction error is highly variable and waxes and wanes over the course of time. Rational adaptation only seems feasible then if comprehenders select which prediction error signals to adapt to. This is perhaps where sentence/story-position becomes relevant, as the sentence-finality might serve as a cue to comprehenders to only adapt to the prediction error associated with that final word. Perhaps, then, people adapt their predictions about sentence-final words only if they pay special attention to these words. While this might work in experiments whose goals are easily inferred by or even explicitly stated to participants (e.g., Brothers et al., 2017; Huettig & Guerra, 2019), this clearly raises the question of how useful (and therefore widespread) rational adaptation might really be. Even if speakers in real world conversations produce many predictable or unpredictable words, they are unlikely to do so for a fixed sentence position or for predominantly sentence-final words. This merely underlines the importance of the current study. Embedding target words in a richer mini-story is a feature, not a bug, because it allows us to investigate the generalizability of rational adaptation.

In this spirit, we should also consider some caveats regarding the generalizability of the pre-nominal prediction effect. We observed this effect even when most materials contained an unpredictable article-noun combination. However, our materials were constructed in a particular way that could limit generalizability. For example, we could not investigate a potential role of predictive cue reliability of article gender. Pre-nominal article gender was always a reliable cue to the expected noun, because expected gender always heralded the expected noun and unexpected gender always heralded an unexpected noun. Pre-nominal prediction effects might weaken once participants learn that article gender is not a reliable cue to the noun (for example, if the articles sometimes herald plural and/or diminutive nouns of either

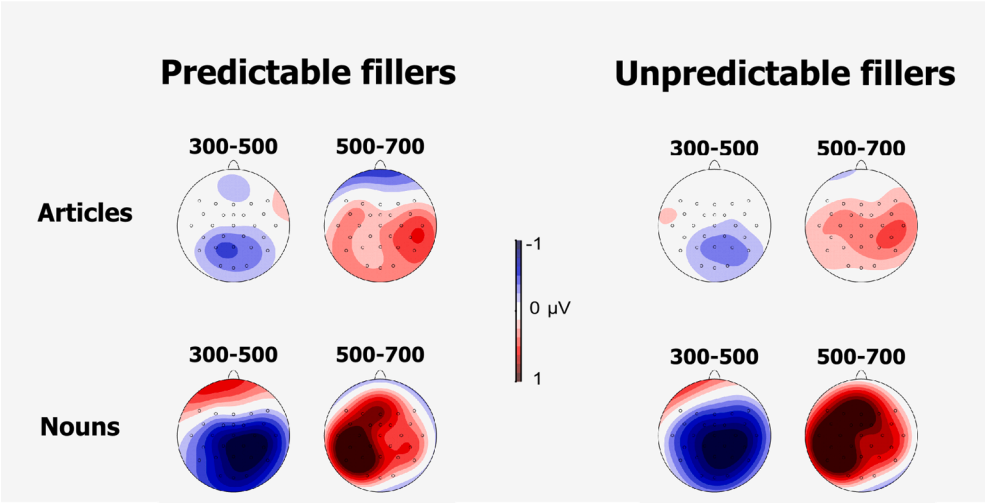

Fig. A5. Scalp distribution of article and noun expectancy effects (unexpected minus expected) when fillers were predictable (left) or unpredictable (right).

gender). Although predictive cue validity might underlie general processing differences between types of articles and possibly between languages (Ito et al., 2020; Nieuwland et al., 2018; Rabovsky, 2020), it is unclear whether cue validity is tracked and used on a trial-by-trial basis.

We also could not investigate a potential role of constraint, because all mini-stories in our experiment were highly constraining. In this regard, our study could still be considered 'prediction-encouraging' (Huettig & Mani, 2016; see also Heyselaar, Peeters & Hagoort, 2021). Maybe people do adapt their predictions, not to the likelihood of error, but to constraint, i.e., to the likelihood that a prediction can be generated in the first place. Participants possibly adapt their comprehension strategies once they learn that the materials frequently make them think of a specific word that either does or does not appear. Were this assumed learning process to be delayed or counteracted (e.g., in an experiment with only non-constraining filler stories), observed prediction effects may change accordingly.

No rational adaptation to prediction error: Theoretical implications

In our study, prediction strength was primarily determined by 'local' linguistic constraint (the discourse context), not by 'global' probabilistic constraints (statistical regularities in the wider experimental context). Our results therefore call into question rational adaptation as a crucial component of predictive processing. In particular, we question that people generally adapt unfolding predictions to their estimated reliability. Without evidence of rational adaptation, we are left to conclude that either our participants did not estimate the likelihood of prediction error to begin with, they estimated it wholly inaccurately, or they estimated it (approximately) correctly without adapting their predictions accordingly. Our data do not distinguish between these interpretations. We think it is reasonable to assume participants realized that the second sentence of each story (or of most stories) contained a relatively expected or unexpected word (anecdotally, remarks from our participants seem to confirm this), and maybe even that unexpected words were more common than expected words, or vice versa. But even if participants picked up on such statistical regularities in the stimuli, they did not adapt their predictions accordingly. In other words, our results show that the language comprehension system does not always engage prediction in an 'optimal' way. This calls for a reconsideration of the assumptions of the rational adaptation hypothesis. If people sometimes do adapt their predictions, how would this work to begin with?

The rational adaptation hypothesis stipulates two separate predictions, one *content*-prediction generated by the language comprehension system and one *meta*-prediction (or 'hyperprior'). Bayesian adaptor models weigh these two predictions by a formula (e.g., Delaney-Busch et al., 2019), but a formula is not a mechanism and does not specify how *meta*-predictions are generated and used to modulate *content*-predictions (e.g., Kaplan, 2011; Kaplan & Craver, 2011). Advocates of this hypothesis even waver in their claims, sometimes stating that "we are not claiming that the brain literally computes probabilities" (Kuperberg & Jaeger, 2016), and at other times claiming evidence that "the brain combines immediate contextual constraints with global probabilistic constraints" (Delaney-Busch et al., 2019). This lack of clarity resonates with the broader discussion in psychology on whether Bayesian models are mere statistical tools to provide teleological explanations of cognition or whether they capture the psychological reality of Bayesian computations (e.g., Bowers & Davis, 2012a,b; Jones & Love, 2011; Marcus & Davis, 2013).

We distinguish three possible explanations based on which processes supposedly adapt, none of which are strongly supported to date. One possibility is that frequent prediction error causes a more general adaptation of comprehension. For example, maybe people start putting greater emphasis on bottom-up stimulus evaluation and engage top-down processing mechanisms less (Brothers et al., 2017; Lupyan & Clark, 2015), or maybe people reduce covert engagement of the language production system to internally generate anticipated lexical items ahead of time (e.g., Pickering & Garrod, 2013). Importantly, if adaptation impacts comprehension, this should in principle be detectable (for example, in increased reading times or N400 amplitudes) *before* the target word. That is, unless people are able to focus their adaptation to specific (e.g., sentence-final) words.

Another possibility is that frequent prediction error causes people to inhibit or change predictions after generating them (Dave et al., 2021). Dave et al. suggested this could take place via non-linguistic, domain-general executive resources such as 'cognitive control' (for related discussion, see Ryskin, Levy & Fedorenko, 2020; Shain, Blank, Van Schijndel, Schuler & Fedorenko, 2020).

A third possibility is that frequent prediction error does not change people's predictions, but only how they deal with prediction confirmation and/or disconfirmation. In this account, one would expect to see adaptation effects somewhat downstream from the target word. In the current study, for example, we did not find effects of the frequency of prediction error, yet responses to expected and unexpected words may change during an experiment (increase of the noun-elicited PNP effect) even without clear evidence that the predictions themselves change (article-elicited effects, noun-elicited N400s). In addition, we think that changes in how people respond to prediction (dis)confirmation are particularly likely when they are engaging in a prediction-relevant task (e.g., through an explicit instruction to predict sentence-final words or through a meta-linguistic judgment task).

The outlined three possibilities, which are not mutually exclusive, differ in the assumed locus of adaptation: in comprehension, in processing (generating or evaluating) predictions, or in dealing with prediction (dis)confirmation. If adaptation impacts comprehension, people may indeed generate weaker predictions; not necessarily because generating predictions is metabolically costly, but possibly because people generate weaker discourse representations that pre-activate the relevant semantic content. If adaptation impacts the processing of predictions and their (dis)confirmation, initial predictions are generated regardless of adaptation, i.e., relatively automatically. If adaptation causes people to inhibit their own predictions (e.g., Dave et al., 2021), this would suggest that to some extent predictions are under strategic control. We emphasize that these possibilities do not require assuming that predictions are metabolically costly to generate nor that they are generated 'actively' (cf. Kuperberg & Jaeger, 2016).

Regardless of the presumed mechanism for rational adaptation, and regardless of the current results, we acknowledge that people - in principle - may be able to rationally adapt their predictions. However, along with other recent demonstrations (Nieuwland, 2021a,b), the results of the current study show that frequent prediction error does not necessarily dampen predictive processing. Therefore, rational adaptation and the notion of expected utility does not suffice as an overarching explanation of predictive processing, at least not in the way outlined by Kuperberg and Jaeger (2016) and others (Delaney-Busch et al., 2019; Ness & Meltzer-Asscher, 2021). Possibly, then, predictions may remain 'useful' even when repeatedly wrong, simply because any amount of (relevant) pre-activated semantic content benefits processing. This fits the view wherein predictions naturally emerge from a representation of the discourse context, without requiring a separate, active prediction mechanism.

We conclude that prediction does not always operate in an optimal, Bayesian fashion (see also Marcus & Davis, 2013). Of course, one could attempt to rescue the rational adaptation hypothesis with an explanation of why people might disregard expected prediction error in one situation but not another (e.g., people adapt to sentence constraint rather than predictability, or people only adapt to prediction error on sentence-final words). However, these explanations are at this point merely post-hoc.

Conclusion

We observed pre-nominal prediction effects regardless of the ratio of prediction-disconfirming stimuli. Moreover, the little evidence supporting an effect of this ratio was primarily observed for unexpected, neuter-gender articles ('het'), which is inconsistent with the rational adaptation hypothesis of prediction (Kuperberg & Jaeger, 2016; Yan et al., 2017). We conclude that our participants generated predictions chiefly based on discourse-contextual constraints, regardless of expected utility and prediction error. In line with recent demonstrations (Nieuwland, 2021a,b), our study calls into question whether people estimate the reliability of their predictions and adapt their predictions accordingly. Linguistic prediction may be less 'rational' or Bayes optimal than is often suggested.

Data Availability

In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey, Chambers, Etchells, Harris, Hoekstra, Lakens et al., 2016), all stimuli, data and analysis scripts associated with this manuscript were available during the review process and remain available on OSF project "Lexical prediction in high and low predictive validity contexts" at https://osf.io/p8rta.

CRediT authorship contribution statement

Elise van Wonderen: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Mante S. Nieuwland:** Conceptualization, Methodology, Supervision, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Birgit Knudsen and Janniek Wester for help with data collection and three anonymous reviewers for helpful comments on a previous draft.

Appendix

See Figs. A1-A5.

References

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*(3), 471–485.
Altmann, G. T., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609.
Baggio, G. (2018). *Meaning in the brain*. MIT Press.
Bañón, J. A., & Martin, C. (2019). Anticipating information structure: An event-related potentials study of focus assignment via the it-cleft. *Neuropsychologia, 134*, Article 107203.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.
Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389.
Bowers, J. S., & Davis, C. J. (2012b). Is That What Bayesians Believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin, 138*(3), 423–426.
Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia, 135*, Article 107225.
Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language, 93*, 203–216.
Brouwer, S., Sprenger, S., & Unsworth, S. (2017). Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology, 159*, 50–65.
Brown, C. M., Hagoort, P., & Chwilla, D. J. (2000). An event-related brain potential analysis of visual word priming effects. *Brain and Language, 72*, 158–190.
Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.
Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 396–411.
Cevoli, B., Watkins, C., & Rastle, K. (2022). Prediction as a basis for skilled reading: Insights from modern language models. *Royal Society Open Science, 9*(6), Article 211837.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.
Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42–45.
Dave, S., Brothers, T., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2021). Cognitive control mediates age-related changes in flexible anticipatory processing during listening comprehension. *Brain Research, 147573*.
Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When Redundancy Is Useful: A Bayesian Approach to "Overinformative" Referring Expressions. *Psychological Review, 127*(4), 591–621.
Delaney-Busch, N., Morgan, M., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition, 178*, 10–20.
Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120394.
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), 1117–1121.
Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 1362–1376.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE, 8*(10), e77661.
Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition, 204*, Article 104335.
Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1461.
Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language, 95*, 200–214.
Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences, 360*(1456), 815–836.
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*(20), 8051–8056.
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.
Grodner, D. J., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In E. A. Gibson, & N. J. Pearlmutter (Eds.), *The processing and acquisition of reference* (pp. 239–272). MIT Press.
Heyselaar, E., Peeters, D., & Hagoort, P. (2021). Do we predict upcoming speech content in naturalistic environments? *Language, Cognition and Neuroscience, 36*(4), 440–461.
Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language, 35*(1), 66–85.
Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review, 116*(4), 717–751. https://doi.org/10.1037/a0017187
Hope, R. M. (2013). Rmisc: Ryan miscellaneous. *R package version, 1*(5).
Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.
Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research, 1706*, 196–208.
Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80–93.
Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31*, 19–31.
Hutchison, K. A. (2007). Attentional control and the relatedness proportion effect in semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 645.
Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition, 21*(2), 251–264.
Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language, 86*, 157–171.
Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia, 136*, Article 107291.
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience, 32*(8), 954–965.
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). Why the A/AN prediction effect may be hard to replicate: A rebuttal to Delong, Urbach, and Kutas (2017). *Language, Cognition and Neuroscience, 32*(8), 974–983.
Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*(4), 169.
Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124*, 66–71.
Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese, 183*(3), 339–373.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science, 78*(4), 601–627.
Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148.
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*, 239–253.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*, 32–59.
Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A Look around at What Lies Ahead: Prediction and Predictability in Language Processing. In *Predictions in the Brain: Using Our Past to Generate a Future*, Editor M. *Bar, Oxford University Press, 2011*, 190–207.
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*, 621–647.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161.
Kwisthout, J., & van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior, 3*(2), 174–188.
Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience, 25*, 484–502.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.
Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*.
Loerts, H., Wieling, M., & Schmid, M. S. (2013). Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing. *Journal of psycholinguistic research, 42*(6), 551–570.
Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60.
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.
Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Wcience, 24*(12), 2351–2360.
Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology, 7*, 120.
Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience, 32*(8), 1407–1427.
Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports, 8*, 1079. https://doi.org/10.1038/s41598-018-19499-4
Molinaro, N., Giannelli, F., Caffarra, S., & Martin, C. (2017). Hierarchical levels of representation in language prediction: The influence of first language acquisition in highly proficient bilinguals. *Cognition, 164*, 61–73.
Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61–64.
Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., et al. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science, 3*(1), Article 150547.
Myslín, M., & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition, 147*, 29–56.
Ness, T., & Meltzer-Asscher, A. (2021). Rational adaptation in lexical prediction: The influence of prediction strength. *Frontiers in Psychology, 12*.
Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia, 107427*.
Nieuwland, M. S. (2018). *An attempt to publish a replication attempt in a Nature journal, part 2 [Internet]. Retraction Watch*. Available from: https://retractionwatch.com/2018/05/09/an-attempt-to-publish-a-replication-attempt-in-a-nature-journal-part-2.
Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews, 96*, 367–400.
Nieuwland, M. S. (2021a). Commentary: Rational adaptation in lexical prediction: The influence of prediction strength. *Frontiers in Psychology, 12*, Article 735849. https://doi.org/10.3389/fpsyg.2021.735849
Nieuwland, M. S. (2021b). How 'rational' is semantic prediction? A critique and re-analysis of Delaney-Busch, Morgan, Lau, and Kuperberg (2019). *Cognition, 215*, Article 104848. https://doi.org/10.1016/j.cognition.2021.104848
Nieuwland, M. S., Arkhipova, Y., & Rodríguez-Gómez, P. (2020a). Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex, 133*, 1–36.
Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020b). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 375*, 20180522. https://doi.org/10.1098/rstb.2018.0522
Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*, e33468.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology, 47*(2), 204–238.
Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience, 31*(1), 4–18.
Otten, M., & Van Berkum, J. J. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research, 1291*, 92–101.
Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience, 8*(1), 89.
Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. Available online at: https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.
Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*, 1002–1044.
Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105–110.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*, 329–347.
Prasad, G., & Linzen, T. (2020). February 21). Rapid syntactic adaptation in self-paced reading: Detectable, but requires many participants. *PsyArXiv*. https://doi.org/10.31234/osf.io/9ptg4
R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia, 143*, Article 107466.
Roettger, T. B., & Franke, M. (2019). Evidential strength of intonational cues and rational adaptation to (un-) reliable intonation. *Cognitive Science, 43*(7), e12745.
Ryskin, R., Levy, R. P., & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia, 136*, Article 107258.
Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia, 138*, Article 107307.
Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: Revisiting the N400. *Journal of Cognitive Neuroscience, 25*(8), 1383–1395.
Szewczyk, J. M., & Wodniecka, Z. (2020). The Mechanisms of Prediction Updating That Impact the Processing of Upcoming Word: An Event-Related Potential Study on Sentence Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(9), 1714–1734.
Van Berkum, J. J. A. (1996). *The psycholinguistics of grammatical gender: Studies in language comprehension and production.* Nijmegen: Nijmegen University Press.
Van Berkum, J. J. A. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.
Van Berkum, J. J. A., Brown, C. M., 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.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190.
Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189.
Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003a). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters, 346*(3), 165–168.
Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003b). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex, 39*, 483–508.
Wicha, N. 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*, 1272–1288.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of open Source Software, 4*(43), 1686.
Wijnen, F., & Verrips, M. (1998). The acquisition of Dutch syntax. In S. Gillis, & A. De Houwer (Eds.), *The acquisition of Dutch*. Amsterdam: John Benjamins.
Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook. Chapman and hall/CRC. ISBN 9780367563837*. URL https://bookdown.org/yihui/rmarkdown-cookbook.
Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017, May 30). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *BioRxiv*, 143750. doi:10.1101/143750.
Zhang, W., Chow, W. Y., Liang, B., & Wang, S. (2019). Robust effects of predictability across experimental contexts: Evidence from event-related potentials. *Neuropsychologia, 134*, Article 107229.