Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Registered Report

Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials

Mante S. Nieuwland [a,b,c,*], *Yana Arkhipova* [a,d] *and Pablo Rodríguez-Gómez* [e,f]

[a] *Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands*
[b] *Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands*
[c] *Heinrich-Heine-University, Düsseldorf, Germany*
[d] *Cognitive Neuroscience Lab, Potsdam University, Potsdam, Germany*
[e] *Instituto Pluridisciplinar, Universidad Complutense de Madrid, Madrid, Spain*
[f] *Facultad de Psicología, Universidad Complutense de Madrid, Madrid, Spain*

ARTICLE INFO

Article history:
Protocol received 26 July 2018
Protocol approved 28 November 2018
Received 8 June 2020
Reviewed 31 July 2020
Revised 8 September 2020
Accepted 9 September 2020
Action editor Chris Chambers
Published online 30 September 2020

Keywords:
Language comprehension
Prediction
Replication
Grammatical gender

ABSTRACT

Numerous studies report brain potential evidence for the anticipation of specific words during language comprehension. In the most convincing demonstrations, highly predictable nouns exert an influence on processing even before they appear to a reader or listener, as indicated by the brain's neural response to a prenominal adjective or article when it mismatches the expectations about the upcoming noun. However, recent studies suggest that some well-known demonstrations of prediction may be hard to replicate. This could signal the use of data-contingent analysis, but might also mean that readers and listeners do not always use prediction-relevant information in the way that psycholinguistic theories typically suggest. To shed light on this issue, we performed a close replication of one of the best-cited ERP studies on word anticipation (Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005; Experiment 1), in which participants listened to Dutch spoken mini-stories. In the original study, the marking of grammatical gender on pre-nominal adjectives ('groot/grote') elicited an early positivity when mismatching the gender of an unheard, highly predictable noun, compared to matching gender. The current pre-registered study involved that same manipulation, but used a novel set of materials twice the size of the original set, an increased sample size (N = 187), and Bayesian mixed-effects model analyses that better accounted for known sources of variance than the original. In our study, mismatching gender elicited more negative voltage than matching gender at posterior electrodes. However, this N400-like effect was small in size and lacked

Author Note: The protocol for this report was approved and pre-registered at https://osf.io/tj4aw/. In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey, Chambers, Etchells, Harris, Hoekstra, Lakens, et al., 2016), all materials associated with this manuscript were available for review and remain available along with all data and scripts on OSF project "Gender Inflection Replication" at https://osf.io/jqhpz/.

* *Corresponding author*. Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, the Netherlands.
E-mail address: mante.nieuwland@mpi.nl (M.S. Nieuwland).

https://doi.org/10.1016/j.cortex.2020.09.007

support from Bayes Factors. In contrast, we successfully replicated the original's noun effects. While our results yielded some support for prediction, they do not support the Van Berkum et al. effect and highlight the risks associated with commonly employed data-contingent analyses and small sample sizes. Our results also raise the question whether Dutch listeners reliably or consistently use adjectival inflection information to inform their noun predictions.

1. Introduction

According to current theories of language comprehension, people implicitly and routinely anticipate upcoming words by activating their meaning and possibly other features in advance (e.g., Altmann & Mirkovic, 2009; Dell & Chang, 2014; Levy, 2008; Pickering & Gambi, 2018; Pickering & Garrod, 2013). The most convincing evidence for word anticipation or prediction is when a neural effect of a predictable word is obtained before that word is presented, for example, on a preceding article or adjective. When readers or listeners predict a specific noun, pre-nominal words that mismatch the predicted word elicit an enhanced event-related potential (ERP) response compared to matching words (for a review, see Kutas, DeLong & Smith, 2011). Several studies have provided such evidence in the past decades (e.g., DeLong, Urbach, & Kutas, 2005; Otten & Van Berkum, 2008, 2009; Van Berkum, Brown, Zwitserlood & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004). However, the replicability and consistency of the observed patterns have recently come into question (Ito, Martin, & Nieuwland, 2017a,b; Kochari & Flecken, 2019; Nieuwland et al., 2018), which has highlighted the need for replication of key findings. The current study aims to replicate one such highly influential ERP study on linguistic prediction (Experiment 1 of Van Berkum et al., 2005, henceforth VB05), which tested for effects of prediction on pre-nominal adjectives during comprehension of spoken Dutch mini-stories.

It has long been known that the language comprehension system works highly incrementally by incorporating novel input into the unfolding interpretation as soon as possible (e.g., Marslen-Wilson, 1975; Marslen-Wilson & Tyler, 1980). Rather than waiting for a word or sentence to finish, listeners can use the initial sound of a word to identify the potential meaning or reference of a word (e.g., Connolly & Phillips, 1994), and can relate each incoming word to the wider discourse context even before the word is completely and uniquely identifiable (e.g., Van Petten, Coulson, Rubin, Plante, & Parks, 1999). Moreover, research in the past two decades has reported that language comprehension is not merely incremental, but sometimes even predictive. Comprehenders can predict a word's meaning before it becomes physically available to them (e.g., Federmeier & Kutas, 1999), and maybe even its grammatical, orthographic, or phonological features (VB05; DeLong et al., 2005; Wicha et al., 2004; but see Nieuwland, 2019). Such word predictions, when correct, can facilitate processing by enabling people to activate word meaning more easily, reducing the 'workload' incurred by the predicted word.

Prediction can manifest itself in amplitude reduction of the N400 (Kutas & Hillyard, 1984; Lau, Almeida, Hines, & Poeppel, 2009; for a review, see; Kutas & Federmeier, 2011; Lau, Phillips & Poeppel, 2009; Van Berkum, 2009), a well-known event-related potential (ERP) component thought to reflect the access to or activation of semantic information in long-term memory (Kutas & Hillyard, 1980, 1984). Amplitude of the N400 is strongly correlated with the predictability of a word given its context (commonly operationalized as 'cloze probability', the probability of being used in a non-speeded, offline sentence completion test; Kutas & Hillyard, 1984). The prediction of an upcoming word or concept involves the pre-activation of associated semantic information, leading to easier access to the meaning of a word once it appears (e.g., Federmeier & Kutas, 1999). This assumed process of semantic pre-activation explains why words that are semantically related to an expected word or to the sentence context elicit reduced N400 amplitude, compared to unrelated words, despite being equally unexpected or plausible (Kutas & Hillyard, 1984).

But an amplitude reduction of the noun-elicited N400 alone does not provide clear evidence that a specific word was predicted ('lexical prediction'), because it is also compatible with a more passive pre-activation of related semantic content (which may emerge naturally from comprehension of the preceding context; for discussion, see Baggio, 2018; Kutas, DeLong, Smith, & Bar, 2011; Van Berkum, 2009). In addition, the observed effect could mean that the predictable noun is easier to integrate with the sentence than an unpredictable word, simply because it is a more plausible sentence continuation (for discussion, see Federmeier & Kutas, 1999; Nieuwland et al., 2020), regardless of whether it was actually predicted. For these reasons, researchers typically argue that evidence for lexical prediction is strongest when it is observed before the predicted noun is heard or read, and is obtained by comparing ERPs to words that themselves have little semantic meaning (e.g., the English articles 'a/an') and/or do not differ in meaning (e.g., the Dutch adjectives 'groot/grote', which have the same meaning but differ in the presence of the inflectional suffix '-e' to mark grammatical gender). Differential effects elicited by these prenominal critical words cannot be due to a difference in the meaning of the words themselves, and are therefore thought to arise from the phonological or grammatical relationship between the pre-nominal word and the predicted noun. ERP evidence for

lexical prediction has been reported from various prenominal manipulations in different languages, primarily Spanish, English and Dutch.

1.1. *Previous ERP studies with pre-nominal manipulations*

In a pioneering study by Wicha, Bates, Moreno, and Kutas (2003), native Spanish speakers listened to sentence pairs in which a relatively predictable (i.e., moderate-to-high cloze probability) or an unexpected and incongruent noun was replaced with a drawing. Crucially, gender-marked prenominal articles elicited an enhanced N400 ERP when their gender did not match that of the predictable nouns. In a follow-up experiment with written Spanish sentences, Wicha, Moreno, and Kutas (2003) found a very similar N400 effect of gender-mismatch with a predictable noun. In a subsequent experiment with written Spanish sentences but without accompanying drawings, Wicha et al. (2004) found that gender-mismatching articles elicited a different pattern, namely a positive ERP effect (P600) compared to matching articles. In more recent studies on comprehension of written sentences, gender-mismatch on prenominal articles was associated with N400-like effects, i.e., an enhanced negativity in the typical N400 time window (Dutch: Fleur, Flecken, Rommers, & Nieuwland, 2020; Otten & Van Berkum, 2009; Spanish: Foucart, Martin, Moreno, & Costa, 2014; Martin, Branzi, & Bar, 2018; Molinaro, Gianelle, Caffarra & Martin, 2017), although sometimes with a time course or scalp distribution unlike the typical N400 effects elicited by nouns.

In a series of studies on comprehension of Dutch mini-stories, VB05 reported evidence for prediction from a different manipulation also involving grammatical gender. These studies capitalized on the Dutch grammatical rule by which adjectives are marked with an inflectional suffix when they modify an indefinite noun of common gender but not of neuter gender. For example, the suffix on 'grote' in 'een grote boekenkast' (a large bookcase) agrees with the common gender of 'boekenkast', whereas lack of the inflectional suffix '-e' in 'groot' in 'een groot schilderij' (a large painting) agrees with the neuter gender of 'schilderij'. In Experiment 1 of VB05, participants listened to a two-sentence context that presumably led people to predict a specific noun (e.g., schilderij – painting). This context was followed by an adjective phrase that contained two gender-marked adjectives and either the predictable noun (e.g., 'groot maar onopvallend schilderij' – big but unobtrusive painting) or a less predictable noun of a different gender (e.g., 'grote maar onopvallende boekenkast' – big but unobtrusive bookcase). VB05 found that a mismatch between the inflectional suffix (or lack of thereof) on the first adjective and the gender of a predictable noun (e.g., 'grote' when the predictable noun was 'schilderij') elicited a more positive ERP than a match. This positive ERP effect had a very early onset, namely about 50–250 msec after the first acoustic difference between words with and without inflection, although when ERPs were time-locked to adjective onset the ERP difference had a later time-course (500–800 msec). In Experiment 2, participants listened to only the target sentences, which by themselves presumably did not lead to a specific noun prediction. Consistent with this hypothesis, no statistically significant ERP effect was obtained for the adjectives. In Experiment 3, participants read a subset of the materials from Experiment 1 in a self-paced reading study, where participants press a button to make each next word appear on the screen. The participants slowed down when the second adjective mismatched the predictable noun, compared to a match. Although no such effect was obtained on the first adjective (which would be a behavioral equivalent of the ERP results obtained in Experiment 1), these reading time results nevertheless supported the hypothesis that people anticipated the predictable noun (but see Guerra, Nicenboim, & Helo, 2018, for a recent failure to find prediction-related reading time results for Spanish sentence comprehension).

In two follow-up studies (Otten, Nieuwland, & Van Berkum, 2007; Otten & Van Berkum, 2008), the same suffix-based manipulation elicited ERP effects that were different from those obtained in Experiment 1 of VB05. In a study with spoken materials (Otten et al., 2007; henceforth OT07), prediction-mismatching adjectives elicited a negative ERP effect at right-frontal electrodes that started at about 300 msec and lasted until 600 msec (based on the associated scalp distribution, the authors were reluctant to interpret this effect as an N400 modulation), time-locked to the adjective onset. In a study with written materials (Otten & Van Berkum, 2008), prediction-inconsistent adjectives elicited a negative ERP effect that appeared as late as 900–1200 msec after adjective-onset.

The gender-effects reported by Wicha and colleagues, by OT07 and VB05, as well as by various others, indeed suggest prediction of a specific noun, but various questions about the functional significance of these effects remain. For example, it is unclear whether pre-activated information (which is presumably already available before an article or adjective is presented) includes grammatical gender (e.g., Pickering & Gambi, 2018; Wicha et al., 2004). It is possible that the initial prediction is limited to word meaning, and that people use gender information to evaluate whether the specific word can still appear. The second question is whether effects of gender-mismatch reflect the detection of a prediction mismatch or (also) the updating or revision of a prediction (for discussion, see Fleur et al., 2020; Nieuwland et al., 2018). Importantly, aside from these questions about interpretation, a major obstacle to any unitary interpretation of the available results is that qualitatively different types of effects have been obtained with (sometimes very) similar gender-based manipulations (for discussion, see Ito, Martin, & Nieuwland, 2017b; Kochari & Flecken, 2019). This could signal something meaningful, namely that different processes are engaged in each of these studies. However, it could also signal the problem with statistically significant effects obtained in noisy, small-sample settings, which are associated with increased probability of overestimated or wrong-sign effect estimates (e.g., Gelman & Carlin, 2014; Vasishth, Mertzen, Jäger, & Gelman, 2018).

This problem with small-sample effects has also surfaced in another well-known demonstration of prediction, a study on English sentence comprehension by DeLong et al. (2005), who capitalized on the phonological rule for indefinite articles (i.e., 'a/an' signals that the next word will start with a consonant or a vowel, respectively). Indefinite articles that

mismatched a predictable noun in terms of phonology (e.g., 'an' if the predictable word was 'kite') elicited an enhanced N400 compared to matching articles. This effect is therefore sometimes taken to demonstrate phonological prediction (Pickering & Garrod, 2013).

However, the DeLong et al. results have proven controversial. A study by Martin, Thierry, Kuipers, Boutonnet and Costa (2013) reported a similar effect for mismatching articles, but differences in the analysis complicated a quantitative and qualitative comparison to the DeLong et al. results (for discussion, see Ito et al., 2017a,b). Another study with this manipulation (Ito et al., 2017a,b) did not obtain a reliable effect of gender mismatch. Moreover, a recent, large-scale (N = 334) direct replication study (Nieuwland et al., 2018) failed to replicate the result of DeLong et al. in an analysis that duplicated the original, and found no statistically significant effect in an additional analysis that took into account subject- and item-level variance. Nieuwland et al. concluded that the 'a/an' article-effect may indeed be non-zero, but that it is likely far smaller than originally reported and too small to observe without very large sample sizes. Nieuwland et al. further speculated that the a/an manipulation does not elicit reliable or strong prediction effects because these articles are diagnostic of the next word, which need not be a noun (e.g., 'an old kite'). Unexpected articles thus do not actually refute the upcoming noun altogether, but signal that the noun cannot appear immediately after the article. Stronger or more reliable effects might therefore be obtained with gender-marked articles or adjectives, which can disconfirm the predicted noun because they agree with that noun in gender irrespective of intervening words.[1]

Given the strength of gender agreement relationships, the apparent lack of consistent patterns across studies with gender-based manipulations may seem disconcerting. Establishing the nature and timing of such effects is critical to developing hypotheses about the mechanisms that underlie the generation and evaluation of predictions. For example, prediction-mismatching suffixes elicited an early onset, positive ERP effect in VB05, who did not commit to a specific functional interpretation of this effect beyond the conclusion that the effect demonstrated lexical prediction (see also Van Berkum, 2004). However, this positive ERP response could be related to P600 effects seen for syntactically unexpected information (e.g., Osterhout, Holcomb, & Swinney, 1994) and for morphosyntax agreement mismatch (e.g., Tanner, Grey, & van Hell, 2017; Wicha et al., 2004). Such P600 effects are thought to reflect a reanalysis or syntactic integration process (e.g., Kaan, Harris, Gibson, & Holcomb, 2000; Kaan & Swaab, 2003). In contrast, studies reporting N400 or N400-like effects suggest that predictions impact the activation of word meaning (lexical access), and have led some authors to argue that people even predict the specific form of the prenominal article itself along with the noun (DeLong et al., 2005). However, before attempting to explain the different effects of prediction and their association with specific linguistic manipulations or experimental procedures, the field needs to establish which of the key findings can be replicated with similar methods and materials, in a sufficiently large sample to obtain a sufficiently reliable and plausible effect estimate. This is not a trivial issue in a research field where it has long been and still is rather common to select a dependent variable based on visual inspection of low-sample ERP data (Kilner, 2013; Luck & Gaspelin, 2017), a practice that leads to over-estimated effect sizes and higher rates of false positives (e.g., Gelman & Carlin, 2014; Gelman & Loken, 2013; Vasishth, Mertzen, et al., 2018; Vul, Harris, Winkielman & Pashler, 2009). Moving away from ERP analysis based on visual inspection, recent ERP studies on language comprehension have pre-registered data processing steps and statistical analyses, and explicitly distinguish between confirmatory and exploratory analyses (e.g., Coopmans & Nieuwland, 2020; Fleur et al., 2020; Nieuwland et al., 2018; Sassenhagen & Bornkessel-Schlesewsky, 2015).

1.2. *The current study*

The current study tries to replicate the main result obtained in Experiment 1 of VB05, along with that of OT07, as it used the same manipulation and similar materials. Like DeLong et al. (2005), VB05 is an influential and highly cited ERP study (at time of writing, 843 citations on Google Scholar) that features in major theoretical reviews on linguistic prediction (e.g., Altmann & Mirkovic, 2009; Kutas & Federmeier, 2011; Pickering & Garrod, 2013). However, like DeLong et al., VB05 has yet to be successfully replicated. The only available study with the same gender-based inflection-manipulation found an effect in the opposite direction (OT07). Moreover, the key evidence reported by VB05 came from an analysis in a time window that was based on visual inspection of the grand-average ERP waveforms (p. 448). This procedure has long been, and probably still is common (see Kilner, 2013; Luck & Gaspelin, 2017; see also Nieuwland, 2019, for related discussion), although it is not always explicitly mentioned in Methods sections (for example, in some work by the first author of this paper, see Nieuwland, Ditman, & Kuperberg, 2010; Nieuwland & Kuperberg, 2008). Selection of a spatial and/or temporal region-of-interest is an appropriate method to sidestep the requirement for multiple comparison correction. However, it is only robust and valid when the selection is independent of the data, whereas it has an increased risk of false positives if the selection is based on where an effect looks strongest in grand-average ERPs (e.g., Kilner, 2013; Luck & Gaspelin, 2017; see also http://deevybee.blogspot.com/2013/06/interpreting-unexpected-significant.html). Therefore, we deem it important to perform a pre-registered attempt to replicate the effect of key results of VB05 in a sufficiently powered study (see Methods for power analyses).

The ERP effect of prediction-mismatching inflections is also important because this manipulation tests for online prediction in a unique, possibly quite subtle way. Dutch adjective-suffix inflection is a better cue to noun gender than definite articles. In Dutch, 'het' is also used for diminutive nouns, whereas 'de' is used for plural nouns, irrespective of noun gender. This pattern is different for adjectives once a plural noun has already been ruled out by the preceding indefinite, singular article 'een'. The absence of a suffix is compatible with a diminutive noun irrespective of gender

[1] This holds true for Spanish, whereas Dutch definite articles are not fully reliable cues to noun gender because they also precede plural and diminutive nouns irrespective of noun gender.

('een leuk boekje/tafeltje', a nice little book/table), but suffix-presence is only compatible with a common gender noun. Absence and presence of the suffix therefore have different repercussions for whether the general semantic meaning of the predicted noun can still follow: if one predicted 'boek', then 'een leuke' disconfirms that lexical meaning; if one predicted 'tafel, then 'een leuk' does not entirely disconfirm that meaning, as the diminutive 'tafeltje' could follow. Even in an experiment where no such diminutives appear, participants may be sensitive to the possibility of the expected word appearing in diminutive form and therefore do not take 'missing' inflection as a cue that the predicted meaning is wrong (see also Nieuwland et al., 2018).

Aside from this issue of 'cue reliability', gender-mismatching inflections might be relatively hard to detect compared to prediction-mismatching gender-marked articles ('el/la' in Spanish, 'de/het' in Dutch), and therefore be less likely to yield prediction effects. The inflection manipulation relies on the detection of an absent or present inflection within a short time frame, whereas the article manipulation relies on detecting two entirely different lexical items. In addition, it involves the detection of prediction-relevant information from an adjective that itself is relatively unexpected[2] and contains novel semantic content. This differs from detection of a gender-marked article that itself might be predicted along with the noun and therefore generate stronger effects. For these reasons, the inflection-based gender manipulation might pose a stronger test of the predictive use of gender-relevant information during language comprehension than the more commonly used article-based gender manipulation (e.g., Foucart et al., 2014; Martin et al., 2018; Wicha et al., 2004).

The current study aims to replicate previously observed patterns for prediction-mismatching adjective inflections (VB05; OT07). We use a novel set of experimental materials that is twice the size of that in VB05, and based on materials from a recent ERP study on lexical prediction (Fleur et al., 2020). Fleur et al. constructed two-sentence mini-stories that either suggested a definite noun phrase (e.g., 'het boek', the book) or an indefinite noun phrase ('een boek', a book) as its most likely continuation. Following these contexts, participants saw a definite noun phrase with either the expected noun ('het boek') or an unexpected, different-gender noun ('de roman'). Using pre-registered data preprocessing procedures and statistical analyses, Fleur et al. found that gender-mismatching articles elicited an enhanced N400 compared to gender-matching articles[3] (see also Otten & Van Berkum, 2009, for similar results). These findings are relevant for the current study, because they show that we use materials that have already demonstrated relevant ERP effects of prediction.

In the current study, we only used a subset of the story contexts of Fleur et al., namely those in which an indefinite noun phrase was the expected continuation, with a minimum cloze value of 75% for both articles and nouns (the cut-off used for the main analysis in VB05). To match the manipulation of VB05, we added two adjectives to each target noun phrase (see Table 1 for an example story). Using Bayesian mixed-effects model analyses, we take a spatiotemporal region of interest approach to test for effects of prediction-match on average voltage values for each trial. The choice of ROIs and time windows was based on VB05 and OT07. These analyses aim to answer the question of whether previous neural evidence of lexical prediction from gender-marked, pre-nominal adjectives can be replicated. We pre-register further exploratory analyses to test the effect of prediction-match with traditional ANOVAs (on average values per condition per subject, following VB05 and OT07), and to test whether the prediction-match effect is similar for common and neuter gender nouns.

2. Methods

All aspects of the Methods were identical to those of VB05, except as noted.

2.1. Participants

A total of 189 right-handed native speakers of Dutch (range 18–40 years) were recruited through the subject pool of the Max Planck Institute for Psycholinguistics (Nijmegen, the Netherlands), which is the current version of the subject pool used by VB05. Our participants had been raised in a monolingual household, and were free from neurologic impairments, neurologic trauma, neuroleptics use, and any known language or hearing difficulties. The participants had not taken part in any of the completion norm tests or in Fleur et al. (2020). All participants gave informed written consent to take part in the experiment, which was approved by the Ethics Committee for Behavioral Research of the Social Sciences Faculty at Radboud University Nijmegen in compliance with the Declaration of Helsinki. Likewise, all participants consented to making their anonymous EEG and behavioral data publicly available. Participants were paid 8 Euro per hour for their participation. Our participant recruitment (and possibly payment) thus differed from VB05, mostly because we imposed more requirements for participation (monolingual household, no participation in pre-tests, consent to data availability).

2.2. Initial sample size calculation

An initial, minimum sample size was determined by a mixed-effects a priori power analysis with the SIMR package (Green & MacLeod, 2016). Because no single-trial data were readily available from VB05 and OT07, single-trial data were adapted from a previously published study by Ito et al. (2017a,b; Experiment 1), which tested for predictive effects using the prenominal a/an manipulation in native speakers of English.

[2] Here, 'unexpected' means that the adjectives did not appear in cloze test responses. However, it is possible that in an experiment that contains many pre-nominal adjectives, participants may start to expect pre-nominal adjectives through the course of the experiment, either as an expectation of a word category or perhaps even of specific adjectives.

[3] This prediction-effect was larger when participants expected a definite noun phrase compared to when they expected an indefinite noun phrase (in the latter case, the definiteness of the article was itself unexpected and associated with overall enhanced N400 amplitude).

Table 1 – Dutch example mini-story with prediction-matching and -mismatching continuations, plus approximate translation.

Context	Critical adjective phrase	Ending
Tv-kijken en social media vindt Cas maar niks. Hij leest eigenlijk het liefste gewoon een	**Match:** *dik$_{neu}$ en spannend$_{neu}$ boek$_{neu}$* **Mismatch:** *dikke$_{com}$ en spannende$_{com}$ roman$_{com}$*	*van Stephen King.*
Approximate translation		
Cas doesn't really like television and social media. Actually he prefers to just read a	**Match:** *thick$_{neu}$ and exciting$_{neu}$ book$_{neu}$* **Mismatch:** *thick$_{com}$ and exciting$_{com}$ novel$_{com}$*	*by Stephen King.*

Ito et al. had a similar number of items and subjects as the Dutch gender studies (Ito et al., 64 items and 23 subjects; VB05, 75 items and 24 subjects; OT07, 80 items and 29 subjects), and used a categorical variable for expected/unexpected conditions with the following model:

voltage ~ condition + (condition | subject) + (condition | item)

Of note, the effect sizes in VB05 and OT07 may overestimate the true underlying effects because the analyses were based on visual inspection of the data, and because a result that is statistically significant in a noisy, low-powered experiment is likely to have an overestimated effect size (e.g., Gelman & Carlin, 2014; Vasishth, Mertzen, et al., 2018). Therefore, for current purposes voltage values in the expected and unexpected conditions of the Ito et al. data were tweaked such that they yielded an estimated difference that was smaller than the smallest reported significant effect in VB05 and OT07 (following the guidelines of this journal). In this data, the effect of condition was not statistically significant at the alpha = .05 level [$\beta = .37$, 95% Wald CI = (−.18, .93), $t = 1.32$, $p = .19$], and quite similar to the patterns obtained in a large-scale (N = 334) replication study using the same a/an manipulation (Nieuwland et al., 2018). For the simulation, the number of items for the model was extended to 150 to match the number used in the current study. Power analysis by simulation (number of simulations = 1000) showed that 90 subjects was sufficient to detect an effect at a significance level of alpha = .02 with 90% power. We set the initial sample size slightly higher at N = 100, which is more than 4 times the sample size of VB05, and refers to the participants ultimately used for statistical analysis, and thus to the minimum number of participants that were tested. Participants were excluded from the statistical analysis using pre-defined criteria (a response accuracy under 75%, or insufficient trials after artefact rejection, described in the data pre-processing section). Each excluded participant was replaced by another participant.

However, this sample size was set as a minimum, because the simulation did not take into account the potential effects of the absence/presence of the inflection, and does not guarantee the Bayes Factor evidence strength required by this journal. In the original studies, which used ANOVAs, the absence/presence of inflection was approximately balanced across items. In the current study, however, this factor is explicitly accounted for in the model (see Sassenhagen & Alday, 2016), using a more powerful analysis that simultaneously takes into account sources of variance (subjects, items, presence of inflection) that were not included in the original study's ANOVAs. This is important because, not only may the effect of match differ for different-gender adjectives (see the pre-registered exploratory analysis), unaccounted-for variation that is orthogonal to the effect of interest (e.g., random intercept variation) can reduce power, while unaccounted-for variation that is confounded with our effect of interest (e.g., random slope variation) can drive differences between means, with increased risk of false positives and overestimation of effect size (for discussion, see Barr, Levy, Scheepers, & Tily, 2013). Therefore, our final sample size was not based on the a priori power analysis and we continued to increase our sample size from 100 to 200 in steps of 20 participants until we reached the Bayes Factor evidence strength required by this journal (see Statistical Analysis). However, our laboratory was shut down due to the unfolding covid-19/coronavirus pandemic when we reached 189. Because we had no view of a continuation of testing in the foreseeable future, and because we were confident that the remaining 11 to-be-tested participants would not change our conclusions, we were granted an early sampling finish by the editor.

2.3. Materials

Instead of the materials used in the original study (VB05), we used a suitable set of stimuli readily available from a previous study (Fleur et al., 2020). This set of materials was created following a similar procedure as that of the original, was larger than that of the original, and had already been normed for cloze probability. Moreover, as stated in the introduction, Fleur et al. had already demonstrated a prediction-consistency N400 effect on pre-nominal articles using these stimuli (see also Otten & Van Berkum, 2009).

The critical stimuli for the current study consisted of 150 mini-stories, each of which had one context sentence and two possible target sentences. Each mini-story was written to suggest a specific combination of an indefinite article plus predictable noun in the target sentence. In a cloze test (N = 20), each mini-story was truncated before the article,[4] and participants were asked to complete each story. The stories in the current study were completed by at least 75% of the respondents using the same combination of the expected indefinite article and noun.[5] The average cloze probability of the indefinite articles was 95.9% (SD = 5.9%, range 75–100%), and

[4] Materials in the cloze test of VB05 were truncated after the indefinite articles. Our procedure differed in where the materials were truncated because we also wanted to obtain cloze values for the prenominal articles (Fleur et al., 2020).

[5] Different spellings of the same word were permitted and counted towards the same response, such as 'tattoo/tatoeage' or 'tv/televisie'.

93.4% for the nouns (SD = 6.4%, range 75–100%). These values are numerically higher than those of the original studies.[6] Of the 150 predictable nouns, 80 were common gender 'de' words, and 70 were neuter-gender 'het' words.[7]

After the norming, we added two adjectives (separated by a function word, e.g., 'groot en sterk' or 'grote en sterke', which both mean 'big and strong') in between the indefinite article and the critical noun. If one or more participants in the cloze test had used an adjective to complete a sentence fragment (which was rare), that adjective was not used for that item in the EEG experiment. For each mini-story, we created a prediction-matching and -mismatching condition. In the matching condition, the absence or presence of an overtly realized suffix '-e' on both adjectives matched the gender of the predictable noun ('dik en spannend' when the predictable noun is 'boek'), and the second adjective was followed by the predictable noun. In the mismatching condition, the absence or presence of the suffix mismatched the gender of the predictable noun ('dikke en spannende' when the predictable noun is 'boek'), and the second adjective was followed by a different-gender noun that was semantically possible but less predictable (e.g., 'roman'). These alternative nouns had appeared at most only once in the cloze responses (average cloze value .2%, SD = .9%, range 0–5%). Like in the materials of VB05 and OT07, there was no overlap in the set of predictable nouns and alternative nouns (which means that a direct comparison between these nouns may be confounded by lexical variables). All sentences were grammatically correct. The critical nouns were never sentence-final, and all subsequent words were identical for the matching and mismatching condition of a given story.

An additional set of 120 filler mini-stories of 2 sentences each was added to the materials. Sixty fillers were similar to the prediction-match condition: they also contained high cloze nouns (average cloze 74.5%, SD = 17.7%, range 29.4–95%), preceded by pre-nominal adjectives. Due to these fillers, highly-constraining stories in the entire experiment were almost twice as likely to end in a predictable than an unpredictable noun. We added these fillers to counter the argument that participants will adapt to unexpected syntactic information (i.e., start expecting unexpected information) and therefore not show prediction-consistency effects. Such adaptation-effects have been reported, albeit only for frequent repetition of an unexpected syntactic structure (see Fine, Jaeger, Farmer, & Qian, 2013; but see also a recent failure to replicate this type of adaptation effect,; Stack, James, & Watson, 2018), not for varied sentences structures as used in the current experiment. We have also added 60 relatively non-constraining stories to increase the variability of our materials and to make the ratio between the experimental and filler stories more similar to that of VB05. These were adapted from a subset of low-constraint materials used by Otten and Van Berkum (2009; 'prime control' stories), and they did not contain nouns preceded by two adjectives in the second sentences.

The current study thus used different fillers and a slightly different ratio between experimental and filler items (150/120) than VB05 and OT07 (120/150 and 160/90, respectively), although it was similar to these previous studies in using only grammatically correct and semantically coherent/plausible mini-stories as fillers. Although the possible effect of the number and type of fillers on comprehension is not precisely known, some ERP studies suggest that a high proportion of prediction-licensing materials actually boosts predictive processing (e.g., Brothers, Swaab, & Traxler, 2017; Lau, Holcomb, & Kuperberg, 2013). In addition, Fleur et al. (2020) obtained N400 evidence for prediction on prenominal articles in a study that only included high-constraint sentences, of which 66% contained a critical 'de/het' article. For these reasons, we do not think there is a convincing a priori argument that our materials will elicit less predictive processing than those of VB05 and OT07.

The mini-stories were recorded with a normal speaking rate and intonation by a female native speaker.[8] Recordings followed the procedure described in VB05. Target sentences of the critical stories were recorded in both conditions. The context sentence was recorded once, together with either the prediction-matching target sentence or the prediction-mismatching target sentence (counterbalanced over stories). We ensured that the critical inflections were always clearly distinguishable from the subsequent word. The recordings of context and target sentences were stored separately. From the target sentence recordings, we identified the acoustic onset of the critical adjectives, of the critical inflections therein, and of the critical noun. Inflection onset was determined as the moment at which adjectives begin to differ between conditions in terms of their respective phonemes, following VB05.

Unlike VB05 and OT07, we included comprehension questions to encourage participants to pay attention to the meaning of the stories, and as a means to exclude participants who did not pay sufficient attention. A potential null effect of prediction is then unlikely to result from participants not paying attention to the meaning of the stories. In our view, the importance of ruling out such a 'lack of attention' account balances out the slight deviation from the original studies. It is not known whether and how comprehension questions – that are orthogonal to the manipulation of interest – change the way that people process the meaning of linguistic stimuli. To

[6] For comparison, the expected nouns in VB05 had an average cloze of 86% (SD = 6%, minimum = 75%) while the unexpected nouns had an average of 2% (SD = 3%). The expected nouns of OT07 had an average of 74% (SD = 14%, range 53–100%), whereas the unexpected nouns had an average of 3% (SD = 6%). Although the difference in noun cloze-values between the current study and the original study is small, it is unlikely to make it harder to find a prediction-effect in the current study, given that higher noun-cloze is associated with more, not less predictive processing (e.g., Kutas & Federmeier, 2011).

[7] This asymmetry was also present in the original materials (e.g., 40/30 de/het words in VB05, and 98/62 in OT07).

[8] The total set of story materials and recordings are available on our OSF page. For comparison, two sample recordings of VB05 are available online on http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:60092:12. Automated articulation rate analysis (number of syllable/speaking time; see De Jong & Wempe, 2009) on the two samples from VB05 and 15 samples of the current recordings suggested that the articulation rate in VB05 was higher than in the current study (5.48 compared to 3.48). Lack of a clear prediction effect in the current study is therefore unlikely to result from participants not being able to keep up with the articulation rate.

the best of our knowledge, there is no evidence to suggest that comprehension questions cause participants to process stimuli less predictively compared to when there are no comprehension questions. Moreover, our materials including the comprehension questions largely overlap with and are highly similar to those of Fleur et al. (2018), who obtained relevant ERP evidence for prediction on pre-nominal, gender-marked articles, as has been also reported in experiments without comprehension questions (Otten & Van Berkum, 2009). In our study, 80 of the 270 stories (29%) were followed by a yes/no comprehension question (38 follow a filler story, 42 follow a critical story; 40 questions were designed to elicit a 'yes' response, and the other 40 to elicit a 'no' response). Each question was answerable from the preceding mini-story, irrespective of the critical manipulation when it followed a critical story. Participants who answered fewer than 60 from the 80 questions correctly (75%) would have been replaced by new participants who are presented with the same materials. However, none of our participants met this exclusion criterion, and they had on average 96% correct answers (SD = 3, range = 83–100).

Materials were organized into two trial lists. In one list, one half of the critical stories was presented in the matching condition and the other half in the mismatching condition, and vice versa in the second list. Conditions within each list were pseudo-randomized such that no more than 2 filler stories were presented successively and no more than 3 matching or mismatching critical stories were presented successively. Both lists start with 5 practice stories, including two practice comprehension questions.

2.4. Procedure

After electrode application, subjects sat in a sound-attenuating booth and listened to the stories from a speaker placed in front of them. Although participants in VB05 listened to the stories over earphones, we decided to present the stories over speakers, which was also done in later work by Van Berkum and colleagues, including OT07. This method of delivery is more comfortable for participants and, unlike a headphone setup, does not introduce additional artefacts. Our participants were asked to listen attentively to the stories and answer the comprehension questions. After the 5 practice stories, the 270 stories were presented in 5 blocks of 54 items, separated by rest periods.

Participants self-paced through the experiment by button press. Upon pressing, each trial started with a 300 msec auditory tone, followed by a 700 msec silence, the spoken context sentence, a 1000 msec of silence, and the spoken target sentence. Participants were instructed to sit still and to refrain from eye-movements and blinks during the second, target sentence. To signal to subjects when to sit still and refrain from eye-blinks and eye-movements, an asterisk was displayed from 500 msec before the onset of the target sentence to 1000 msec after the sentence offset. If a trial was followed by a comprehension question, the question appeared in its entirety on the screen upon disappearance of the asterisk. The yes/no answer options were presented below the question, on the left or right side of the screen, and participants gave their response by clicking on a corresponding left or right button. The estimated, minimal 'time-on-task' for the EEG experiment was about 65 min (compared to the reported 75–80 min in VB05/OT07).

After the EEG experiment, participants completed four short computerized tests. Because these tests were included for exploratory analyses that are not reported in this paper, we only briefly discuss them here. More detailed descriptions can be found in the cited references. The first test was a Dutch version of the Author Recognition Test (Vander Beken & Brysbaert, 2017; Stanovich & West, 1989; see also; Hartung, Burk, Hagoort & Willems, 2016), a measure of print exposure that has been shown to correlate with reading skills (e.g., Acheson, Wells, & MacDonald, 2008). Participants indicated which ones out of 132 names correspond to authors they know (90 are author names, 42 are made up names). The second test was a Dutch prescriptive grammar test, adapted from Hubers, Snijders, and De Hoop (2016) and Favier, Meyer, and Huettig (2018). Participants heard 40 spoken sentences and indicated for each of them whether or not it is a correct Dutch sentence. Half of the sentences contained grammatical norm violations (expressions that are frequently used by some groups of speakers but that are considered ungrammatical by others). The third test was the Peabody vocabulary test (Dunn & Dunn, 1997; Schlichting, 2005), in which participants hear words and have to select matching pictures from sets of four options. The fourth test was the Dutch version of STAIR-S4WORDS, an adaptive test for assessing receptive vocabulary size (Hintz, Jongman, Dijkhuis, van 't Hoff, Damian, Schröder et al., 2018). On each trial, the participant saw a word or a non-word foil (ratio 3:1) and indicated whether or not they knew the item.

2.5. EEG data recording and pre-processing

The EEG was recorded from 27 active electrodes (Fz, FCz, Cz, Pz, Oz, F7/8, F3/4, FC5/6, FC1/2, T7/8, C3/4, CP5/6, CP1/2, P7/8, P3/4, O1/2) relative to a left-mastoid reference electrode, along with activity at a right-mastoid reference channel and 4 EOG channels. The electrode locations were similar but not identical to those of VB05 and OT07, however they still allowed for a similar quadrant-based ROI analysis as reported in these earlier studies. Data was recorded with a BrainAmp DC amplifier, at a sampling rate of 1000 Hz, using a time constant of 10 sec (.016 Hz) and high cut-off of 250 Hz in the hardware filter (this high cut-off differed from the 70 Hz used in VB05 but matched that of OT07 and allowed for later analysis in the 30–100 Hz gamma frequency band), with an additional high cut-off of 100 Hz in the recording software.[9] Electrode impedance was kept below 20 kΩ where possible, which differed from the procedure in VB05 (who used passive electrodes), but it is under the guidelines of the hardware manufacturer, and lowering impedances to under 3 kΩ takes prohibitively long and could cause too much discomfort to

[9] Because the BrainRecorder software could only achieve the pre-registered filter-outcome by a combination of hardware and software filtering, we deviated from the pre-registered protocol in the filter settings. To err on the safe side, we furthermore doubled the pre-registered sampling rate. We matched the precise VB05 hardware filter settings during offline filtering.

participants. Because we had a large number of trials and participants, and because the recordings took place in an air-conditioned room to ensure a cool and dry environment, such impedance differences were very unlikely to meaningfully impact statistical power (see Kappenman & Luck, 2010). In addition, our sample size calculation was based on high impedance (Biosemi) data and we had already demonstrated prediction ERP effects in our lab with a less restrictive impedance threshold (<25 kΩ, Fleur et al., 2020). Regardless, impedance values were stored before and after the experiment for potential checks.

Data pre-processing was performed using BrainVision Analyzer. First, bad EEG channels were identified through visual inspection (as electrodes showing poor signal for at least half of the experiment due to blocking, faulty connectivity or other large-amplitude artefact) and interpolated through spherical splines. Our pre-registration stipulated interpolation of maximally 4 EEG channels per participant. We ended up interpolating 1 channel from 13 participants, 2 channels from 6 participants, 3 channels from 3 participants, and 4 channels from 1 participant. An interpolation procedure was not used or mentioned in VB05 but we used it to avoid unnecessary data loss. Then, the continuous data were filtered with a .02–70 Hz (24 dB) Butterworth IIR band-pass filter (we also included a 50 Hz notch filter, which was not pre-registered) to match the hardware filter of VB05. We then segmented the data into epochs from 150 before to 2100 msec after the onset of the first critical adjective. However, because this epoch did not always extend 1000 msec beyond the later noun as in VB05, we created separate segments for the nouns ranging from 150 msec before to 1000 msec after noun onset. We used separate segments instead of longer segments that included all the critical adjective and noun data, because longer segments would have also included artefact-rich post-sentence data in many of the trials. Each segment was then screened for large muscle artifacts, electrode drift, and amplifier blocking. We corrected for artefacts in the segments (due to eye-movements, blink, cardiac or steady muscle activity[10]) using Independent Component Analysis (trained on segments extracted from continuous data that was filtered with a .1–70 Hz zero phase shift Butterworth band-pass filter plus notch filter). This procedure was not included in VB05 but was used in OT07 and avoids unnecessary data loss, which is important because participants might find it hard or distracting to avoid blinking and eye-movements.

We subsequently extracted smaller segments running from 150 msec before to 1000 msec after the onset of adjectives, inflections and nouns. For each segment, baseline correction was performed by subtracting average voltage in the relevant (of three) 150 msec prestimulus interval from the entire segment. We then applied an automated artefact rejection procedure that rejects epochs with values exceeding ± 100 μV. Although no such rejection procedure was applied in VB05, we felt it was important to apply one objective artefact criterion instead of only relying on visual inspection, also to remove segments with artefacts that had been overlooked during visual inspection.

In VB05, no participant exclusion criteria were mentioned, but it was stated that, on average, 20.5% of all trials were rejected (based on visual inspection of the data), and that there were no asymmetries between conditions. Here, participants were excluded if their comprehension question accuracy was under 75%, if they had fewer than 40 remaining trials from the initial 75 trials (53%) in any of the 6 conditions (matching/mismatching adjectives time-locked to onset or inflection, matching/mismatching nouns), or if they had fewer than 50 remaining trials (66.7%) on average across all 6 conditions. Two participants were excluded, which left us with a total of 187 participants, who had, on average, 72 match and 72 mismatch trials time-locked to inflection, 72 match and 72 mismatch trials time-locked to onset, and 73 match and 73 mismatch noun-trials (corresponding to a trial rejection rate of approximately 4% for inflection and adjective onset trials, and 3% for noun trials). All raw data and pre-processed data are available on https://osf.io/jqhpz. Before statistical analysis, we downsampled the data segments to 500 Hz, matching that of VB05.

2.6. Statistical analysis

Based on the results reported in VB05 and OT07, we performed analysis on the average voltage within pre-defined spatio-temporal regions of interest (ROIs) for each trial. We defined five spatial ROIs: 4 different quadrant-selections of 4 electrodes each (left-anterior: F7/F3/FC5/FC1, right-anterior: F8/F4/FC6/FC2, left-posterior: P7/P3/CP5/CP1, right-posterior: P8/P4/CP6/CP2) and 1 midline selection of 5 electrodes (Fz/FCz/Cz/Pz/Oz). For each ROI, we averaged activity per trial within a 50–250 msec time window after inflection onset (VB05) or a 300–600 msec time window after adjective onset (OT07). We performed separate tests at each spatiotemporal ROI, so that we did not miss effects at ROIs where VB05/OT07 did not observe statistically significant effects. We defined three spatiotemporal ROIs for the nouns (based on VB05), such that we averaged activity within a 300–500 msec time window after noun onset within the left-posterior quadrant, the right-posterior quadrant, and the midline.

2.7. Adjective analyses

Using the trial-level data from these ROIs, we performed Bayesian linear mixed effects model analysis using the 'brms' package (Bürkner, 2017) in the R software (R Core Team, 2018), which fits Bayesian multilevel models in the Stan programming language (Stan Development Team, 2018) with formula syntax that is similar to that of the 'lme4' package (Bates, Maechler, Bolker, & Walker, 2014). In addition to Bayes Factor hypothesis testing, this analysis allows for Bayesian parameter estimation: an estimation of the uncertainty about the magnitude of the observed effect (by computing the point estimate along with a credible interval of effect sizes), including the uncertainty about the effect being greater than

[10] The pre-registered procedure only covered ICA correction for blink artefacts but we noticed that this left too many artefact-related components in the data, in particular artefacts from horizontal eye movements. For comparison, we re-ran our main analyses using the originally pre-registered ICA procedure. This led to exclusion of two additional subjects but yielded results that are very similar to the results reported here. Relevant data and results are available on our OSF page.

zero, while including previous results as constraints on plausible values. For each adjective-ROI, we constructed a model that included the deviation-coded fixed factor 'match' (match/mismatch with the predictable word) and the deviation-coded fixed factor 'gender' (common/neuter suffix). The factor 'gender' captures the potential effect of the absence or presence of the inflectional suffix '-e', which is not manipulated independently from 'match' within each item (and therefore not included as a random slope for 'item').

voltage ~ match + gender + (match + gender | subject) + (match | item)

We followed the suggestions for replication studies by Dienes (2014) and Dienes and McLatchie (2018), namely to use, for the effect of interest, a prior with a zero mean and a standard deviation that is the previously reported effect size. The previous effect sizes went in both directions (i.e., a positivity in VB05, a negativity in OT07) and corresponded to roughly a .75 μV difference. To perform replication tests of those studies, the prior on the effect of match was a normal distribution with a zero mean and SD = .75 μV (here and elsewhere, the normal distribution was used lacking an obvious reason to assume non-normality). In other words, there is a 95% prior probability that the parameter lies between −1.5 and 1.5 μV (of note, these are two-sided tests, but we make up for the associated overall lower prior density in our sampling plan described below).

We also included a prior for the effect of gender, which centered on mean zero with a normal distribution because the effect could be positive or negative, and a prior SD that was the same as for match (assuming the effect of gender is unlikely to be bigger than that of match), such that there is a 95% probability that the parameter lies between −1.5 and 1.5 μV.

We included an intercept prior with a normal distribution, mean zero and SD of 1.5, such that there is a 95% probability that the intercept parameter lies between −3 and 3 μV. This decision was informed by intercept parameters in previous studies (Fleur et al., 2020; OT07; VB05), and appeared suitable for analyses time-locked to inflections, adjectives or nouns.

We did not include priors for the standard deviations of group-level ('random') effects, but used the corresponding default priors, which "are used (a) to be only very weakly informative in order to influence results as little as possible, while (b) providing at least some regularization to considerably improve convergence and sampling efficiency" (https://rdrr.io/cran/brms/man/set_prior.html; Bürkner, 2017). Likewise, we also did not include a prior for the standard deviation of the residual error. We did include a prior for the correlations of group-level ('random') effects using as the LKJ(2) prior (Bürkner, 2017; Lewandowsky, Kurowicka & Joe, 2009; for discussion, see; Vasishth, Beckman, Nicenboim, Li & Kang, 2018).

Analysis scripts with prior definitions are available on our OSF page. Normal(x,y) denotes a normal-distribution prior centered on x and with SD = y, here always in μV. Each model was run with at least 10.000 iterations (2000 warm-up) in at least 4 chains, which is advised for calculating Bayes Factors (e.g., Vasishth, Mertzen, et al., 2018; https://rdrr.io/cran/brms/man/bayes_factor.brmsfit.html; see also https://mvuorre.github.io/post/2017/bayes-factors-with-brms/).

From the Bayesian mixed-effects model, we calculated a Bayes Factor using the Savage–Dickey method (e.g., Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) to quantify the obtained evidence for the alternative hypothesis (H1) that there is a non-zero effect of prediction-match on inflection-elicited ERPs, or for the null hypothesis (H0) that there is no such effect. The Bayes Factor was calculated as the ratio between the posterior and prior distribution at an effect size of 0 μV, for each spatiotemporal ROI. Strength of the obtained evidence was interpreted following the convention of Jeffreys (1939).

To sum up, we performed 10 prediction-match replication tests (5 using inflection time-locked data, 5 involving the adjective-onset time-locked data), after having collected data from 100 participants who met the inclusion criteria. The required strength of the evidence for statistical inference was set at a Bayes Factor of at least 12, either for the alternative hypothesis or the null-hypothesis. This was double the evidence strength required by this journal, which we used to 'make up' for the lower prior density of our normal prior (i.e., a two-sided test) compared to a half-normal prior (i.e., a one-sided test). We took sufficiently strong evidence (Bayes Factor > 12) for the alternative hypothesis at any ROI selection as a successful replication in the sense that it demonstrates the use of inflection information. In that scenario, the observed effect was either positive or negative, which we would take as a replication of either VB05 or OT07, respectively. Based on VB05, we expected the effect polarity to be the same in the two time windows. If this was not the case, this could signal a problem with the inflection time-locked analysis, perhaps due to application of a baseline correction in an already unfolding effect (for example, because participants pick up on relevant acoustic differences between conditions before inflection onset, see discussion in VB05). The adjective-onset time-locked effect then receives priority in guiding our conclusions, since baseline differences are less likely to be a problem for that analysis. Importantly, we took sufficiently strong evidence for the null hypothesis at all these selections as a failure to replicate both VB05 and OT07.

Our Bayesian sampling plan was such that if none of the obtained Bayes Factors for the alternative hypothesis reached 12 or all obtained Bayes Factors for the null hypothesis stayed below 12, we followed the guidelines of this journal and tested additional participants until that evidence strength is reached (one Bayes Factor that sufficiently supports the alternative hypothesis, or all Bayes Factors sufficiently supporting the null hypothesis). Sample size was increased in steps of 20 participants, to be capped at a maximum of 200 participants for practical considerations. However, we were forced to halt testing before reaching that number because of a covid-19 pandemic related lockdown of our institute, and we therefore report analyses on a total sample size of only 187 participants.

After completion of data collection, we performed additional analyses with different priors for the effect size of 'match' to investigate the robustness of the obtained results [normal(0,1) and normal(0,5) to cover a wider/narrower range of plausible values].

2.8. Noun analyses

The noun analyses were performed once the data collection had stopped, and served as positive controls because N400 effects of predictable versus unpredictable words are highly common throughout the psycholinguistic literature (for review, see Kutas & Federmeier, 2011) and typically of a relatively large effect size compared to prenominal manipulations. If no effect is observed for the adjectives *and* no N400 effect of match is observed for the nouns, no valid inference about predictive processing can be drawn from the adjective data (see also DeLong et al., 2005; Nieuwland et al., 2018; VB05).

For the three noun-ROIs, we tested the following model: voltage ~ match + (match | subject) + (match | item). Similar to the adjective analyses, the prior on the estimated effect size had a normal distribution and a zero mean (although unexpected nouns were expected to elicit more negative voltage than expected nouns). The prior standard deviation of the match parameter was set at 2 μV, corresponding to the approximate effect sizes reported by VB05 and OT07. The intercept prior was the same as for the adjective analysis, as were the other priors (but no prior was included for 'gender').

Like the nouns in VB05 and OT07, there was no overlap between predictable and unpredictable nouns, which means that the noun-comparison was confounded by lexical variables (e.g., frequency) and contextual variables (e.g., plausibility) that are known to influence N400 amplitude (e.g., Kutas & Federmeier, 2011; Nieuwland et al., 2018). In addition, VB05, OT07, and this study used different nouns altogether. Despite these caveats about between-study differences, an additional model was tested with a stronger noun prior, to test whether the obtained N400 effect had changed the support for the specific noun effect-size reported by VB05. For this analysis, we used a normal prior for 'match' with a mean at −2.2 μV and a standard deviation of .50 μV, corresponding to the strongest effect reported in VB05 (at the left-posterior quadrant). This prior defines a 95% probability that the 'match' parameter lies between −1.2 and −3.2 μV.

2.9. A pre-registered exploratory analysis of adjective-gender effects

The factor 'gender' was included in the analyses because, in principle, it was considered a nuisance variable. Instead of only counterbalancing gender across items, our confirmatory analyses explicitly accounted for the associated variance when testing the effect of prediction-match (see also Sassenhagen & Alday, 2016). However, we further considered an effect of 'gender' in an exploratory analysis, because the effect of prediction-match could depend on gender (see also Loerts, Wieling, & Schmid, 2013). For example, the prediction effect could be greater when the predictable noun has common gender compared to neuter gender, perhaps because it is easier to detect a mismatch on an overt suffix than on the absence of a suffix, or perhaps because an overt suffix rules out the expected noun meaning, whereas the absence of the suffix does not rule out entirely the expected noun meaning (it could signal an upcoming diminutive noun irrespective of gender; see also Loerts et al., 2013, for relevant discussion). An alternative scenario is also possible, namely that the prediction-match effect is greater when the predictable noun has neuter gender as opposed to common gender, perhaps because people find it harder to detect a mismatch on an overt suffix. This could be related to the fact that language learners tend to add the suffix incorrectly more often than they omit it incorrectly; both young and old language learners tend to overgeneralize the suffix like in 'een moeilijke boek' (Weerman, Bisschop & Punt, 2006; for a review on Dutch adjectival inflection, see Van de Velde & Weerman, 2014). It is possible that such overgeneralizations by L2 learners change inflection processing even in L1 speakers, or that overgeneralizations in childhood continue to impact inflection processing later in life, even if only very subtly.

In both these hypothetical scenarios, one could expect to observe an interaction between 'match' and 'gender' on inflection-elicited ERPs. There was no strong a priori reason to assume that this interaction term would yield a strong effect, given that previous studies approximately balanced the 'gender' factor across items. Nevertheless, we considered this possible interaction effect and performed exploratory tests that included the interaction term as a fixed effect and included a by-subject random slope in the brms model.

voltage ~ match * gender + (match * gender | subject) + (match | item)

For this analysis, we add one prior to the brms model for the adjective and inflection analyses, namely a normal distribution prior with a mean zero and SD of 1 for the slope of the interaction parameter. This prior defines a 95% probability that the parameter for the interaction term, i.e., the difference in the match effect for common and neuter gender adjectives, lies between −2 and 2 μV.

2.10. Pre-registered traditional ANOVAs

At the request of the reviewers, we conducted repeated-measures ANOVAs that closely follow the analyses of VB05 and OT07[11]. However, we note that the primary basis for our conclusions is the Bayes mixed-effects analysis approach described previously. The ANOVA analyses were performed after data collection had ended, using the function 'aov_car'

[11] Using G*Power with 'SPSS standards' (Faul, Erdfelder, Lang, & Buchner, 2007), we established that our minimum sample size of N = 100 also yielded sufficient a priori power for these ANOVA analyses. For the lowest relevant F-value from VB05 ($F_{(1, 23)} = 5.16$), i.e., the most conservative estimate, the partial eta-squared measure of effect size was $\eta_p^2 = .183$ (see Lakens, 2013). The lowest F-value in OT07 was ($F_{(1, 28)} = 4.5$), which yielded $\eta_p^2 = .138$. To detect this latter, more conservative effect size with a power of .9 at an alpha level of .02, the required sample size is N = 86. This would be the required sample size if we used only about half of our items.

from the 'afex' R package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2019; in our pre-registration we planned on using JASP, 2018). Averaging over items, we analyzed mean amplitude values per condition per subject in the 50–250 msec time window after inflection onset (VB05), and in the 300–600 msec time window after adjective onset (OT07). For the N400 effects at the noun, we analyzed the 300–500 msec time window. We conducted repeated measures ANOVAs on the five ROIs defined above. In the midline region, Prediction-match (matching *vs* mismatching) was fully crossed with the five electrodes (Fz, FCz, Cz, Pz, Oz). The analysis at the four quadrants was carried out by crossing Prediction-match with Hemisphere (left *vs* right) and Anteriority (anterior *vs* posterior). In all analyses, Greenhouse-Geisser correction was applied to F-tests with more than one degree of freedom (Greenhouse & Geisser, 1959). We report the mean voltage and SD for each condition, the estimated difference between conditions with the 95% CI, and Cohen's d measure of effect size.

3. Results

Fig. 1 displays the grand-average ERPs at each ROI,[12] time-locked to the inflection onset of the gender-matching (solid blue line) and -mismatching (dotted red line) adjectives (for ERPs at individual channels, see Appendix Figure A.1). Both conditions showed a negative peak at anterior ROIs around 200 msec and a positive-going waveform in the first 500 msec at posterior ROIs. Corresponding to these ERP waveforms, the statistical analyses showed that mismatching inflections elicited more negative voltage (plotted upwards) than matching inflections within the 50–250 msec time window (Table 2), instead of the positive ERP effect reported by VB05. The corresponding scalp distribution of the effect in this time window is shown in Fig. 4.

Fig. 2 displays the equivalent grand-average ERPs time-locked to onset of the adjective (see also Fig. 4 and Appendix Figure A.2). Both conditions showed a negative peak at anterior channels around 200 msec after word onset, and a much broader negative peak around 300 msec at midline and posterior ROIs followed by a positive-going waveform. Here, too, there was no sign of a positive ERP effect, and mismatching adjectives elicited a negativity compared to matching adjectives, which was visible in the 300–600 msec and subsequent time window at the midline and posterior ROIs.

Fig. 3 displays the subsequent effects at the nouns (see also Fig. 4 and Appendix Figure A.3), which also showed a negative peak at anterior ROIs for both conditions, in addition to a clear N400 effect, namely enhanced negativity for the mismatch condition compared to the match condition, peaking around 300–500 msec after noun onset, with strongest effects at midline and posterior ROIs.

[12] Figures in this article were created using the following packages: "cowplot" (Wilke, 2019), "dplyr" (Wickham, François, Henry & Müller, 2019), "forcats" (Wickham, 2020), "ggplot2" (Wickham, 2016), "emmeans" (Lenth, 2019), "reshape2" (Wickham, 2007), "Rmisc" (Hope, 2013), "patchwork" (Pedersen, 2019), "stringr" (Wickham, 2019).

Table 2 lists the Bayesian estimates (b) for the mismatch effect (mismatch minus match) for each ROI, along with the credible interval (CrI) and the posterior probability of the effect being negative [$p(b) < 0$, the percentage of sample under zero]. Time-locked to inflection onset and to adjective onset, the mismatch estimate is negative at all ROIs and a majority of the posterior distribution falls under zero. This is also visible in Figs. 5 and 6, which show the results of the main Bayesian hypothesis tests at each ROI for the effects time-locked to inflection and adjective onset, respectively. The graphs show the prior distribution (light blue) and posterior distribution (dark blue), and highlights their corresponding values at a mean effect of zero (indicated by yellow and red dots, respectively), along with the Bayes Factor corresponding to their ratio. Each reported Bayes Factor (BF_{null}) was calculated as the posterior density at zero divided by the prior density at zero, such that values greater than one correspond to increase in evidential strength for the null hypothesis (e.g., a $BF_{null} = 3$ means that evidence for the null hypothesis has increased three-fold), whereas values smaller than one correspond to increased evidential strength against the null hypothesis. For effects time-locked to inflection, all Bayes Factors are greater than 1 but remained low, with only the right-anterior and -posterior ROIs showing values just over 3 ('moderate' evidence for the null-hypothesis) and the rest only yielding 'anecdotal' evidence. A somewhat similar pattern occurred for effects time-locked to adjective onset, with Bayes Factors just over 4 at the anterior ROIs, and only the left-posterior ROI yielding a value under 1.

The nouns elicited strong N400 effects at all three pre-registered ROIs, in fact, so strong that none of the posterior samples contained zero (Table 2, Fig. 7), yielding 'extreme' evidence against the null ($BF_{null} = 0$, although one could say it is more precise to state $BF_{null} < 1/32.000$, as our analysis would not be able to pick up values smaller than 1 divided by the number of posterior samples).

We also computed Bayes Factors with different priors to investigate the robustness of the obtained results (Table 3). For effects time-locked to the inflection or adjective onset, use of a wider prior ($SD = 1$) increased the obtained BF_{null} at each ROI, with half of the ROIs yielding moderate evidence for the null hypothesis and the other half yielding anecdotal evidence. With a narrower prior ($SD = .5$), the tests yielded anecdotal evidence. For comparison, we also performed exploratory analyses on ERPs time-locked to inflection with a prior mean and SD roughly based on VB05 ($M = .75$, $SD = .375$) and OT07 ($M = -.75$, $SD = .375$), reported in Appendix Table A.1. With this 'stronger' VB05 prior, we found strong evidence for the null hypothesis (BFs_{null} ranging from 12.7 to 22.7 for the 5 ROIs). With the OT07 priors, we found anecdotal/moderate evidence for the null hypothesis (BFs_{null} ranging from 2.1 to 6.8), which suggests that while the obtained effect is likely to be a negativity, it is probably much smaller than the effect reported by OT07.

For the noun effects, using a prior that corresponded to the strongest effect reported in VB05 did not impact the results because our posterior samples never included zero. Hence, even though the obtained N400 effects had mean estimates that were lower than those reported in VB05, our results nevertheless yielded extreme evidence against the null due to the high precision of our estimates.

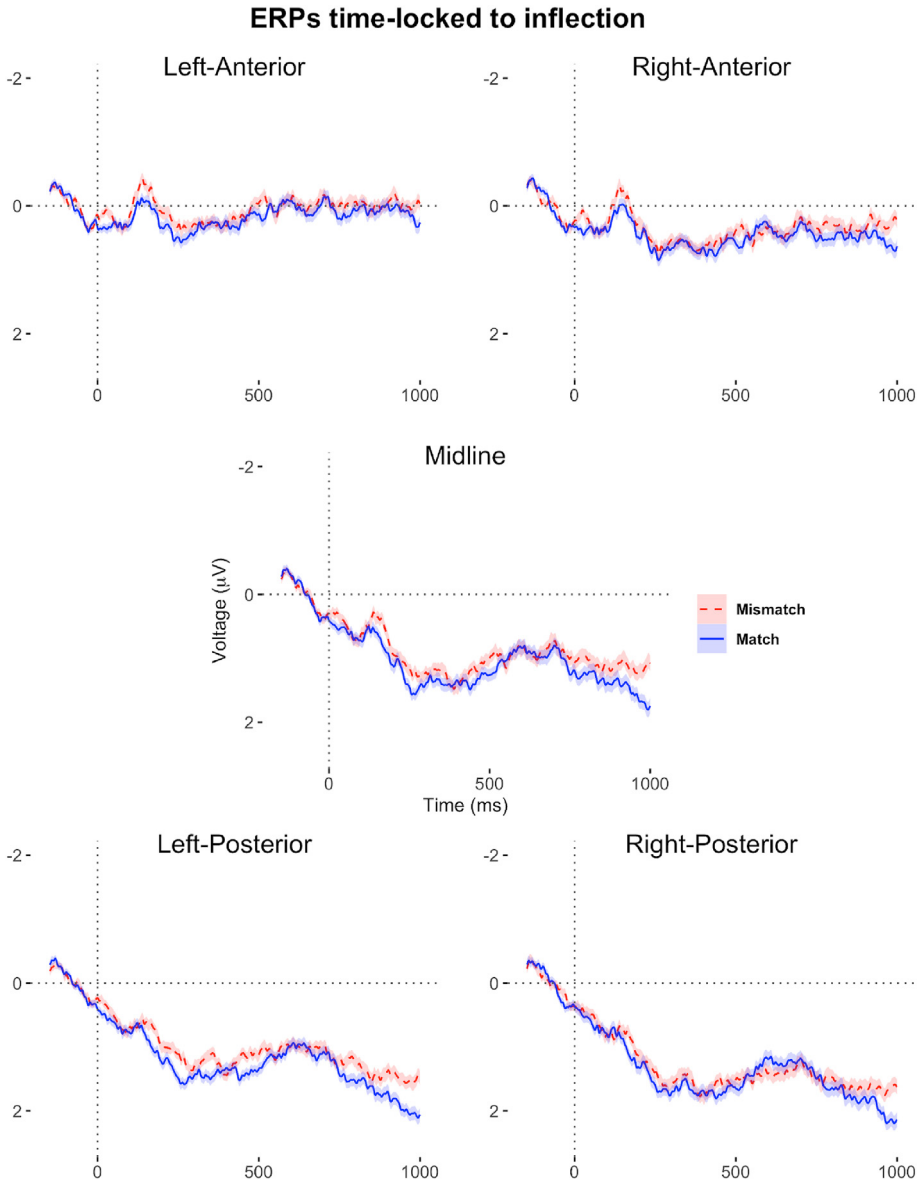

Fig. 1 – Inflection effects. The graphs show the grand-average ERPs elicited by gender-matching (solid blue lines) and gender-mismatching inflection (dotted red lines) at the 5 pre-registered ROIs. In these and following figures, color-shaded areas show the within-subject standard error of the condition mean (Cousineau, 2005; Morey, 2008 calculated with the 'Rmisc' package in R). We emphasize that these ERP plots do not directly correspond to the results of our statistical analyses, which account for variance associated with different items and with common/neuter gender.

Table 2 – Results of the pre-registered Bayesian analyses. Gender-mismatch effect at each ROI for ERPs time-locked to inflection onset, adjective onset and the nouns. Each cell gives the corresponding estimate (b) in μV for mismatch minus match, the credible interval (CrI), and the posterior probability of the effect being negative [*p*(b) < 0, the percentage of posterior samples under zero].

ROI	Inflection onset (50–250 msec)			Adjective onset (300–600 msec)			Nouns (300–500 msec)		
	b	CrI	*p*(b) <0	b	CrI	*p*(b) <0	b	CrI	*p*(b) <0
Left-Anterior	−.16	[−.36 .04]	94	−.10	[−.33 .14]	80			
Right-Anterior	−.13	[−.35 .10]	87	−.10	[−.34 .15]	79			
Midline	−.15	[−.37 .07]	91	−.19	[−.43 .06]	94	−1.68	[−2.02 1.33]	100
Left-Posterior	−.18	[−.36 .01]	97	−.21	[−.41 .02]	98	−1.47	[−1.77 1.18]	100
Right-Posterior	−.12	[−.32 .08]	88	−.21	[−.44 .02]	96	−1.73	[−2.05 1.41]	100

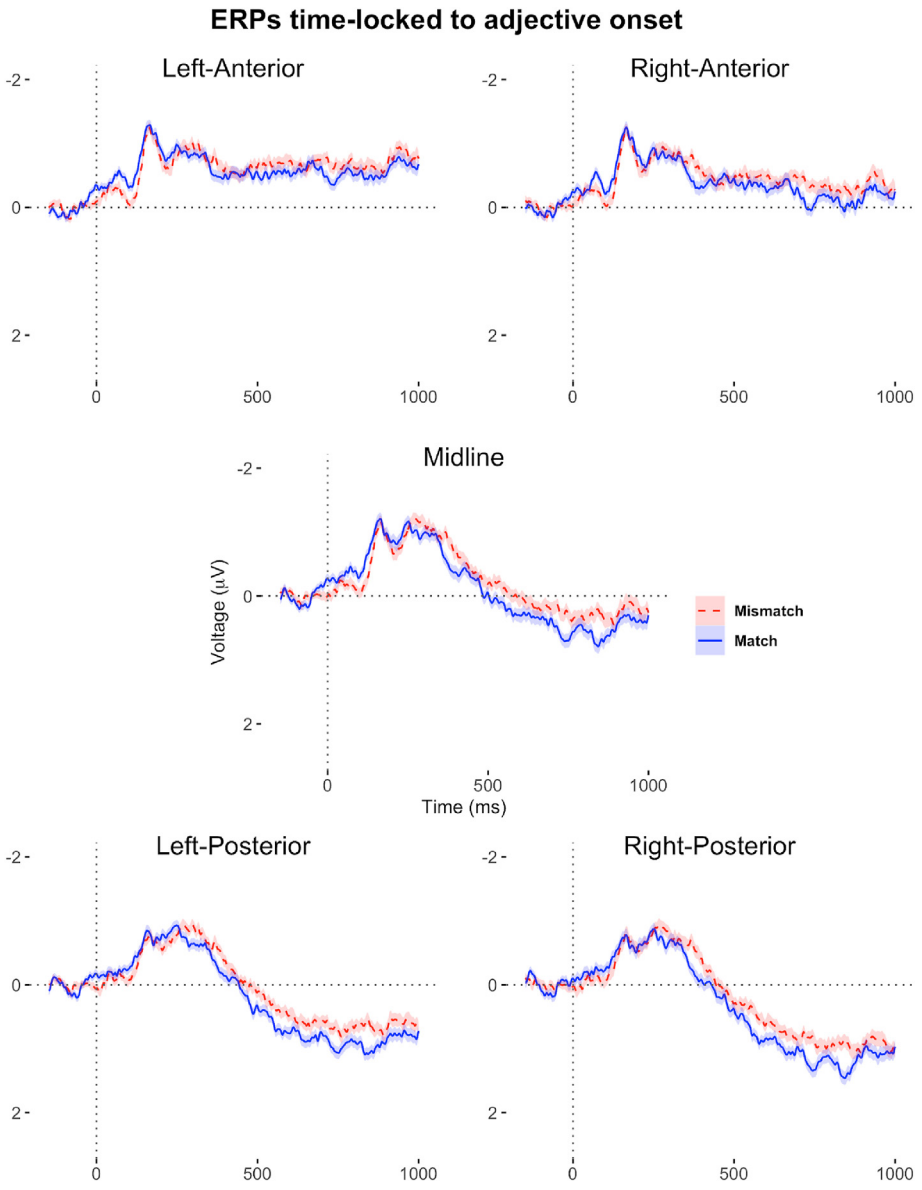

Fig. 2 – Adjective onset effects. The graphs show the grand-average ERPs elicited by gender-matching (solid blue lines) and gender-mismatching adjectives (dotted red lines) at the 5 pre-registered ROIs.

3.1. *Pre-registered exploratory analyses involving adjective-gender*

As shown in Fig. 8, mismatching inflection elicited a more pronounced negativity for common gender than for neuter gender, at least at frontal ROIs (see Appendix Figures A.4 and A.5 for effects at individual ERP channels, and see Fig. 4 for scalp distributions of the mismatch effects). At the left-anterior ROI, this negativity started as early as 0 msec and lasted for 1000 msec, whereas the other ROIs showed a negativity mostly in the 150–600 msec time window. For neuter nouns, the negativity was also visible at posterior ROIs, but anterior ROIs showed a positivity, at least from approximately 300 msec onwards.

The corresponding adjective onset effects are shown in Fig. 9 (see also Appendix Figures A.6 and A.7). For common gender adjectives, mismatch elicited an early positivity, which was most pronounced at frontal and midline ROIs. However, the early onset of this effect raises doubt that it is actually elicited by gender mismatch. For both common and neuter gender adjectives, mismatch elicited enhanced negativity. For neuter gender adjectives, this started at about 300–400 msec after onset and lasted until the end of the segmentation window. For common gender adjectives, the effect started a bit later at about 500 msec after onset.

The corresponding results for the match by gender interaction term are listed in Table 4. The interaction estimates and CrI correspond to [common gender (mismatch minus match) – neuter gender (mismatch minus match)] such that negative values show a greater negativity effect for common gender compared to neuter gender. These results only lend some support to the interaction pattern, albeit weak. In all analyses, the credible interval included zero and the BF_{null} only yielded anecdotal evidence, although the posterior probability of the

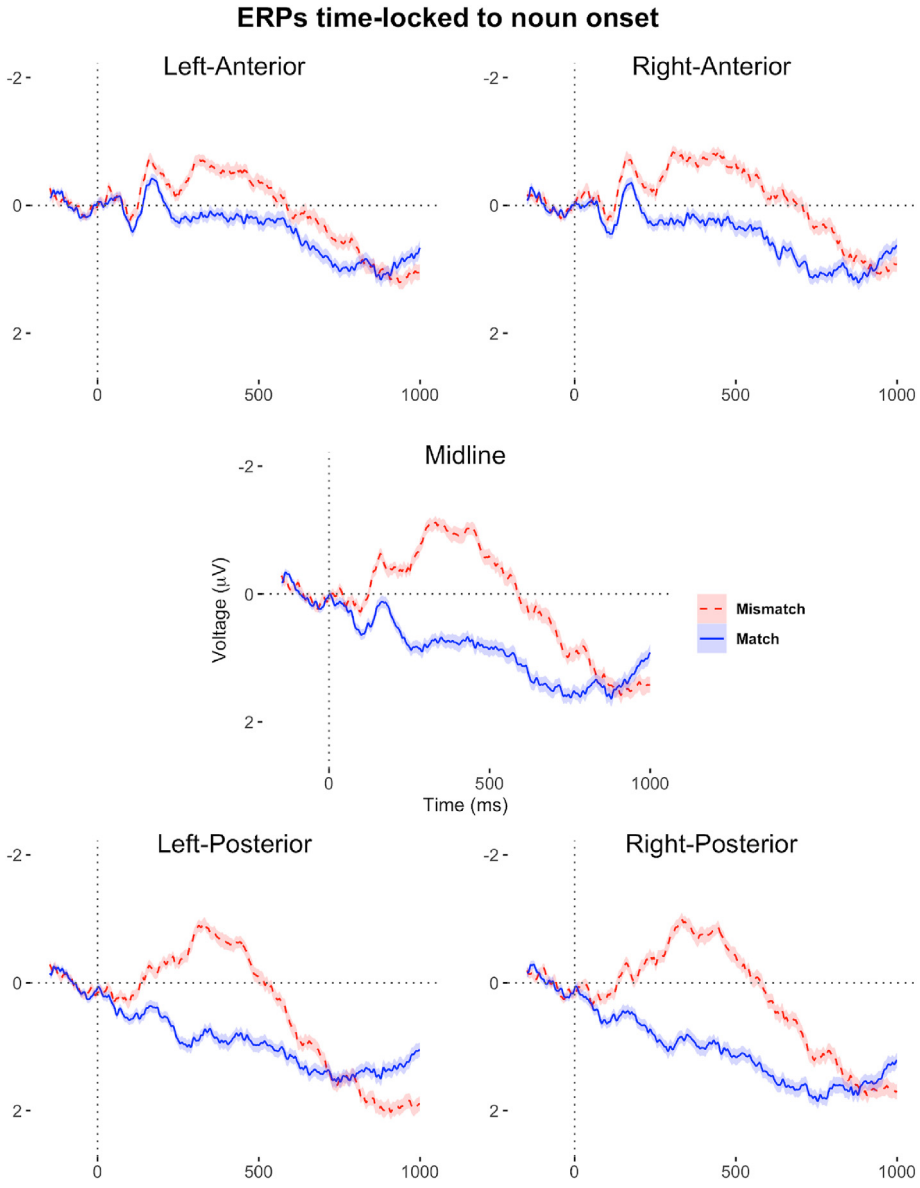

Fig. 3 – Noun effects. The graphs show the grand-average ERPs elicited by prediction-matching (solid blue lines) and -mismatching nouns (dotted red lines) at the 5 ROIs (only the midline and posterior ROIs were pre-registered for statistical analysis).

effect being negative was high for inflection-locked effects at anterior ROIs. Fig. 10 shows the pairwise mismatch effects for common and neuter nouns separately. Not reported in detail here, estimates for the mismatch effect were very similar to those from the models without interaction term.

3.2. *Pre-registered traditional ANOVAs*

For ERPs time-locked to inflection onset, repeated measures ANOVAs on the four quadrants revealed a marginally significant prediction-mismatch effect [$F(1,186) = 3.21$, $p = .074$, mean difference = −.14 μV, 95% CI = (−.01, .29), Cohen's $d = .168$], reflecting, on average, less positive voltage for the mismatching condition than for the matching condition (match, M = .62 μV, SD = 1.17; mismatch, M = .48 μV, SD = 1.05). None of the interactions between prediction-match and the two distributional factors (hemisphere and anteriority) yielded statistically significant effects (all Fs < 1.08, see Appendix Table A2). In the midline ROI, the prediction-match effect was not statistically significant [$F(1,186) = 2.31$, $p = .13$; mean difference = −.14 μV, 95% CI = (−.04, .32), Cohen's $d = .14$; match, M = .84 μV, SD = 1.26; mismatch, M = .7 μV, SD = 1.12].

For ERPs time-locked to adjective onset, the prediction-mismatch effect yielded a marginally significant result in the four quadrants [$F(1,186) = 3.79$, $p = .052$; mean difference = −.16 μV, 95% CI = (−.00, .36), Cohen's $d = .165$; match, M = −.19 μV, SD = 1.28; mismatch, M = −.36 μV, SD = 1.34]. None of the interactions between prediction-match and hemisphere and anteriority yielded statistically significant effects (all Fs < 2.45, see Appendix Table A2). Similarly, the prediction-match effect was marginally significant in the

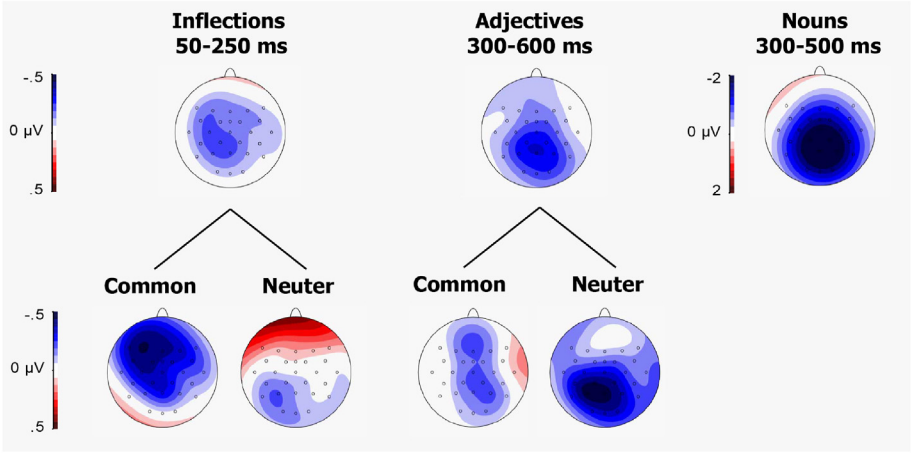

Fig. 4 – Scalp-distribution of the mismatch effects. The upper graphs show the scalp-distribution of the mismatch effect (mismatch minus match) for ERPs time-locked to onset of the inflections (left), adjectives (middle) and nouns (right), in the corresponding, pre-registered time-window used for statistical analysis. The bottom graphs show the mismatch effect for common and neuter gender separately.

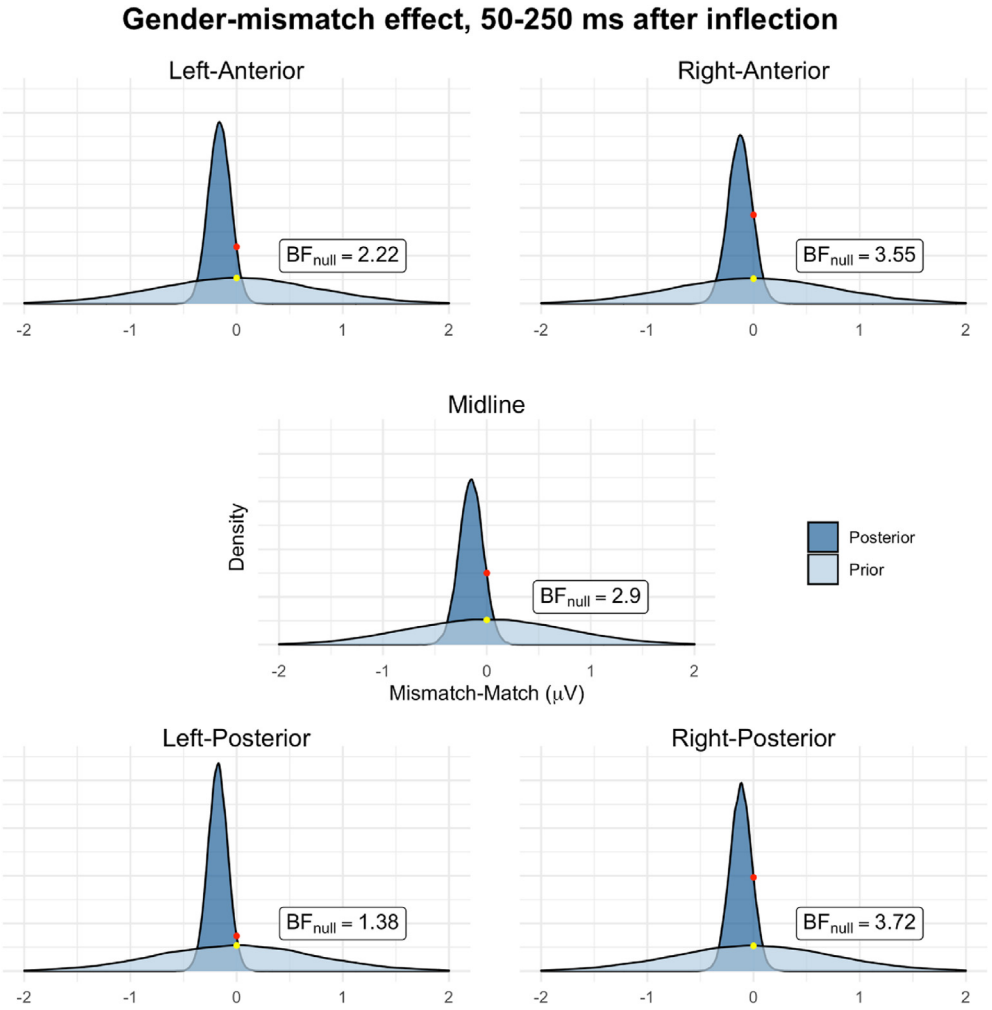

Fig. 5 – Results from the Bayesian hypothesis tests for the gender-mismatch effect time-locked to inflection onset. Graphs depict the prior (light blue) and posterior (dark blue) density, with prior and posterior density at zero marked by a yellow and red dot, respectively. The ratio of the density values at zero, the Bayes Factor, is labeled on each graph, here showing the Bayes Factor evidence in support of the null hypothesis (BF_{null}), with higher values corresponding to the increased belief in the null-hypothesis given our data.

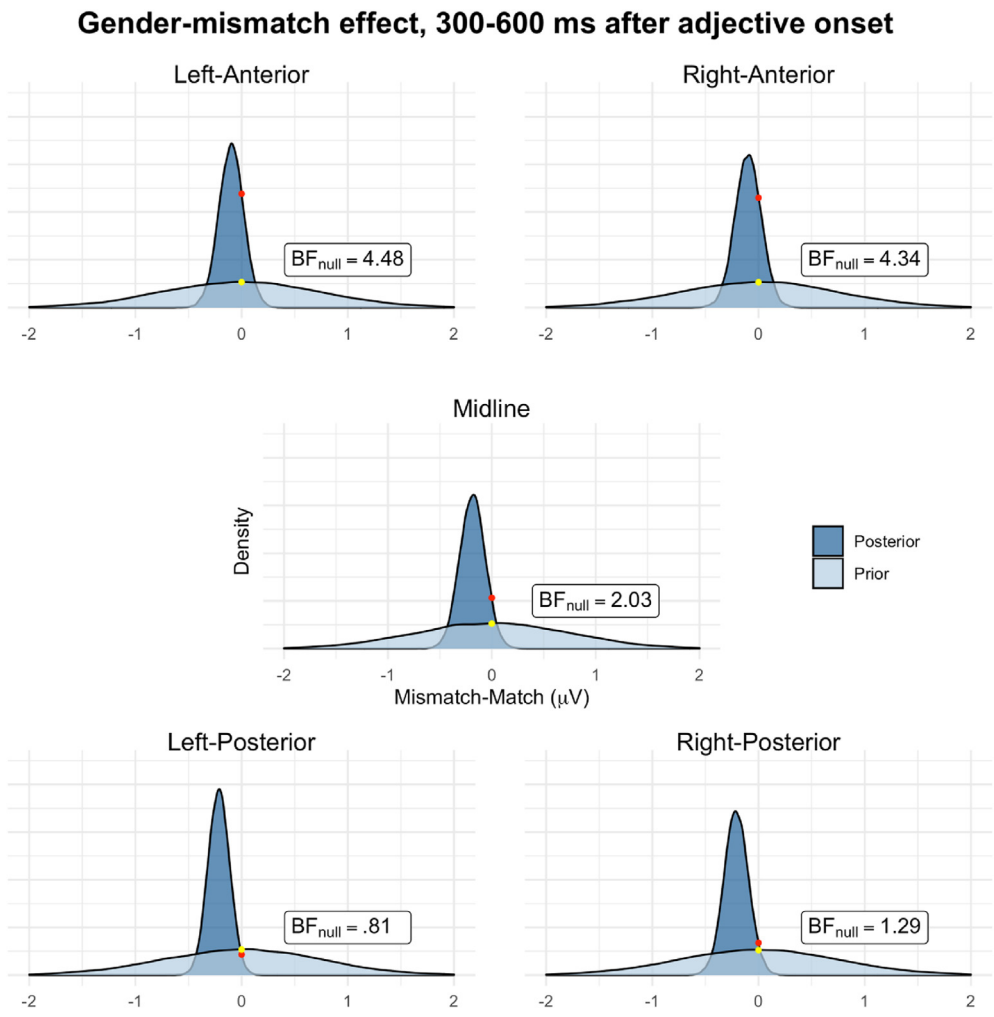

Fig. 6 – Results from the Bayesian hypothesis tests for the gender-mismatch effect time-locked to adjective onset.

midline region [F(1,186) = 3.56, p = .06; mean difference = −.20, 95% CI = (−.01, .41), Cohen's *d* = .168; match, M = −.26 μV, SD = 1.43; mismatch, M = −.46 μV, SD = 1.49].

For ERPs time-locked to noun onset, analyses in both ROIs showed a significant prediction-mismatch effect [F(1,186) = 221.34, p < .001; mean difference = −1.6 μV, 95% CI = (1.38, 1.81), Cohen's *d* = 1.42], with more negative amplitudes for mismatching nouns compared to matching nouns (mismatch, M = −.68 μV, SD = 1.27; match, M = .92 μV, SD = 1.4). This effect was also significant in the midline region [F(1,186) = 197.87, p < .001; mean difference = −1.66 μV, 95% CI = (1.53, 1.78), Cohen's *d* = 1.36; mismatch, M = −.97 μV, SD = 1.41; match, M = .73 μV, SD = 1.53].

3.3. *Exploratory mass regression analysis*

Our pre-registered analyses averaged activity from selected electrodes and time points within spatiotemporal ROIs based on VB05/OT07. To better characterize the effects of interest inside and outside of the ROIs, we performed exploratory mass regression analyses. First, we downsampled the pre-processed, segmented adjective onset and inflection data to 100 Hz (i.e., one sample for every 10 msec) to speed up the analysis. Then, we performed a mixed-effects model analysis using the 'lme4' package (Bates et al., 2014) for each channel, and for each sample between −150 and 500 msec relative to inflection onset (this shorter window minimized distortion from effects associated with noun onset) and between −150 and 1000 msec relative to adjective onset. We first tried analyses with the same fixed and random effects as in the pre-registered exploratory analysis, but because all models failed to converge, even after removing random correlations, we opted for simpler models, also to further speed up the analysis. We here report results from a model with the main effect and interaction between match and gender as fixed effects, a by-subjects random slope for match, and only a by-items random intercept (convergence failures occurred in approximately 10 percent of the inflection and adjective onset models). For each model, we extracted a coefficient estimate with a standard error, t-value and *p*-value associated with 'mismatch', 'gender' and the 'mismatch:gender' interaction term. For ERPs time-locked to inflection (Appendix Figure A.8), the mismatch effect appeared strongest around 200–300 msec at left-posterior channels. Parietal and occipital channels show prominent effect fluctuations of approximately 10 Hz, suggesting sensitivity of the mismatch effect to alpha

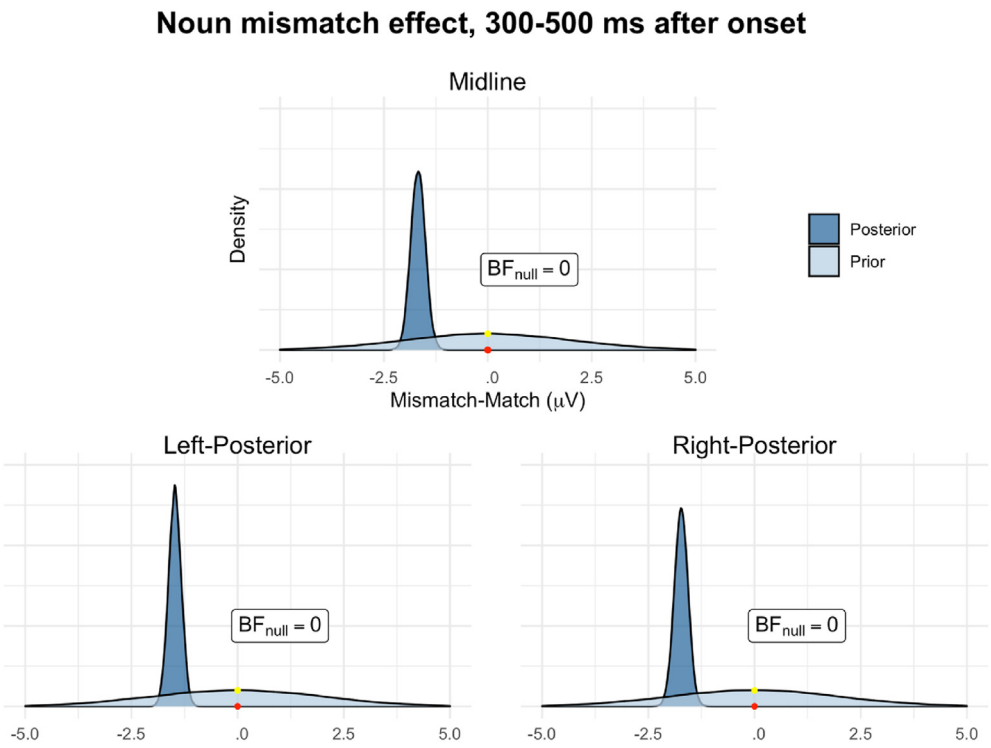

Fig. 7 – Results from the Bayesian hypothesis tests for the noun mismatch effects at the three pre-registered ROIs. Although the graphs highlight posterior density at zero with a red dot, the posterior samples did not contain the value zero, which is why BF_{null} is labeled as zero.

Table 3 – Bayes Factor results (BFs_{null}) from analyses with different priors. For effects time-locked to inflection or adjective onset, the new prior for the standard deviation for the mismatch effect was either wider (1) or narrower (.5) than in the main analyses. For noun effects, the prior corresponded to the strongest effect reported in VB05.

ROI	Inflection onset prior		Adjective onset prior		Noun prior
	SD = 1	SD = .5	SD = 1	SD = .5	Mean = −2.2 SD = .5
Left-Anterior	2.89	1.56	6.02	3.06	
Right-Anterior	4.41	2.47	5.84	3.05	
Midline	3.48	1.86	2.45	1.35	0
Left-Posterior	1.81	1.00	1.05	.58	0
Right-Posterior	4.72	2.52	1.69	.95	0

fluctuations (see also VB05; Nieuwland, 2019, for discussion on 'residual alpha' effects in experiments on spoken language comprehension). For ERPs time-locked to adjective onset (Appendix Figure A.9), the mismatch effect appeared strongest between 500 and 800 msec at posterior channels. Here too, the effect fluctuated at an alpha-range frequency, especially between 700 and 1000 msec after onset. Corresponding results for the interaction terms are available on our OSF page.

Taken together, these results suggest that the effect of gender-mismatch was strongest towards the end or after the pre-registered time windows, and had a posterior scalp distribution. However, we emphasize that the results of these exploratory analyses only have a descriptive purpose. Using a rather conservative method to control the false discovery rate that does not take into account spatio-temporal contingencies in the data (Benjamini & Hochberg, 1995), the tested samples did not survive correction for multiple comparisons.

4. Discussion

We performed a pre-registered, close replication of Van Berkum et al. (2005, Experiment 1), a canonical ERP study on lexical prediction during spoken discourse comprehension. In the original study, the marking of grammatical gender on prenominal adjectives ('groot/grote') elicited an early positivity when it mismatched the gender of an unseen, highly predictable noun, compared to matching gender. In our large-scale (N = 187) replication effort, we did not obtain this pattern of effects, but, if anything, a reverse pattern: mismatching gender elicited enhanced negativity compared to matching gender, reminiscent of the effects reported by Otten et al. (2007). We observed enhanced negativity at all spatio-temporal ROIs, whether time-locked to onset of the inflection or the adjective. However, this enhanced negativity was generally very small (approximately between −.15 and −.20 μV at the different ROIs), and our Bayes Factor hypothesis tests either anecdotally or moderately favored the null hypothesis. In contrast, we successfully replicated VB05's prediction-mismatch N400 effect for the nouns, observing extreme evidence against the null hypothesis even when our prior

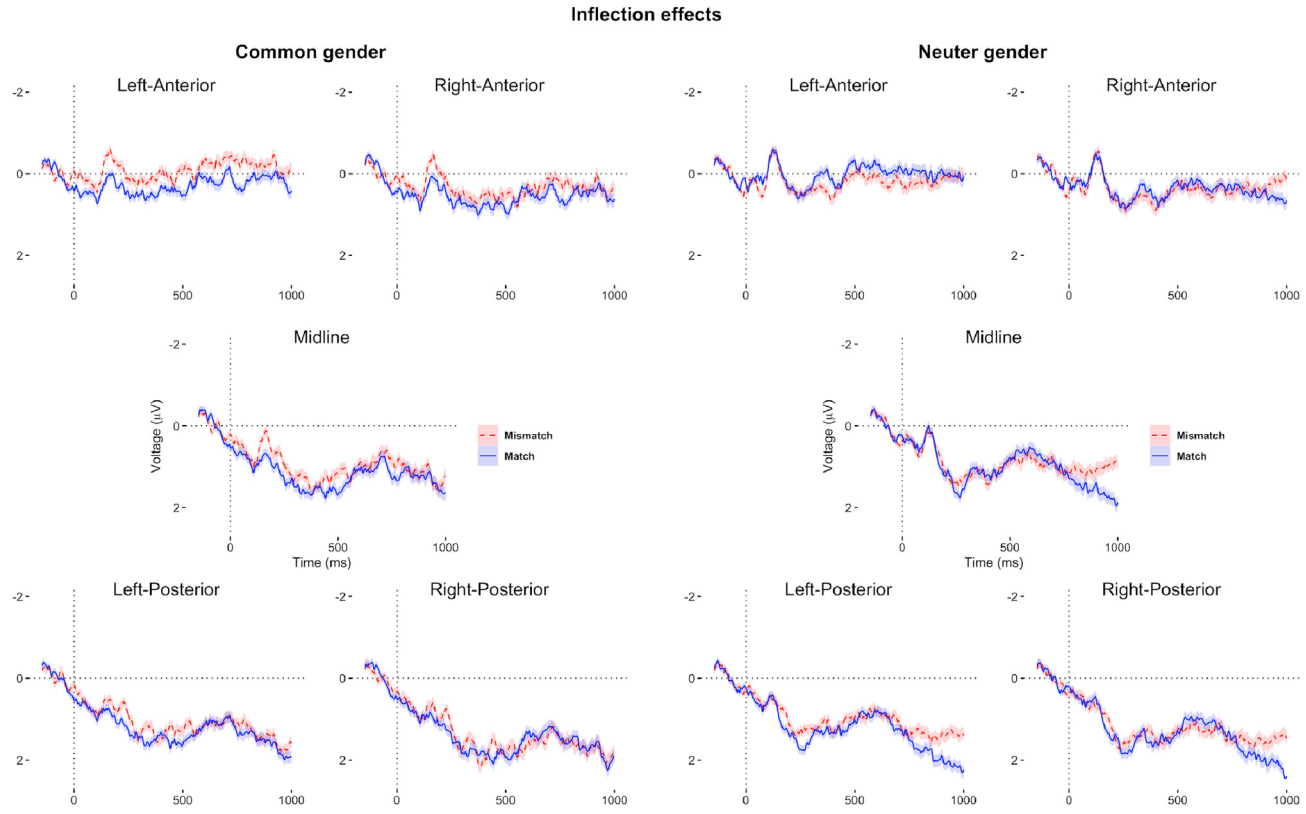

Fig. 8 – Inflection effects for common and neuter gender. The graphs show the grand-average ERPs elicited by common gender inflection and neuter gender inflection (with and without suffix, respectively, e.g., 'groot' and 'grote') that matched (solid blue lines) or mismatched (dotted red lines) the gender of the predictable noun.

corresponded to the strongest noun-elicited effect reported in VB05.

Pre-registered exploratory analyses showed that, at the anterior and midline ROIs, the negativity obtained in the inflection time-locked analysis was primarily generated by common gender adjectives ('grote') and close to zero for neuter gender adjectives ('groot'). However, like the main effect of gender-mismatch, the observed gender by mismatch interaction effect was weak and not supported by our Bayes Factor tests. Further exploratory analyses suggested that the main effect of gender-mismatch was most pronounced at posterior electrodes, where it was similar for common and neuter gender, and strongest near the end or even after the pre-registered time windows.

Taken together, these results do not support the effect reported by VB05. However, the results did not yield clear evidence against lexical prediction more generally, and in fact yielded some evidence in support of prediction. In the sections below, we discuss our results in more detail and briefly consider their implications for theory and research on predictive language comprehension.

4.1. *Weighing the evidence: Bayes Factors versus estimation*

Interpreting our pre-nominal results is not entirely straightforward because different sources of evidence point in different directions. As our primary and pre-registered source, the obtained Bayes Factors weakly favored the null hypothesis. From the 10 BFs_{null} (quantifying support for the null-hypothesis at each of the 5 ROIs, time-locked to inflection or adjective), 9 were over 1, and 4 were over 3 ('moderate evidence'). However, these values were generally low, and nowhere near the pre-registered threshold (BF = 12) that would have allowed us to halt sampling and claim replication success or failure. At the same time, the gender mismatch effect estimates themselves were clearly suggestive of a negativity. This was evident from the posterior probabilities of the effect being negative, which were consistently higher than 78% across all pre-registered tests, and from the exploratory analyses.

This discrepancy is primarily caused by the prior, which influences the Bayes Factor much more strongly than the estimate, as also demonstrated by our results obtained with varying priors. With pre-registered, widened and narrowed zero-mean priors, Bayes Factor support for the null hypothesis increased and decreased, respectively, without noticeable effect on the obtained estimates. With exploratory priors centered on the estimates reported by VB05, we obtained strong Bayes Factor support for the null hypothesis while the estimate became less negative by only about .05 μV. For this reason, it is generally advisable to pre-register a range of informative and plausible priors. The influence of the prior can be considered either a bug or a feature of Bayesian null-hypothesis testing, depending on your perspective (for discussion, see Kruschke, 2011; Kruschke & Liddell, 2018; Rouder,

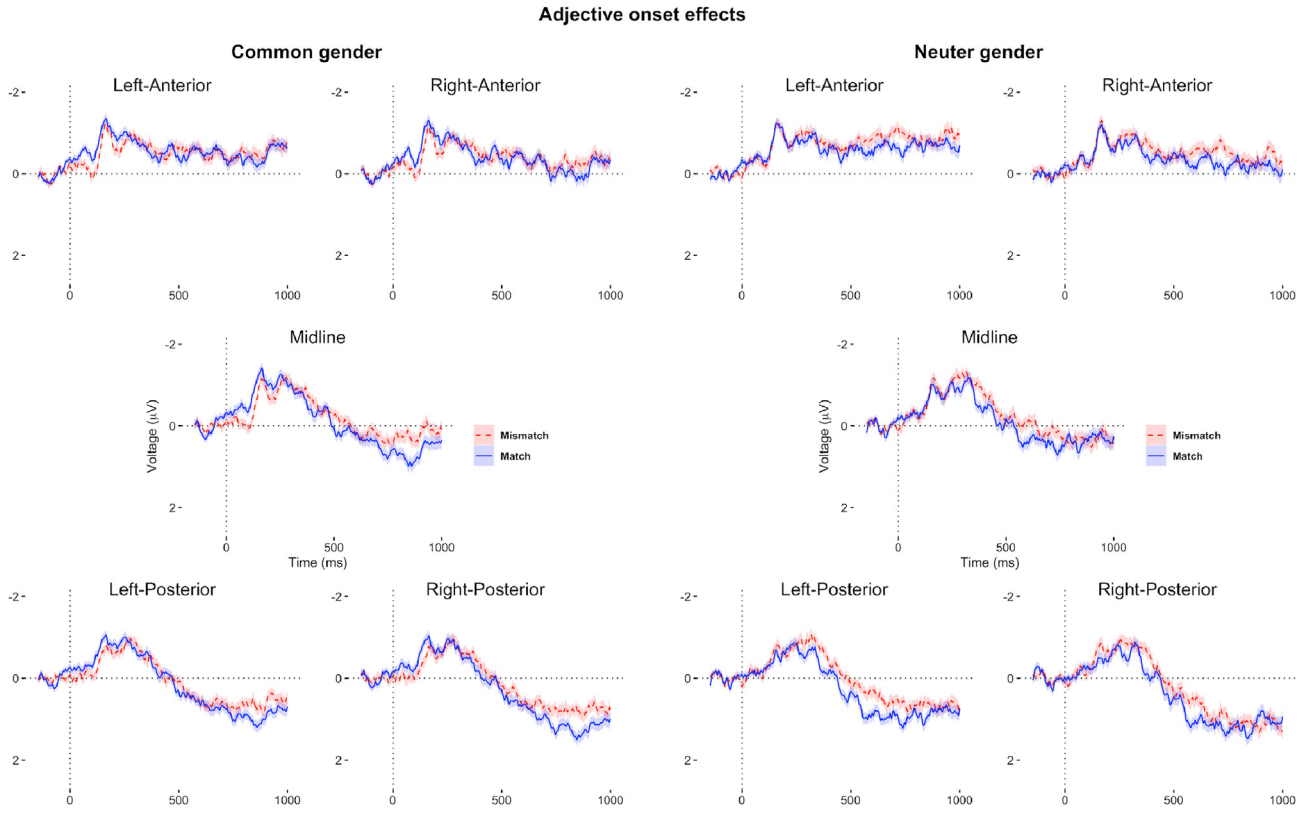

Fig. 9 – Adjective onset effects for common and neuter gender. The graphs show the grand-average ERPs elicited by the onset of common gender adjectives (with suffix, e.g., 'grote') and neuter gender adjectives (without suffix, e.g., 'groot') that matched (solid blue lines) or mismatched (dotted red lines) the gender of the predictable noun.

Table 4 – Results from the pre-registered exploratory analysis of the interaction between gender and gender-mismatch. Each cell gives the corresponding estimate (b) in μV for the interaction term (negative values correspond to a more negative mismatch effect for common gender than for neuter gender), the associated credible interval (CrI), and the posterior probability of the effect being negative [$p(b) < 0$, the percentage of posterior samples under zero], and the BF_{null}.

ROI	Inflection onset				Adjective onset			
	b	CrI	$p(b) < 0$	BF_{null}	b	CrI	$p(b) < 0$	BF_{null}
Left-Anterior	−.43	[−.99 .11]	94	1.10	−.18	[−.87 .51]	70	2.44
Right-Anterior	−.32	[−.89 .26]	85	1.80	−.03	[−.76 .68]	54	2.69
Midline	−.24	[−.91 .42]	55	2.25	−.09	[−.91 .73]	59	2.37
Left-Posterior	−.03	[−.69 .62]	53	3.03	.01	[−.75 .76]	49	2.71
Right-Posterior	−.10	[−.75 .55]	61	2.90	.10	[−.71 .91]	41	2.33

Haaf, & Vandekerckhove, 2018; van Ravenzwaaij & Wagenmakers, 2019). Weighing the two sources of evidence is ultimately a judgment call.

Despite insufficient Bayes Factor support to claim a replication failure, we are fairly confident that our results do not replicate VB05's positivity. However, we are much less confident regarding a replication of OT07's negativity, because our effect appears quantitatively and qualitatively different. Our effect was approximately only one-third of the OT07 effect size. Whereas the OT07 effect had a clear right-anterior maximum, our effect was not particularly lateralized and most prominent at posterior channels (and therefore not unlike an N400 effect in terms of scalp distribution and timing). Nevertheless, and despite the Bayes Factor evidence supporting the null hypothesis, our results do suggest that if a true population-level effect exists at all, it is likely small and negative.

4.2. *Differences between our study and VB05/OT07*

Although defining a close or exact replication remains controversial (e.g., Simons, 2014; Zwaan, Etz, Lucas, & Donnellan, 2018), we consider our study to be a close replication of VB05 and OT07, not an exact one. Readers might therefore be tempted to attribute the difference in results to a difference in methods. While influences of methodological differences cannot be ruled out, we consider it unlikely that they are the primary cause of the different results. For example, we used a different and larger set of prediction-inducing stories, but our stories were constructed in the

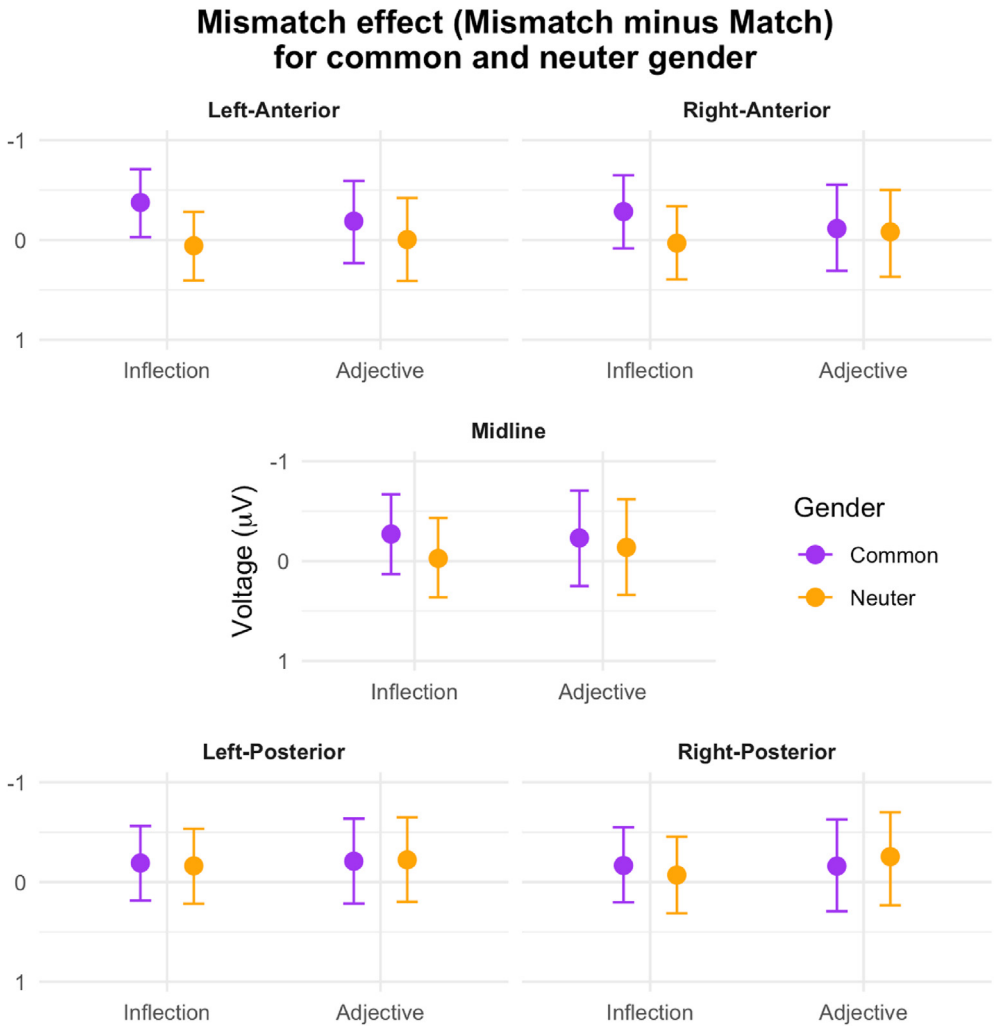

Fig. 10 – Results from the pre-registered exploratory analyses. The graphs show the mismatch effects (mismatch minus match) at each ROI, time-locked to inflection and adjective onset, for common gender (purple) and neuter gender (orange). Dots represent the marginal mean, whiskers represent the 95% credible interval.

same way as those of VB05/OT07, and had an equally strong as, if not stronger cloze probability manipulation than the originals. Moreover, our items were based on a set of items that have twice demonstrated a prenominal prediction effect on gender-marked articles ('de/het') with a much smaller sample size (N = 48 and N = 80; Fleur et al., 2020). We also used different filler items, but retained a similar experimental to filler ratio as VB05/OT07. We used a different speaker, but this speaker was not faster than that of VB05. Differences between our study and VB05/OT07 are detailed and justified in the Methods section, and also summarized in the online Supplementary Table 1. We also summarize differences between our study and our pre-registration in Supplementary Table 2.

Perhaps the strongest argument against the role of methodological differences does not involve a comparison between our study and VB05/OT07, but between VB05 and OT07. These studies were highly similar to each other, but nevertheless yielded two different types of effects. This discrepancy has previously been discussed in terms of as-yet-unidentified differences. For example, when discussing VB05, OT07 and other studies, Otten and Van Berkum (2009, p.96) note that "a systematic inventarization across all studies shows that this variability cannot be accounted for by differences in language, stimulus modality, type of prediction probe, or differences in working memory capacity of participants. One possibility is that perhaps the broader context in which stimuli are presented (i.e., the type of filler that is used, the length of the experiment) matters more than commonly assumed, but we refrain from speculating about specific other factors that could critically influence the way people make predictions, or process prediction-inconsistent data". Otten and van Berkum thus discussed the discrepant effects as two meaningful demonstrations of lexical prediction (i.e., as two 'true' effects), as is typical for the broader psycholinguistic literature (e.g., Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Kutas et al., 2011; Pickering & Gambi, 2018).

The current study, however, was premised on the assumption that only one of the original effects can be a 'true' effect, and that the other effect therefore is likely a false positive. False positives and wrong-sign estimates are to be expected in noisy, small-sample settings (Gelman & Carlin,

2014), especially when analysis choices are contingent on the data (e.g., based on visual inspection of ERP waveforms, as was the case in VB05 and OT07). Our results suggest that the positivity reported by VB05 is more likely to be a false positive finding than the negativity reported by OT07.

4.3. The role of adjective-gender

We anticipated a potential role of adjective-gender and the concomitant inflection in shaping the neural response to gender-mismatch. We considered two potential scenarios. The mismatch effect could be greater for common gender than for neuter gender, either because people find it easier to detect mismatch on overt suffixes than on absent suffixes, or because overt suffixes have a bigger impact than absent ones because only overt suffixes rule out the expected noun entirely. Alternatively, the mismatch effect could be greater for neuter gender than for common gender, possibly because detection of a mismatch on overt suffixes is more difficult for language developmental reasons (e.g., Weerman et al., 2006).

The results from our pre-registered exploratory analyses did not conclusively favor one scenario over the other. ERPs time-locked to inflection suggested a stronger mismatch effect at the left-anterior ROI for common gender than for neuter gender. However, this pattern was very weak and its significance remains unclear. One obstacle to interpretation is the early positive ERP effect that was visible immediately after adjective onset and therefore not elicited by the inflections (whose onset occurred at least 200 msec after adjective onset). This positive effect may have shown up as a negativity when we time-locked to inflection onset, as an artefact introduced by the baselining procedure.[13] This brings us to two general caveats. First, our design was not optimized for this interaction analysis. Because predictable common and neuter gender nouns were preceded by different contexts and adjectives, the mismatch effects for each gender involve a comparison between different sets of adjectives. How this impacted the results is not known, but it could have generated patterns such as the early positivity for common gender adjectives. Second, while our sample size is much larger than in typical ERP experiments on language comprehension, it was also not optimized (and probably too small) to reliably detect a gender by mismatch interaction.

One additional relevant observation pertains to the left-posterior ROI. From all 5 ROIs, the gender-mismatch effect there was strongest, whereas the interaction effect there was weakest and near zero, reflecting similar gender-mismatch effects for common and neuter gender. We conclude, therefore, that our results are most consistent with a gender-mismatch effect for common and neuter gender adjectives.

4.4. Implications for predictive processing

While our gender-mismatch effect may appear surprisingly weak, a weaker effect than those reported by VB05 and OT07 was to be expected. The original effects were both just statistically significant at the $\alpha = .05$ level. In small-sample, noisy data sets, such effects already tend to have an overestimated effect size and increased chance of a wrong sign (Gelman & Carlin, 2014). Moreover, their effects were based on data selected via visual inspection, a procedure that further overestimates the effect size. In the current study, weakness of the observed effect was partly the result of the pre-registered time windows based on VB05/OT07; the effect appeared strongest towards the end of or even after the pre-registered time windows.

Beyond the comparison to VB05/OT07, the pre-nominal prediction effect on Dutch adjectival inflection may be generally smaller than other pre-nominal prediction effects for several potential reasons. One reason is the unexpectedness of the gender-marked adjectives themselves, as suggested by recent results from our laboratory on written language comprehension (Fleur et al., 2020). In Fleur et al., gender-mismatching definite articles elicited enhanced negativity in the N400 time window compared to matching articles when the context presumably led participants to expect a particular, gender-marked article-noun combination (e.g., 'de' when they expected 'het boek'). This effect was found in two identical experiments with pre-registered analyses and much smaller sizes (N = 48 and N = 80) than the current one. However, when participants expected an indefinite article-noun combination that lacks gender-marking (e.g., 'een boek'), there was only a small gender-mismatch effect on definite articles in the N400 time window. In other words, gender-mismatch effects may be relatively small, and therefore harder to detect, when participants do not expect a gender-marked word in the first place, as may have been the case in our study. That said, there is no principled reason why prediction-effects cannot be obtained on adjectives at all, even when they are unexpected. A highly predictive language comprehension system should be able to make do.[14]

Another reason, one that we find more plausible, could be the difficulty with detecting mismatch on fleeting, relatively subtle information in the spoken signal. We emphasize that we do not claim people predict less when listening than when reading. However, our manipulation on word-final inflections is arguably more subtle than a comparison between two entirely different words (e.g., 'de/het', 'el/la'), because our participants needed to distinguish between a schwa sound or an inter-word 'silent' period. This relatively small acoustic/phonetic difference might be hard to discern (e.g., Bailey & Hahn, 2005), and is sometimes further distorted by coarticulation effects (influences on pronunciation associated with preceding or subsequent sounds).[15] We would expect a large

[13] A positive ERP effect in the baseline window can show up after baselining as a negative ERP effect starting as early as 0 msec.

[14] Unless perhaps the meaning of the adjective is incompatible with the predicted noun or changes the noun prediction. In the current study, VB05 and OT07, the critical adjectives were selected for being semantically compatible or congruent with the high-cloze noun, and it is assumed that the meaning of the adjective does not change the noun prediction. Whether gender-mismatch effects can be obtained on semantically incongruent adjectives (e.g., 'blue' if the predicted noun is 'banana') is an open question, but we think this is unlikely.

[15] In some items, coarticulation might make the conditions phonemically more dissimilar (e.g., /d/ sounds different in 'verkleed' *vs* 'verklede', see also VB05 and our Methods section).

pre-nominal prediction effect when the mismatching condition differs more strongly acoustically from the matching condition (e.g., a spoken version of the 'de/het' manipulation). People typically need only one or two phonemes to detect a deviation from a predicted noun (e.g., Van Berkum et al., 2005; Van Petten et al., 1999). This was also demonstrated by our noun results; prediction-mismatching nouns elicited strong N400 effects starting as early as 100 msec after noun onset.

The weak nature of our pre-nominal prediction effect should not be taken as evidence against lexical prediction more generally. It does raise the question, however, whether listeners reliably or consistently use adjectival inflection information to inform their noun predictions. When a misprediction is evident, people may use the available gender information to revise their initial noun prediction, and perhaps even change their initial prediction to a new noun (as demonstrated by concomitant effects on noun-elicited N400s, e.g., Fleur et al., 2020; Szewczyk & Wodniecka, 2020). However, when evidence for misprediction is less compelling or ambiguous, people might be 'reluctant' to let go of their initial noun prediction (e.g., Nieuwland et al., 2018). Such reluctance could make sense because our comprehension system must deal with or compensate for coarticulation effects, disfluencies and noisy real-world environments (e.g., Corley & Stewart, 2008; Mattys, Davis, Bradlow, & Scott, 2012; Norris, McQueen, & Cutler, 2016). Future research efforts should elucidate which pre-nominal manipulations elicit more reliable spoken language prediction effects than others. Especially when combined with computational modeling (e.g., Norris et al., 2016), such efforts can reveal the speech processing mechanisms involved in evaluating discourse-based lexical predictions.

Furthermore, it remains to be established whether the adjectival inflection manipulation has different effects on predictive processing during reading and listening (e.g., Otten et al., 2007; Otten & Van Berkum, 2008). In VB05's self-paced reading experiment (Experiment 3), readers slowed down upon encountering gender-mismatching adjectives compared to matching adjectives. However, this effect did not occur at the first of two gender-mismatching adjectives but on the second one appearing 3 words downstream (e.g., 'onopvallende' in 'grote maar nogal onopvallende', English translation: 'unobtrusive' in 'big but rather unobtrusive'). In a written language version of OT07, Otten and Van Berkum (2008) observed enhanced negativity for gender-mismatching adjectives compared to matching adjectives, but this effect occurred as late as 900–1200 msec after word onset. In sum, while gender-mismatching adjectives elicited rather weak effects in the current spoken language study, their effects during reading may be even weaker or less consistent.

5. Conclusion

We performed a close replication of one of the best-cited ERP studies on word anticipation (Van Berkum et al., 2005; Experiment 1), in which participants listened to Dutch spoken mini-stories. In the original study, the marking of grammatical gender on pre-nominal adjectives ('groot/grote') elicited an early positivity when mismatching the gender of an unseen, highly predictable noun, compared to matching gender. In our large-scale, pre-registered replication effort, we did not obtain such a positivity, but found enhanced negativity instead. However, this negativity was small and our pre-registered Bayes Factor analyses generally favored the null-hypothesis. Although reminiscent of the right-anterior negativity reported in a similar study by Otten et al. (2007), the current negativity was much smaller and had a posterior scalp distribution. Our results highlight the risks associated with data-contingent analysis. Given that data-contingent analysis has been and still is common in the psycholinguistic literature, especially in EEG research, some key findings in this literature may prove hard to replicate (e.g., Nieuwland et al., 2018; Nieuwland, 2019).

The weak nature of our pre-nominal prediction effect should not be taken as evidence against lexical prediction more generally. Recent work from our laboratory, for example, observed strong pre-nominal prediction effects on gender-marked articles during reading (Fleur et al., 2020), with pre-registered analyses but smaller sample sizes. The weak nature of the current effect may reflect the difficulty in detecting gender-mismatch from fleeting, relatively subtle information in the spoken signal. Our results therefore raise the question whether Dutch listeners reliably or consistently use adjectival inflection information to inform their noun predictions.

Author credits

Mante S. Nieuwland: Conceptualization, Resources, Software, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration.

Yana Arkhipova: Resources, Software, Formal analysis, Writing - Review & Editing, Visualization, Project administration.

Pablo Rodríguez-Gómez: Resources, Software, Formal analysis, Writing - Review & Editing, Project administration.

Open practices

The study in this article earned Open Materials, Open Data and Preregistered badges for transparent practices. Materials and data for the study are available at https://osf.io/jqhpz/.

Acknowledgements

For help with creating the stimulus materials and/or data collection, we thank Birgit Knudsen, Damien Fleur, Joost Rommers, Monique Flecken, Iris Schmits, Maarten van den Heuvel, and Jiska Koemans. For helpful suggestions regarding power analysis and/or Bayesian analysis, we thank Phillip Alday, Bruno Nicenboim, Shravan Vasishth, Christopher Chambers and Zoltan Dienes. For helpful comments on previous drafts, we thank Jona Sassenhagen, Steve Luck, Jos van Berkum and one anonymous reviewer. We thank Jos van Berkum for providing the original SPSS files from VB05.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2020.09.007.

Appendix

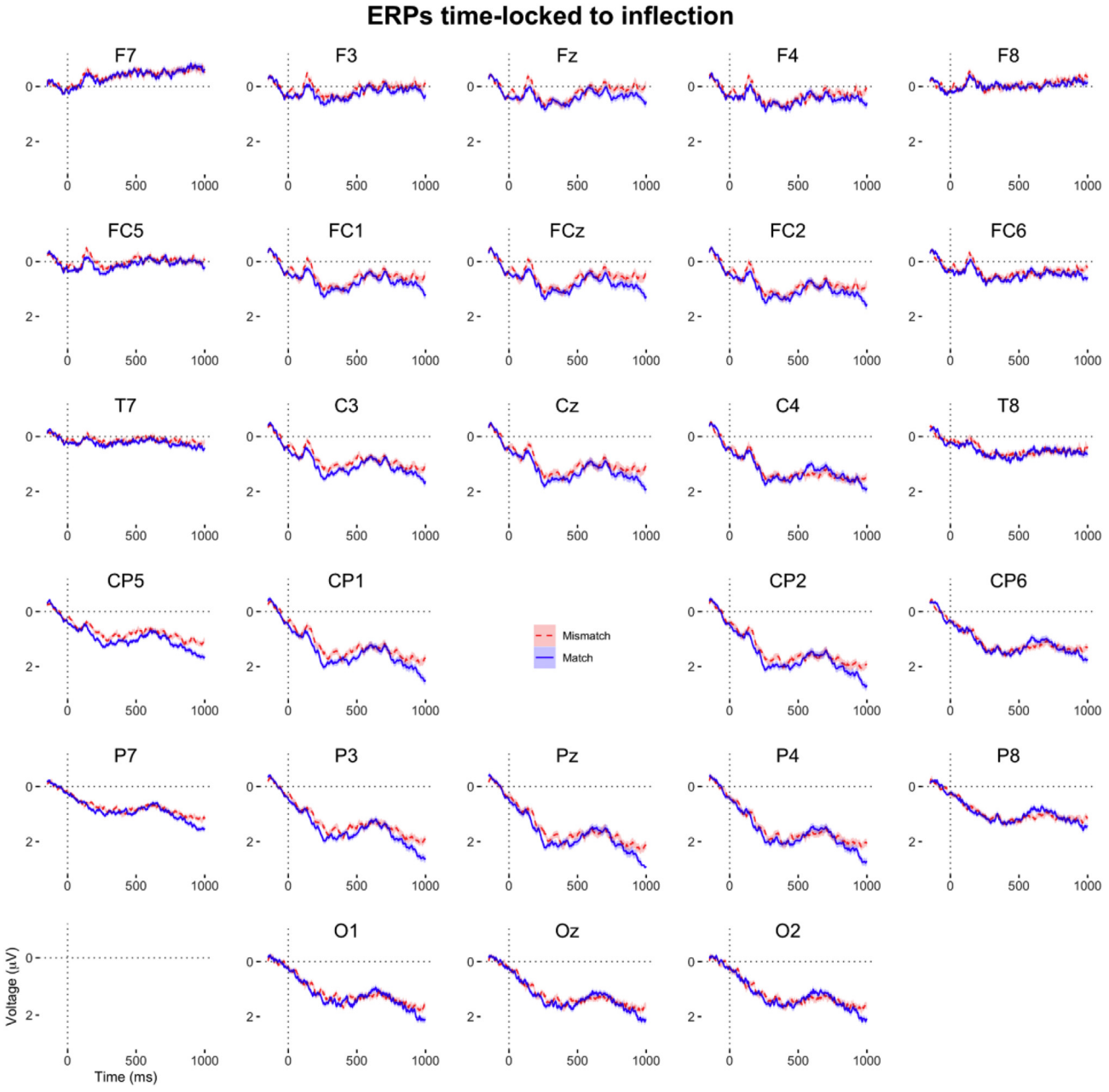

Fig. A.1 – Inflection effects at all individual channels.

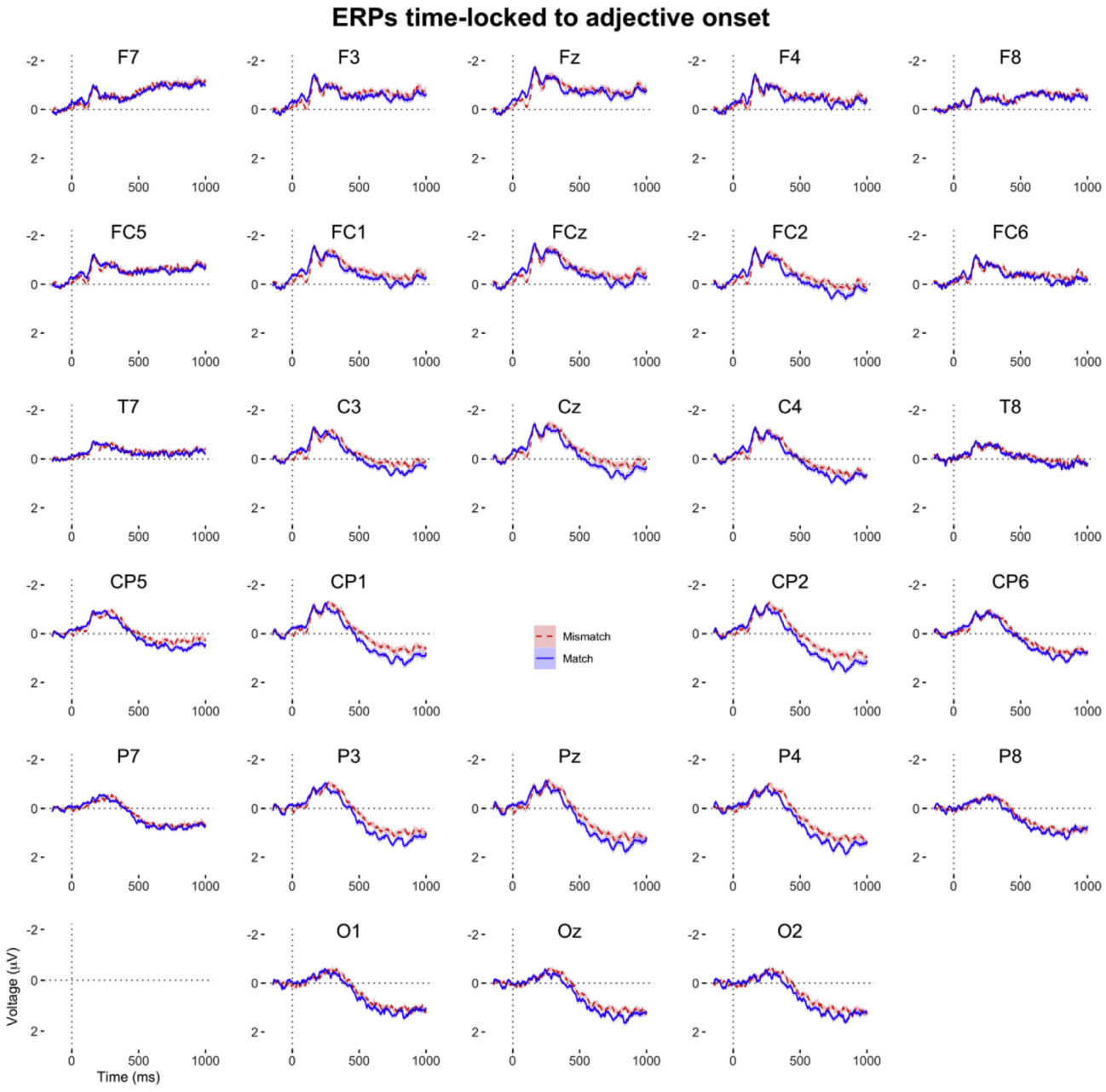

Fig. A.2 – Adjective onset effects at all individual channels.

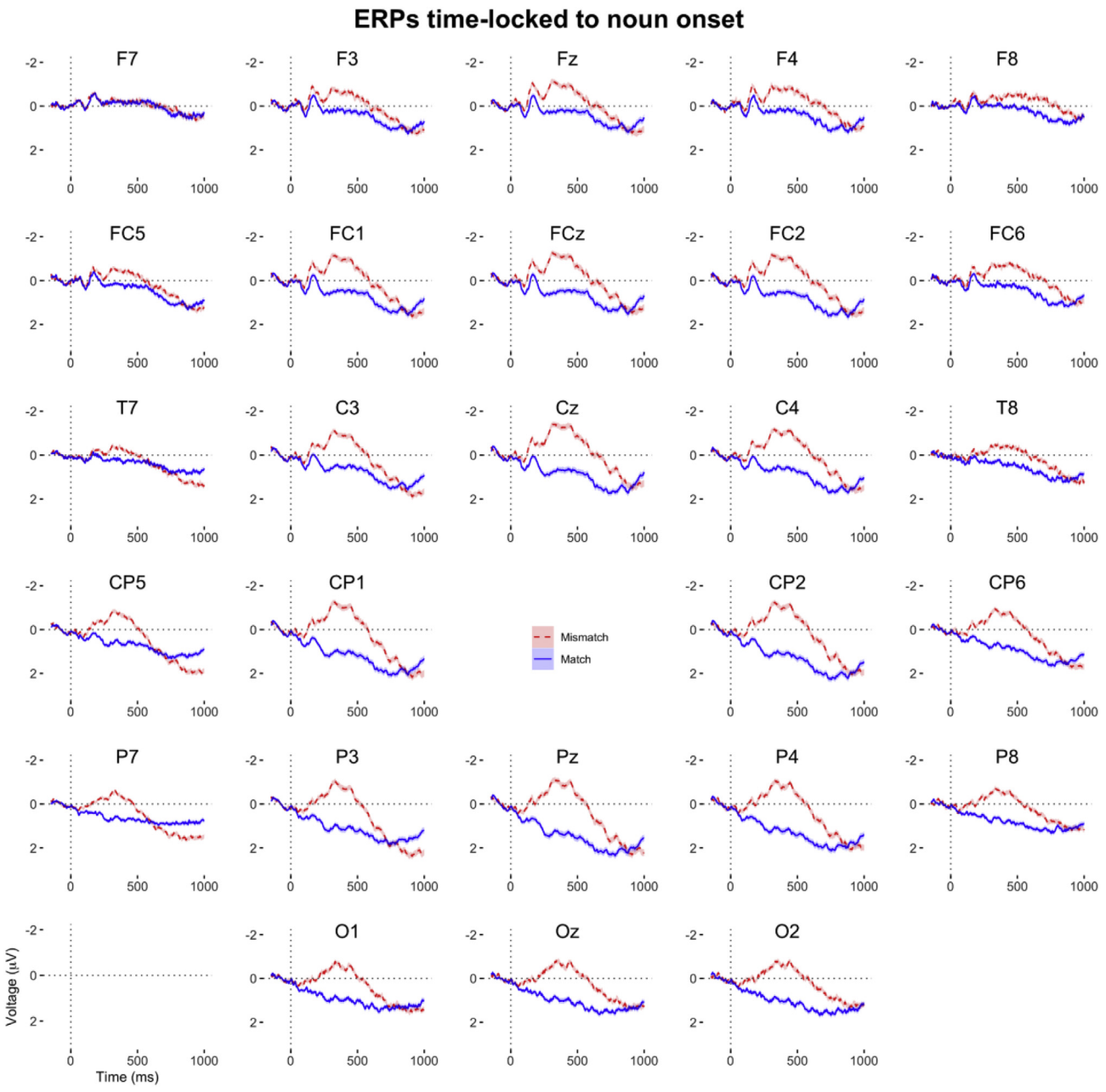

Fig. A.3 – Noun effects at all individual channels.

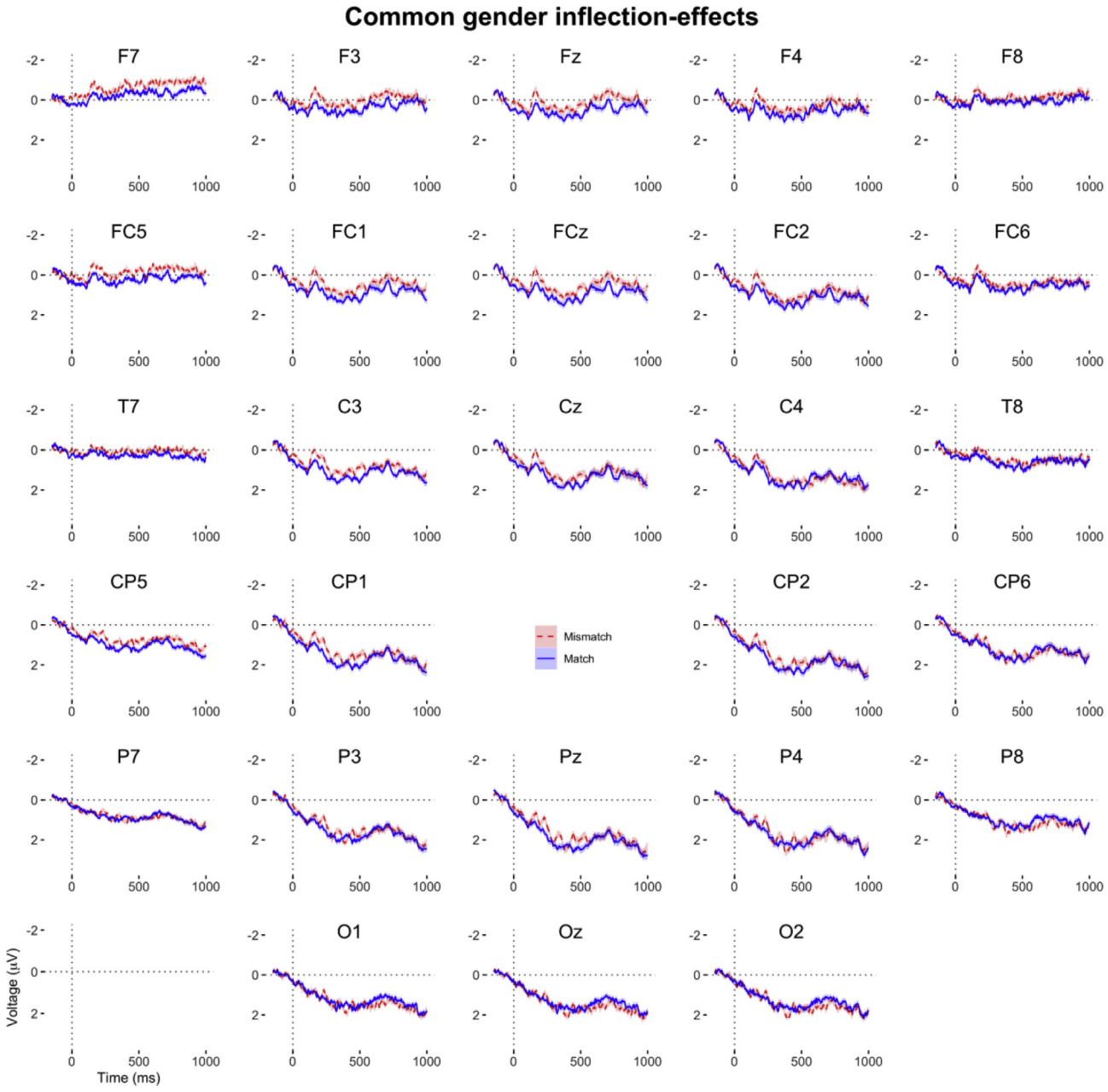

Fig. A.4 – Common gender inflection effects at all channels.

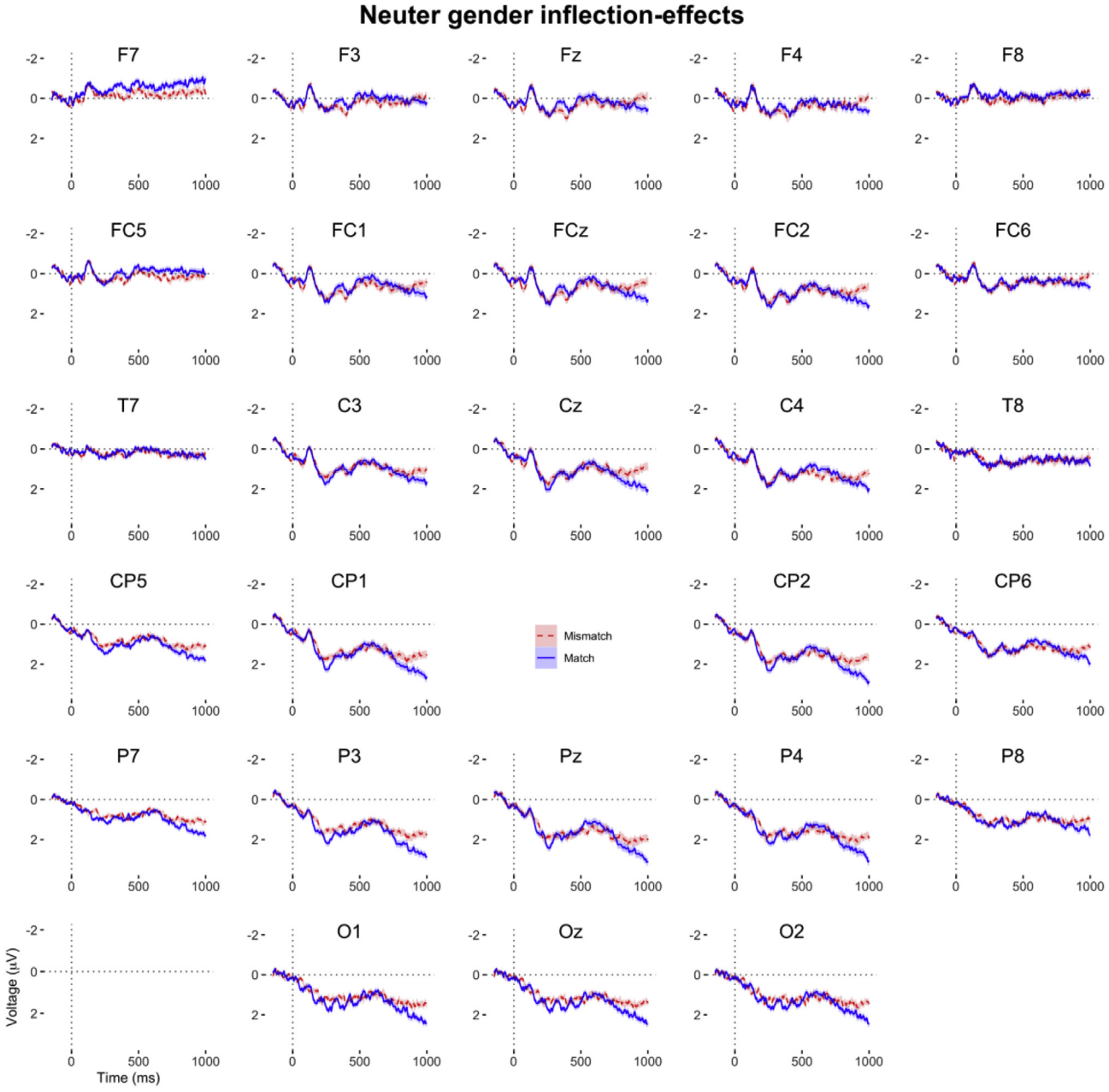

Fig. A.5 – Neuter gender inflection effects at all channels.

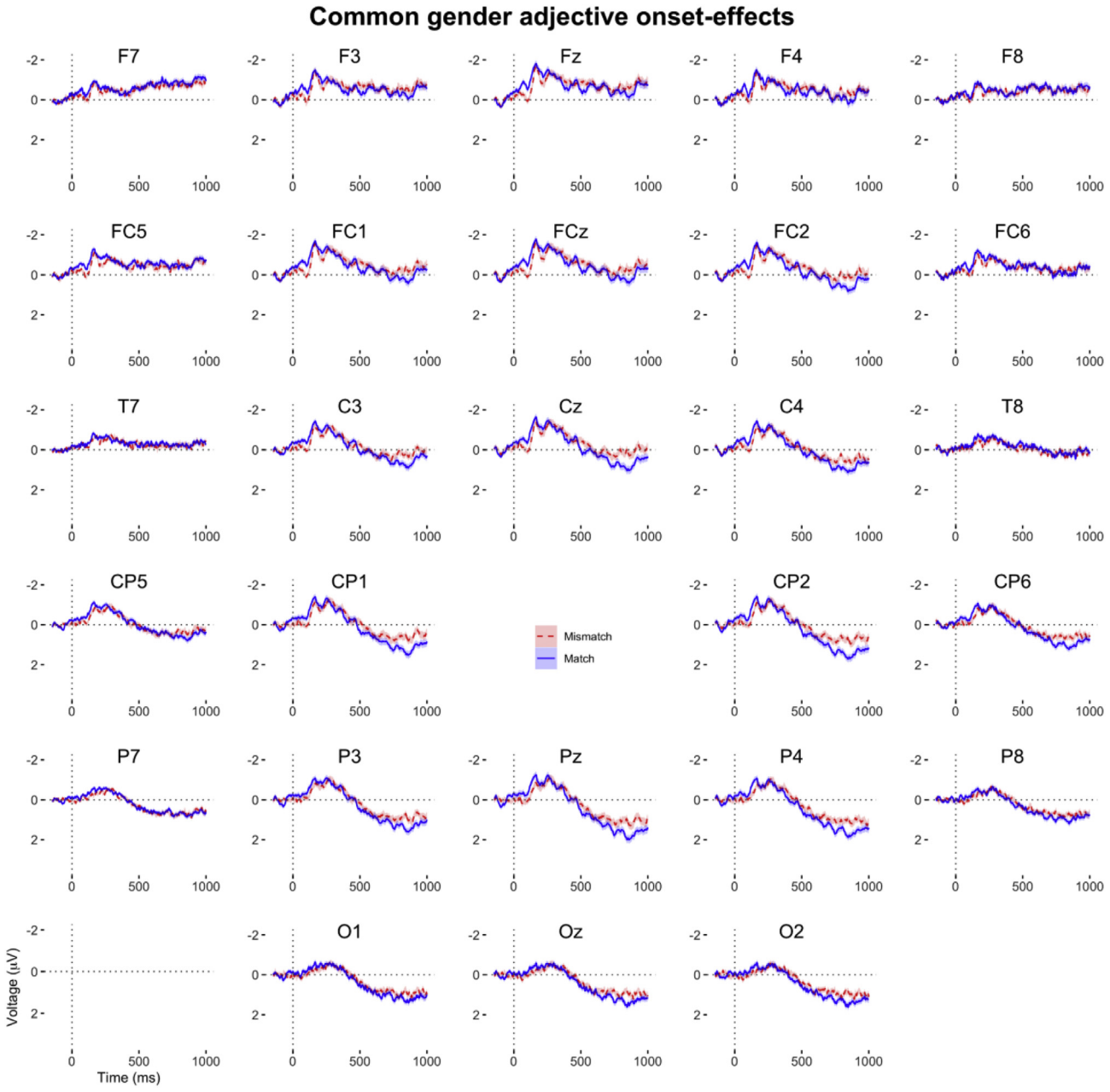

Fig. A.6 – Common gender adjective onset effects at all channels.

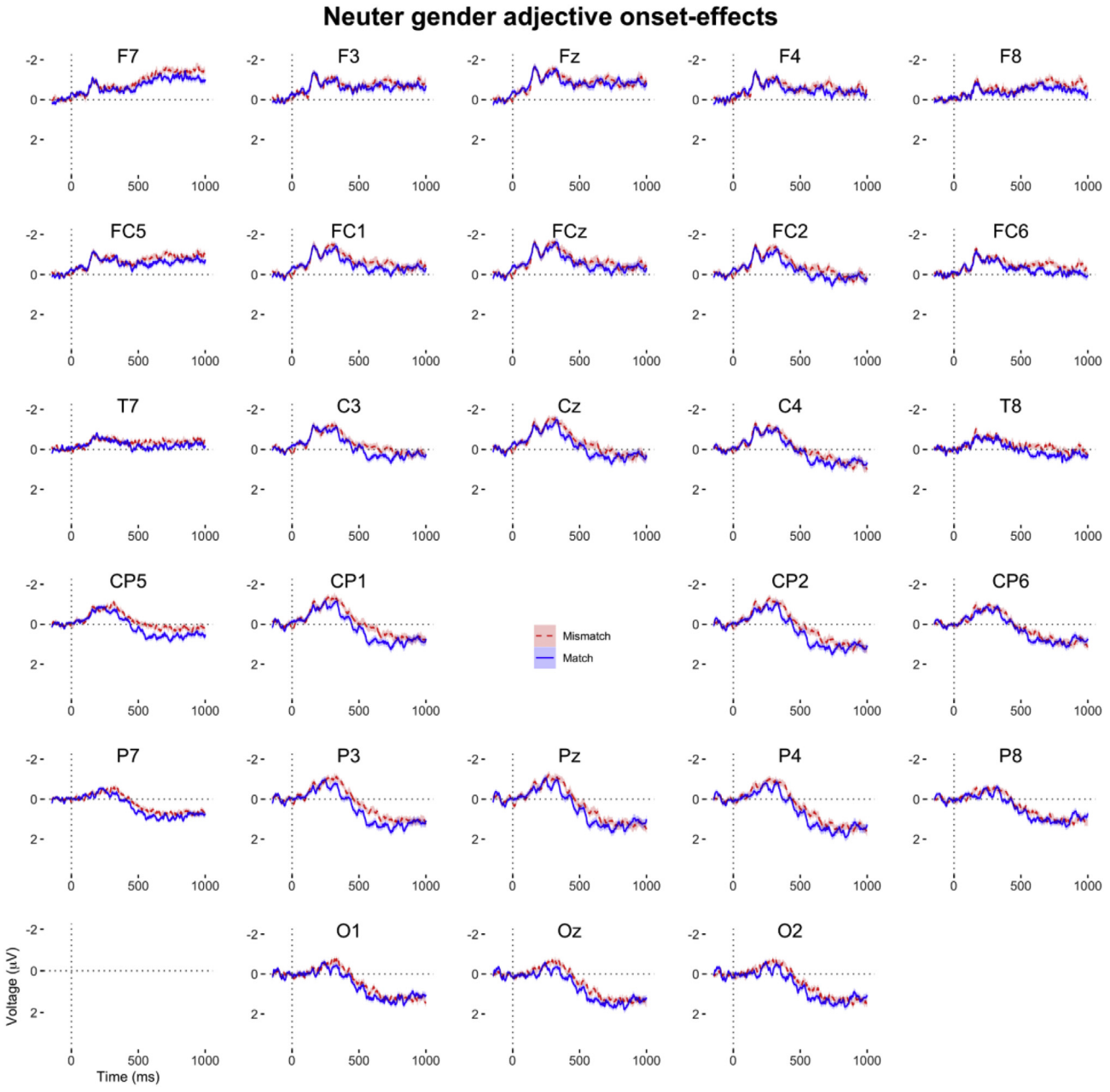

Fig. A.7 – Neuter gender adjective onset effects at all channels.

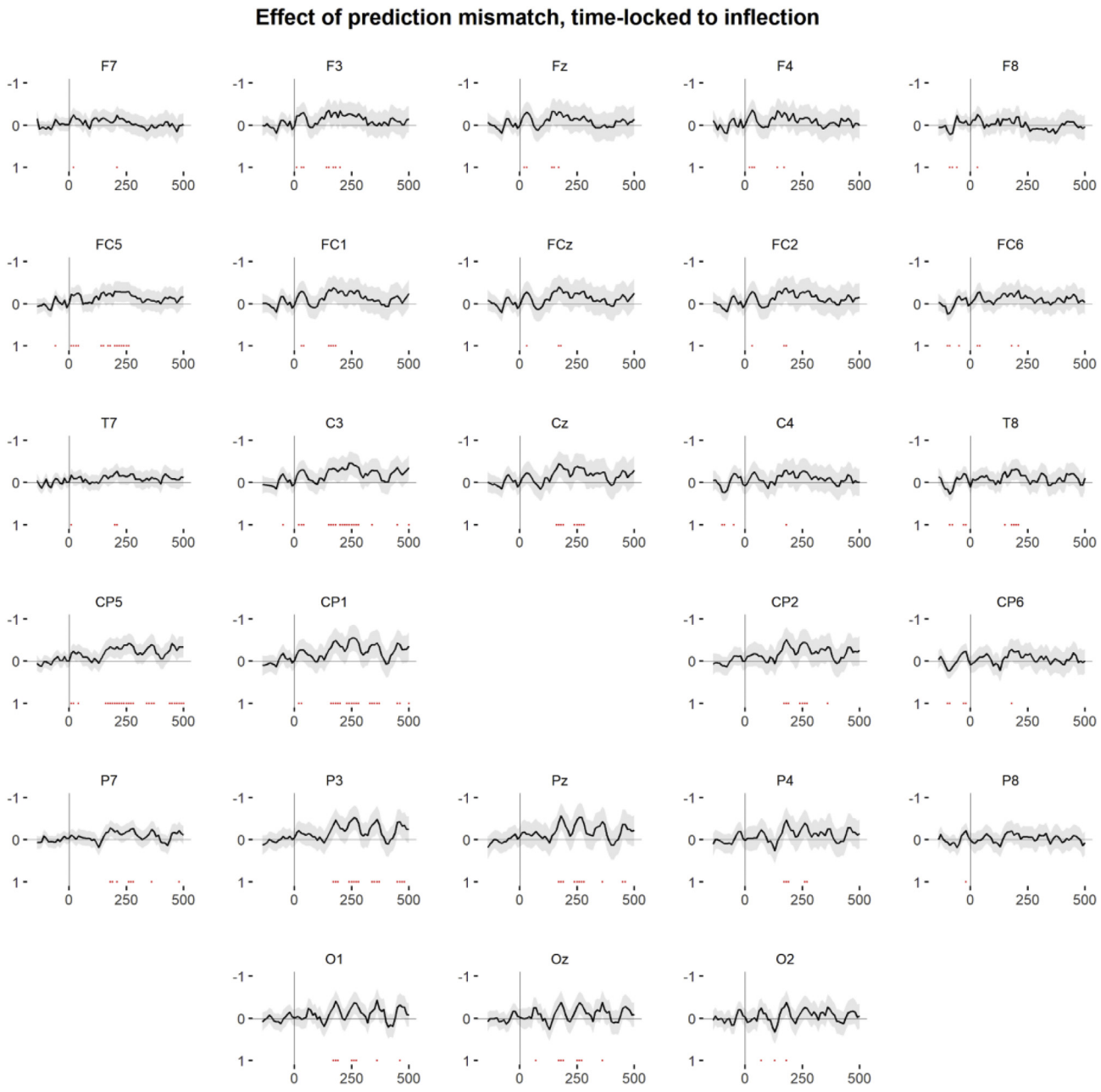

Fig. A.8 – Results from the mass regression analyses. Effect of gender-mismatch (mismatch minus match) on ERP time-locked to inflection onset, plotted as the voltage estimate and corresponding 95% confidence interval (gray area) at each timepoint and channels. Dots underneath the voltage estimates indicate statistically significant samples (not corrected for multiple comparisons). N.B. samples occurring 480 msec after inflection may be distorted by effects associated with noun onset.

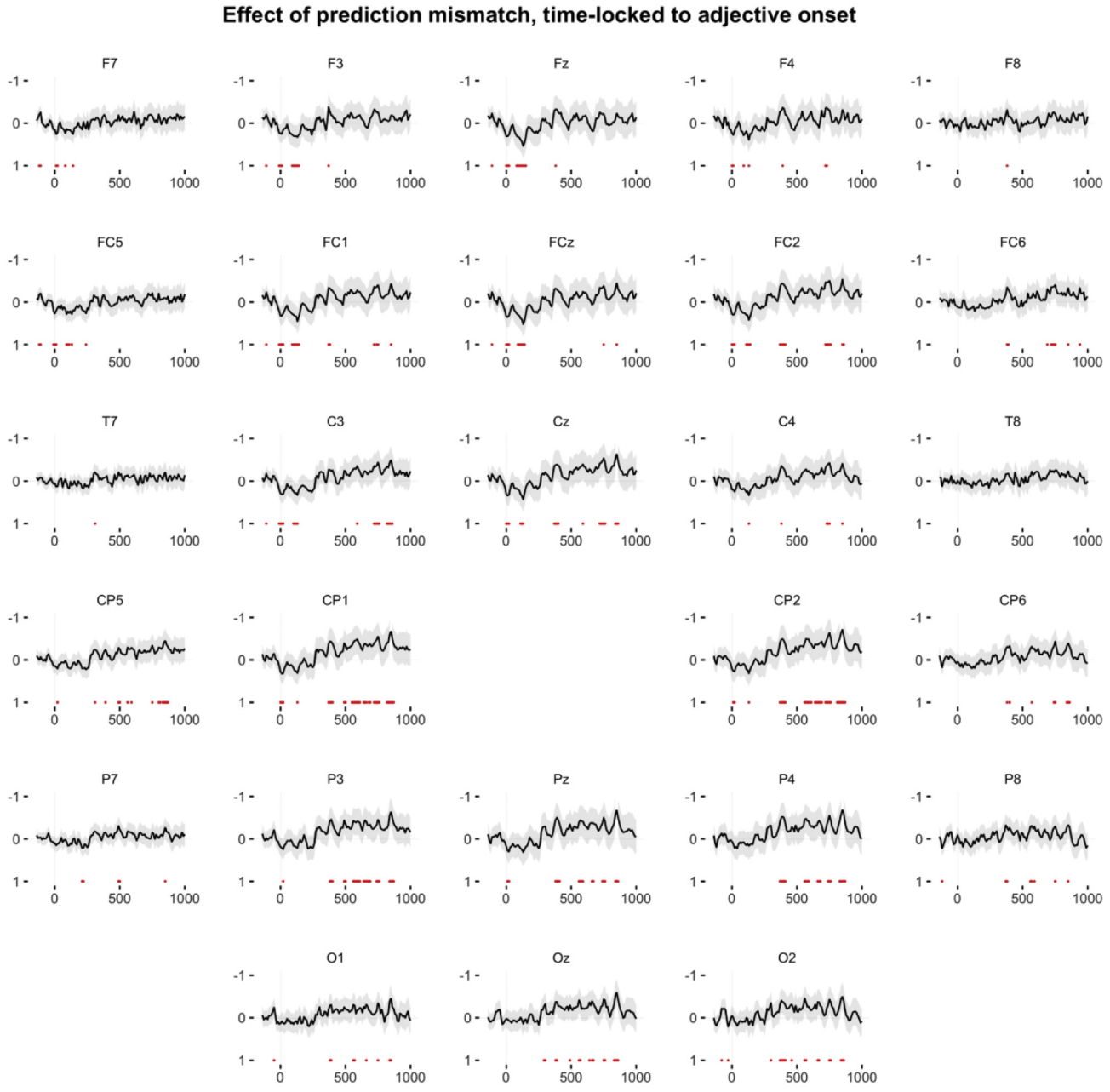

Fig. A.9 – Results from the mass regression analyses. Effect of gender-mismatch (mismatch minus match) on ERP time-locked to adjectives onset, plotted as the voltage estimate and corresponding 95% confidence interval (gray area) at each timepoint and channels. Dots underneath the voltage estimates indicate statistically significant samples (not corrected for multiple comparisons). N.B. samples occurring 800 msec after adjective onset may be distorted by effects associated with noun onset.

Table A.1 – Results from exploratory analyses on ERPs time-locked to inflection, using either a strong positive prior based on VB05 or a strong negative prior based on OT07. Each cell gives the corresponding estimate (b) in μV for the gender-mismatch effect (mismatch minus match), the associated credible interval (CrI), and the posterior probability of the effect being negative ($p(b) < 0$, the percentage of posterior samples under zero), and the BF_{null}.

	Positive prior (VB05) M = .75, SD = .375				Negative prior (OT07) M = −.75, SD = .375			
	b	CrI	$p(b) < 0$	BF_{null}	b	CrI	$p(b) < 0$	BF_{null}
Left-Anterior	−.10	[−.29 .11]	83	16.9	−.20	[−.40 −.01]	98	3.3
Right-Anterior	−.05	[−.27 .17]	69	22.6	−.18	[−.40 .03]	95	6.3
Midline	−.08	[−.30 .14]	76	19.1	−.21	[−.42 .01]	97	4.4
Left-Posterior	−.12	[−.30 .06]	91	12.7	−.21	[−.39 −.03]	99	2.1
Right-Posterior	−.06	[−.26 .14]	72	22.7	−.17	[−.37 .03]	95	6.8

Table A.2 – Interactions between prediction-mismatch and quadrant location (hemisphere, anteriority) for ERPs time-locked to inflection onset (50–250 msec) and adjective onset (300–600 msec). None of the reported F-values reached the traditional alpha = .05 level of statistical significance.

	F values
50–250 msec	
Match * Hemisphere	.6831
Match * Anteriority	.0097
Match * Hemisphere * Anteriority	1.0833
300–600 msec	
Match * Hemisphere	.6333
Match * Anteriority	2.4589
Match * Hemisphere * Anteriority	.1969

REFERENCES

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278–289.

Altmann, G. T., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language, 52*(3), 339–362.

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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.

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.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6*(3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

Coopmans, C. W., & Nieuwland, M. S. (2020). Dissociating activation and integration of discourse referents: Evidence from ERPs and oscillations. *Cortex, 126*, 83–106. https://doi.org/10.1016/j.cortex.2019.12.028

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass, 2*(4), 589–602.

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.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390.

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. https://doi.org/10.1098/rstb.2012.0394

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. https://doi.org/10.1038/nn1504

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218.

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test*. American Guidance Service.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/Bf03193146

Favier, S., Meyer, A., & Huettig, F. (2018). Does reading ability predict individual differences in the syntactic processing of spoken language?. In *Poster presented at the International Meeting of the Psychonomic Society. Amsterdam, the Netherlands*.

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

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *Plos One, 8*(10), Article e77661.

Fleur, D., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition, 204*, 104335. https://doi.org/10.1016/j.cognition.2020.104335

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), 1461–1469. https://doi.org/10.1037/a0036756

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112.

Guerra, E., Nicenboim, B., & Helo, A. (2018). A crack in the crystal ball: Evidence against pre-activation of gender features in sentence comprehension. In *Proceedings of the annual conference on architectures and mechanisms of language processing. Berlin, Germany*.

Hartung, F., Burke, M., Hagoort, P., & Willems, R. M. (2016). Taking perspective: Personal pronouns affect experiential aspects of literary reading. *Plos One, 11*(5), Article e0154732.

Hintz, F., Jongman, S. R., Dijkhuis, M., van 't Hoff, V., Damian, M., Schröder, S., et al. (2018). STAIRS4WORDS: A new adaptive test for assessing receptive vocabulary size in English, Dutch, and German. In *Poster presented at architectures and mechanisms of language processing [AMLaP] conference, Berlin, Germany*.

Hope, R. O. (2013). *Rmisc: Rmisc: Ryan miscellaneous. R package version* 1.5. https://CRAN.R-project.org/package=Rmisc.

Hubers, F., Snijders, T. M., & De Hoop, H. (2016). How the brain processes violations of the grammatical norm: An fMRI study. *Brain and Language, 163*, 22–31.

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., 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. https://doi.org/10.1080/23273798.2016.1242761

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. https://doi.org/10.1080/23273798.2017.1323112

JASP Team. (2018). *JASP (version 0.9)[computer software]*.

Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15*(2), 159–201.

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*(1), 98–110.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology, 47*(5), 888–904.

Kilner, J. M. (2013). Bias in a common EEG and MEG statistical analysis and how to avoid it. *Clinical Neurophysiology, 124*(10), 2062–2063.

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

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178–206.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford University Press.

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. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language, 111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

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*(3), 484–502.

Lenth, R. (2019). *emmeans: estimated marginal means, aka least-squares means. R package v. 1.3. 4*.

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

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001.

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.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189*(4198), 226–228.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1–71.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports, 8*. doi: Artn 107910.1038/S41598-018-19499-4.

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*(4), 574–588. https://doi.org/10.1016/j.jml.2013.08.001

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978.

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), 150547.
Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews, 96*, 367–400.
Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020). 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. Advance Online Publication, 375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522
Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*(3), 324–346.
Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science, 19*(12), 1213–1218.
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. https://doi.org/10.7554/eLife.33468
Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18.
Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 786.
Otten, M., Nieuwland, M. S., & Van Berkum, J. J. A. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience, 8*, 89. https://doi.org/10.1186/1471-2202-8-89
Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes, 45*(6), 464–496. https://doi.org/10.1080/01638530802356463. Pii 905987649.
Otten, M., & Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research, 1291*, 92–101. https://doi.org/10.1016/j.brainres.2009.07.042
Pedersen, T. L. (2019). *patchwork: The Composer of Plots. R package version, 1*, 410.
Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences, 36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495
R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.
Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review, 25*(1), 102–113.
Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001
Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex, 66*, A3–A20.
Schlichting, L. (2005). *Peabody picture vocabulary test-III-NL*. Amsterdam, the Netherlands: Hartcourt Assessment BV.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80.
Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2019). *afex: Analysis of Factorial Experiments. R package version 0.25-1*. https://CRAN.R-project.org/package=afex.
Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, 1–14.
Stan Development Team. (2018). *RStan: the R interface to Stan. R package version 2.17.3*. http://mc-stan.org.
Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402–433.
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 Advance Online Publication*. https://doi.org/10.1037/xlm0000835
Tanner, D., Grey, S., & van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology, 54*(2), 248–259.
Van Berkum, J. J. A. (2004). Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things?. In *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 229–270). Psychology 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), 443–467. https://doi.org/10.1037/0278-7393.31.3.443
Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(2), 394–417. https://doi.org/10.1037//0278-7393.25.2.394
van Ravenzwaaij, D., & Wagenmakers, E.-J. (2019). *Advantages masquerading as 'issues' in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019)*.
Vander Beken, H., & Brysbaert, M. (2017). Studying texts in a second language: The importance of test type. *Bilingualism: Language and Cognition*, 1–13.
Van de Velde, F., & Weerman, F. (2014). The resilient nature of adjectival inflection in Dutch. In P. Sleeman, F. Van de Velde, & H. Perridon (Eds.), *Adjectives in germanic and romance* (pp. 113–145). Amsterdam: John Benjamins.
Vasishth, S., Beckman, M., Nicenboim, B., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics, 71*, 147–161.
Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175.
Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290.

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.

Weerman, F., Bisschop, J., & Punt, L. (2006). L1 and L2 acquisition of Dutch adjectival inflection. *ACLC Working Papers, 1*(1), 5–36.

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters, 346*(3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). 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*(3), 483–508. https://doi.org/10.1016/S0010-9452(08)70260-0

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*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1–20. URL http://www.jstatsoft.org/v21/i12/.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0*. https://CRAN.R-project.org/package=stringr.

Wickham, H. (2020). *forcats: Tools for working with categorical variables (factors). R package version 0.5.0*. https://CRAN.R-project.org/package=forcats.

Wilke, C. O. (2019). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.0.0*. https://CRAN.R-project.org/package=cowplot.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*.