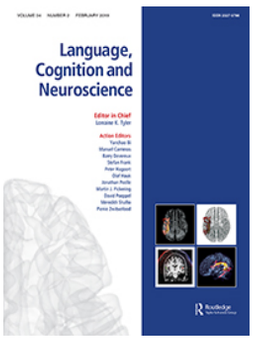

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch

Arnold R. Kochari & Monique Flecken

To cite this article: Arnold R. Kochari & Monique Flecken (2019) Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch, Language, Cognition and Neuroscience, 34:2, 239-253, DOI: 10.1080/23273798.2018.1524500

To link to this article: https://doi.org/10.1080/23273798.2018.1524500

Published online: 21 Sep 2018.

LANGUAGE, COGNITION AND NEUROSCIENCE
2019, VOL. 34, NO. 2, 239–253
https://doi.org/10.1080/23273798.2018.1524500

Routledge
Taylor & Francis Group

REGULAR ARTICLE

OPEN ACCESS Check for updates

Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch

Arnold R. Kochari[a,b] and Monique Flecken[b,c]

[a]Institute for Logic, Language, and Computation, University of Amsterdam, Amsterdam, the Netherlands; [b]Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands; [c]Max Planck Institute for Psycholinguistics, Neurobiology of Language Department, Nijmegen, the Netherlands

ABSTRACT

An important question in predictive language processing is the extent to which prediction effects can reliably be measured on pre-nominal material (e.g. articles before nouns). Here, we present a large sample ($N = 58$) close replication of a study by Otten and van Berkum (2009). They report ERP modulations in relation to the predictability of nouns in sentences, measured on gender-marked Dutch articles. We used nearly identical materials, procedures, and data analysis steps. We fail to replicate the original effect, but do observe a pattern consistent with the original data. Methodological differences between our replication and the original study that could potentially have contributed to the diverging results are discussed. In addition, we discuss the suitability of Dutch gender-marked determiners as a test-case for future studies of pre-activation of lexical items.

ARTICLE HISTORY
Received 7 November 2017
Accepted 11 September 2018

KEYWORDS
Language comprehension; lexical prediction; grammatical gender; ERP; Dutch

Introduction

The process of lexical prediction, i.e. the pre-activation of upcoming words (their meaning, and to some extent, their form) during online sentence comprehension, has been studied in different ways, using different measures and measuring points. For example, visual world eye tracking studies analyse anticipatory eye movements towards visually depicted objects, in relation to specific "cues" in a sentence (a word, morpheme, etc.). These predictive looks are an indicator of the pre-activation of the words referring to the objects depicted. This approach has demonstrated that people, upon encountering specific verbal, nominal, or grammatical forms, and even intonational contours, automatically predict upcoming referents (e.g. Altmann & Kamide, 1999; Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Kamide, Altmann, & Haywood, 2003; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014; Tanenhaus, Spivey-knowlton, Eberhard, & Sedivy, 1995). Using EEG techniques, people have analysed ERPs on specific content words, with the amplitude of the N400 component reflecting the semantic processing of that specific word within the given context (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980; Rommers & Federmeier, 2018). Modulations of the N400 have been obtained for predictability manipulations at different levels, for example, noun semantics (related anomaly paradigm, Federmeier & Kutas, 1999; Federmeier, McLennan, Ochoa, & Kutas, 2002; Kutas & Hillyard, 1984), event semantics (Metusalem et al., 2012; Nieuwland, 2015), but also orthographic or phonological overlap of a presented noun with the predicted noun (Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009). With respect to this type of studies, it is important to note that an N400 on the noun does not necessarily show that the semantics of that word had been pre-activated prior to encountering the noun itself (we will henceforth refer only to the pre-activation of content words as *prediction*; see Kuperberg & Jaeger, 2016 for a discussion). Instead, N400 amplitude modulations may reflect ease of integration of the current content word with the preceding context information (see Kutas & Federmeier, 2011; Van Petten & Luka, 2012 for discussions of what the N400 means for prediction). Hence, these studies do not provide the strongest window onto the process of prediction.

Strong evidence for the pre-activation of lexical-semantic material based on context stems from studies measuring ERPs on forms *preceding* content words, e.g. pre-nominal articles and adjectives. Here, an influential test case concerns a manipulation of the (phonological) form of the indefinite article in English (*a*/*an*; DeLong, Urbach, & Kutas, 2005); the form of the article depends

CONTACT Monique Flecken monique.flecken@mpi.nl

on the first phoneme of the noun following it and this phonological alternation was therefore used to study the degree of prediction of that next word (in particular, the concrete noun following it). Recently, however, this study has been criticised (for discussion and commentary see DeLong, Urbach, & Kutas, 2017; Ito, Martin, & Nieuwland, 2017a, 2017b; Nieuwland et al., 2018) as not providing strong evidence for lexical prediction: the DeLong et al. (2005) findings were not replicated in Nieuwland et al. (2018). It has been suggested that the phonological form of the article in English may not be a good test case for pre-activation of a noun, since it is linked to the next word, which does not have to be a noun (e.g. *an ENORMOUS kite*, Nieuwland et al., 2018).

Other evidence for pre-activation of nominal material stems from languages in which articles and adjectives are marked for grammatical gender (e.g. Dutch, Spanish). In Dutch, nouns either carry common or neuter gender and this is reflected in different forms of the definite article and different adjectival inflections following an indefinite article (een *grote/de* boekenkast [common gender], een *groot/het* schilderij [neuter gender], cf. Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). In Spanish, nouns carry either feminine or masculine gender and this is marked with distinct article forms as well (*el/un* [masculine gender] and *la/una* [feminine gender], cf. Wicha, Moreno et al., 2003; Wicha, Moreno, & Kutas, 2004). All published studies using grammatical gender manipulations to study lexical prediction to date are listed in Table 1. These studies report a difference in ERP amplitudes measured on prediction-consistent vs -inconsistent articles or gender-marked adjectives preceding a noun, suggesting that already at this point a prediction mismatch was detected. Looking closer at these findings, however, we find a mixed picture in terms of ERP latency, polarity and scalp topography.

While a number of studies in Table 1 report unexpected forms eliciting a negativity effect in a time-window overlapping with a canonical N400 time-window (300–500 ms after word onset), the exact time-windows differ, and the topographic distribution of the effects are mostly different from an N400.[1] Moreover, three of the studies in Table 1 report effects that are drastically different from an N400 in terms of time-window and/or polarity. The different test cases, modalities and materials studied complicate a comparison of findings, making it difficult to reflect on the type of processes that may underlie the patterns obtained (see also Kuperberg & Jaeger, 2016; Yan, Kuperberg, & Jaeger, 2017 making a similar point). For example, it is unclear whether the different neural signatures reflect different underlying neural generators, or whether they are caused by specific choices in experimental materials and the design. Moreover, in all of these studies, the time-windows for analyses seem to have been selected based on visual inspection of the data (no other justification is given) which means there could be a higher chance of false-positives with respect to the effects reported (Luck & Gaspelin, 2017). These issues warrant replication studies of the previously reported grammatical gender effects.

Note that, despite this inconsistent picture, all of the studies in Table 1 do report some sort of evidence for lexical prediction. This suggests that grammatical gender might be a suitable test case for exploring the

Table 1. Summary of lexical prediction studies, reporting ERPs on predicted noun gender mismatch.

Study	Manipulation	Polarity	Latency	Topography
Wicha, Bates et al. (2003)	Spanish gender marked articles, presented auditorily, timelocked to onset of the article	Negativity	300–500 ms	Significant across all electrodes; interaction showing stronger effect over frontal and medial sites
Wicha, Moreno et al. (2003)	Spanish gender marked articles, presented visually word by word, timelocked to onset of the article	Negativity	300–600 (700) ms	Significant across all electrodes
Wicha et al. (2004)	Spanish gender marked articles, presented visually word by word, timelocked to onset of the article	Positivity	500–700 ms	Significant across all electrodes, statistically marginally stronger effect over the left hemisphere
Van Berkum et al. (2005), Exp. 1	Dutch gender marked adjectives, presented auditorily, timelocked to the acoustic onset of the inflection at the adjective	Positivity	50–250 ms	Significant across all electrodes
Otten et al. (2007)	Dutch gender marked adjectives, presented auditorily, timelocked to the onset of the adjective	Negativity	300–600 ms	Significant for right frontal quadrant only
Otten and Van Berkum (2008), Exp 1B	Dutch gender marked adjectives, presented visually word by word, timelocked to the onset of the adjective	Negativity	900–1100 ms	Significant across all electrodes
Otten and Van Berkum (2009)	Dutch gender marked definite determiners, presented visually word by word, timelocked to the onset of the determiner	Negativity	200–600 ms	Significant across all electrodes; interaction with hemisphere showing stronger effect over right hemisphere
Szewczyk and Schriefers (2013)	Polish gender/animacy-marked adjectives, presented visually word-by-word, timelocked to the onset of the adjective	Negativity	400–600 ms	Significant across all electrodes; interaction showing centro-posterior distribution

process of lexical prediction further. Nieuwland et al. (2018) support this idea: in a gender-marking system, the form of an article (for example, the definite article in Dutch) is probabilistically strongly linked to the noun following it, regardless of whether it is the immediately following word or not; the article form is not influenced by potential intervening words. It is important to issue a word of caution, however: in Dutch, for example, article forms are not *exclusively* indicative of the gender of the upcoming noun: the article marking common gender also indicates plural nouns (het boek → **de** boeken [the book → the books]; **de** taart → **de** taarten [the cake → the cakes]). Likewise, the neuter gender article (het) also marks all diminutive noun forms (**het** boek → **het** boekje; de taart → **het** taartje). These features make gender-marked article forms (at least, in Dutch) not fully reliable measuring points to study the pre-activation of definite, singular nouns referring to concrete objects. Given this complicating factor, it is again highly important to study whether previously reported grammatical gender effects are robust and replicable.

Here, we present the, to our knowledge, first attempt to replicate one of the previously reported prediction studies using gender-marked articles in Dutch. We use adapted materials from Otten and van Berkum (2009; from now on, O&vB), focusing on ERP modulations on definite articles with expected vs. unexpected gender in high constraining, high cloze sentences. The main focus of the original O&vB study is on a potential modulation of the prediction effect by working memory, taking the existence of this effect as granted. We consider this study as most relevant for our purposes, given that it is, in fact, the only study (in Dutch) measuring the pre-activation of lexical items on forms of articles.[2] We limit our replication attempt to the main ERP effect of prediction reported and will not focus on potential working memory modulations. We also note that the present study is not a direct, but rather a conceptual replication as there are some differences in materials, design and data pre-processing. The dataset we collected initially showed an inconclusive outcome and was submitted for publication. Following the advice of reviewers, data from additional participants were collected. Before the additional round of data collection, we pre-registered the planned EEG data pre-processing and statistical analyses procedures described below. The pre-registration document can be found at https://osf.io/hfwun/.

Method

As mentioned above, we did not do an exact replication of O&vB. Whenever we diverged from the O&vB methods, we mention these differences below.

Participants

Right-handed native speakers of Dutch aged 18–35 were recruited from the participant pool of MPI for Psycholinguistics in Nijmegen. Each participant gave written consent to take part in the study and was paid for their participation according to the local guidelines. The following criteria for exclusion from the analysis were used: comprehension question accuracy rate lower than 75% (note that O&vB did not have comprehension questions); more than 50% of critical trial loss because of EEG artefacts (same criterion as in O&vB); technical problems with the recording or participant's clear unwillingness to cooperate noted in the lab log during data collection (same as in O&vB).

No power analysis to determine the required sample size was performed before the initial data collection. Instead, we aimed to collect data from approximately 30 participants following standard research practices in language ERP research. Because our initially collected data neither clearly rejected the null hypothesis nor showed a clear null finding, peer-reviewers advised us to collect additional data to reach 80% power to detect an effect of the size reported by O&vB. Because at this point we already ran an analysis on part of the data, in order to avoid inflating our Type I error rate, we calculated the required sample size based on the Bonferroni-corrected significance level $\alpha = 0.025$. The effect of determiner expectedness observed in O&vB was of size $\eta_p^2 = 0.15$, 90% *CI* [0.008, 0.336] (computed based on the reported F value for the main effect of expectedness in the ANOVA performed on data from predictive stories, p. 94 of O&vB). This meant that, for 80% power, we needed to have data from 58 participants in total.

Thirty-one participants took part in the study during initial data collection which took place in July and August 2015. For the analysis with determiners,[3] 4 participants were excluded based on our exclusion criteria (see *Results* for details) leaving us with 27 participants with valid data. In the additional round of data collection (February–April 2018), we aimed to stop data collection after data had been collected from 31 participants passing our exclusion criteria. To reach this number, we recorded 39 participants. In total, data were thus collected from 70 participants and 12 datasets were excluded. The mean age of the participants included in the determiner analysis was 22.9 (SD 3.3); 38 were female and 20 male.

For comparison, O&vB collected data from 38 participants and excluded 7, so 31 were included in their analysis.

Materials and design

Participants read 112 two-sentence stories in experimental trials along with 9 other similarly structured stories that were not intended to be used for this project (in total, 121). The first sentence set up the context and the second sentence contained the target determiner and the critical noun. The target determiner and the noun were always separated by 2–5 adjectives. Most of the materials we used were modified versions of the items used by O&vB. Twelve items were directly copied from O&vB, 67 were modified in the region after the target determiner, 16 were modified in the segment preceding the target determiner, and, finally, 17 items were created by us from scratch.

O&vB had an additional, control condition which we did not include; stories in that condition were not meant to lead to prediction of a specific noun (*prime-control* condition; they indeed found no effect of prediction for these). In total, O&vB presented participants with 160 experimental items, of which 80 were predictive and 80 were prime-control stories. Of the 80 predictive stories, 40 trials were in the prediction-consistent condition and 40 in the prediction-inconsistent conditions. In the present study, participants saw 56 trials in each of the prediction-consistent and prediction-inconsistent conditions and almost no filler trials (exactly 9). Thus, there were substantially more stories leading to predictions per participant in the present study.

For each story, there was a prediction-consistent (expected) version, which contained a highly predictable noun and corresponding determiner, and a prediction-inconsistent version, which contained an unexpected, but plausible and coherent noun of the other gender and the corresponding determiner (see Table 2 for examples). The consistent and inconsistent versions of the stories were identical except for the critical region in the second sentence. Fifty-eight experimental sentences contained a highly predictable common gender (DE) noun and determiner and 54 contained a highly predictable neuter (HET) gender noun. The full list of stimuli can be found in the supplemental online materials (http://osf.io/ptqwm).

In order to establish predictability of the target nouns, a cloze task was administered to 15 native Dutch speakers from the same participant pool and of the same age range (4 male; mean age 21.8, SD 3.94, range 19–34) in the form of an online questionnaire. Participants were paid for their time. They saw the first sentence of each story and the second sentence cropped at the position of the determiner (i.e. without the determiner) and were asked to fill in a plausible continuation. In our set of experimental items, mean cloze values of the target noun in the consistent condition was 0.79 (SD 0.11; range 0.60–1). The mean cloze value of the target words in the inconsistent condition was 0.01 (SD 0.04; range 0–0.13). These values are close to the ones reported in O&vB (for prediction-consistent target nouns mean 0.70, SD 0.20; for prediction-inconsistent mean 0.05, SD 0.16; based on 12 Dutch native speakers, 2 males, mean age 22.4 [range 19–26]).

Two lists of stimuli were created, each containing one version of each story. Half of the stories in a list contained the expected noun and determiner and the other half of the sentences contained the unexpected version. The second list consisted of the opposite versions of these sentences. Each participant saw only one of the lists. 25% of the presented sentences were followed by simple yes–no comprehension questions to ensure participants' attention (O&vB did not include comprehension questions, their participants were instructed to simply read the sentences).

Procedure

During the experimental session, participants were comfortably seated in a sound-isolated booth with the

Table 2. Examples of stimulus materials used in our experiment.

Prediction-consistent determiner	Prediction-inconsistent determiner
Nadat hij uren naar het lege doek had gekeken voelde de schilder inspiratie opkomen. Hij greep naar **de** intensief gebruikte **kwast** en smeet de verf op het doek.	Nadat hij uren naar het lege doek had gekeken voelde de schilder inspiratie opkomen. Hij greep naar **het** intensief gebruikte **penseel** en smeet de verf op het doek.
After hours of looking at the blank canvas, inspiration finally struck the painter. He reached for the$_{com}$ heavily used brush$_{com}$ and threw the paint on the canvas.	*After hours of looking at the blank canvas, inspiration finally struck the painter. He reached for the$_{neu}$ heavily used pencil$_{neu}$ and threw the paint on the canvas.*
Het was een prachtige zonnige dag en Thomas en zijn vrienden wilden gaan picknicken. Ze gingen naar **het** grote aangename **park** om daar de middag door te brengen.	Het was een prachtige zonnige dag en Thomas en zijn vrienden wilden gaan picknicken. Ze gingen naar **de** grote aangename **speeltuin** om daar de middag door te brengen.
It was a beautiful sunny day and Thomas and his friends wanted to go on a picnic. They went to the$_{neu}$ big pleasant park$_{neu}$ to spend the afternoon there.	*It was a beautiful sunny day and Thomas and his friends wanted to go on a picnic. They went to the$_{com}$ big pleasant playground$_{com}$ to spend the afternoon there.*

Note: The critical determiner and noun are marked in bold.

experimenter outside. The experiment was run on Presentation® software (Neurobehavioral Systems Inc., www.neurobs.com). Participants were instructed to read stories for comprehension and sometimes answer questions about them. Whereas O&vB presented the whole story word-by-word, we opted for word-by-word presentation of the second sentence only in the interest of saving time and not tiring the participants.

We followed the same procedure for word-by-word presentation, but had different presentation times. Each trial started with a fixation cross that remained on the screen for 2000 ms. The first sentence was presented as a whole on the screen and the participants pressed a key in order to proceed when they were done reading it. This was followed by another fixation cross for 2000 ms. The second sentence was then presented word by word in the centre of the screen. The duration of presentation of each word depended on its length, which was meant to make reading word by word more natural and faster. In the present study, each word except for the critical region was presented for 187 ms plus the number of letters in the word multiplied by 30. For words that were 10 letters long or longer, the presentation time was fixed at 457 ms. The last word of the sentence remained on the screen for an additional 293 ms. The words in the critical region (starting from the target determiner and up to and including the target noun) were presented at a fixed presentation rate of 358 ms (based on the average duration of the word in the critical region across all trials in the same way as it was calculated in O&vB; it was 376 in the O&vB study). The inter-word interval (IWI) during word presentation was set at 400 ms (it was 106 ms in O&vB; we opted for longer IWI for the purpose of the second part of this project irrelevant to the current study).[4] Importantly, longer presentation times have been reported to affect predictive processing, but not negatively (cf. Wlotko & Federmeier, 2015). If the second sentence was followed by a comprehension question, participants answered using a button box.

The sentences were presented in black font (Arial, 21 pt.) against a light grey background (O&vB used Courier New font, 36 pt.). The order of presentation of the trials was randomised and differed for each participant. Participants were asked not to blink during the second sentence (thus, they could blink during the first sentence; in O&vB participants were asked not to blink during the whole story as both sentences were presented word by word). The trials were divided into 4 blocks of equal length with breaks for the participants to rest. Each block lasted approximately 10–15 min depending on the reading speed of the participant. The total duration of the task was around 45 min.

EEG recording and analysis

EEG signal was recorded continuously from 27 active scalp electrodes mounted in an elastic cap (ActiCAP), placed according to the 10–20 convention. In addition to the scalp electrodes, four EOG electrodes were used to detect eye-movements and blinks – to the left of the left eye, above and below the left eye and to the right of the right eye. Finally, two reference electrodes were placed on both mastoid bones. The signal was amplified using BrainAmp DC and recorded using BrainVision Recorder (Brain Products GmbH, www.brainproducts.com). All electrode impedances were kept below 5 kΩ. The recording was done at a sampling rate of 500 Hz, with a time constant of 10 s. The signal was referenced online to the left mastoid and filtered with a low pass filter at 150 Hz.

For offline data processing, we used the Fieldtrip toolbox in Matlab (Oostenveld, Fries, Maris, & Schoffelen, 2011). We followed the procedure for EEG pre-processing reported by O&vB except for the rejection of trials contaminated with eye movements. The signal was band-pass filtered at 0.03–100 Hz and re-referenced to the average of the left and right mastoids. We removed extremely noisy channels from the data (marked during the recording and/or spotted during visual inspection of the raw data before segmentation into relevant trials). Epochs starting 500 ms before the onset of the determiner and ending at 1500 ms after the onset were created. We created two bipolar-referenced EOG channels (one for vertical movements and one for horizontal eye movements) and excluded trials contaminated with artefacts from the data manually (O&vB performed an ICA for eye-movement removal; both of these options are standard procedures for EEG language research, but we considered manual exclusion more appropriate in our case).[5] The remaining trials were baseline-corrected to the mean of −150–0 ms of the determiner onset. Subsequently, all trials in which the signal exceeded ±75 uV were excluded. The remaining trials for each condition were then averaged.

In order to replicate the analysis performed by O&vB, electrodes were divided into four quadrants. The following quadrants were created by crossing *hemisphere* with *anteriority*: left-anterior (F7, F3, FC5, FC1, C3, T7), right-anterior (F4, F8, FC2, FC6, C4, T8), left-posterior (CP5, CP1, P7, P3, O1), and right-posterior (CP2, CP6, P4, P8, O2). O&vB used more (specifically 31) scalp electrodes, but we followed the quadrant locations they specified.

Results

Because our hypothesis concerns the ERP effect on determiners, we will focus on those in this section. As a control

analysis to ensure sensitivity to the Expectedness manipulation, we also report ERPs on the target nouns in a separate sub-section at the end (anticipating the results, we would like to note that we find a large N400 effect for nouns). The statistical analyses presented below were done using R environment (R Development Core Team, 2016). Because this is a replication study, we will focus on comparability of the effect sizes in our study and O&vB.

Two participants were excluded from the analysis due to comprehension question accuracy below 75% (60% and 75%), 2 because of technical problems with the recording, 7 due to more than 50% critical trial loss and, finally, one due to unwillingness to cooperate during the experimental session (12 in total). The average comprehension question accuracy of the participants included was 88.3% (range 80–100%). The average proportion of trials that was filtered out during pre-processing for the determiner was 19.7% (range 2–50%). For comparison, O&vB report a rejection rate of 11% [range 1–27%] (note that they did not exclude trials contaminated with eye movements, but corrected the data using ICA).

Grand averaged ERP waveforms on determiners are shown in Figure 1 and scalp topography of the effect of expectedness is shown in Figure 2.

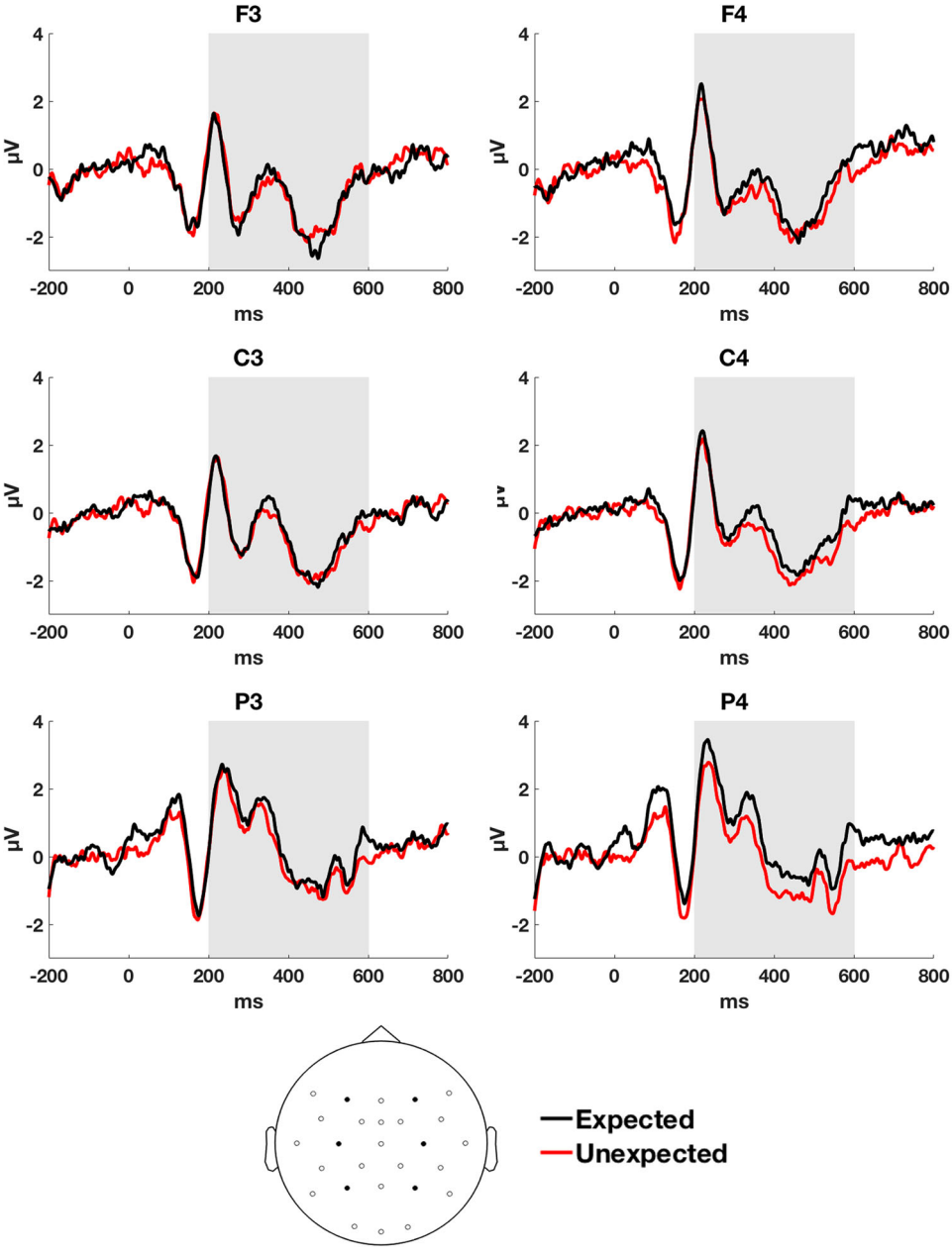

Figure 1. ERPs elicited by target determiners in the expected (black line) and unexpected (c) conditions on individual channels. NB: negative polarity is plotted downwards.

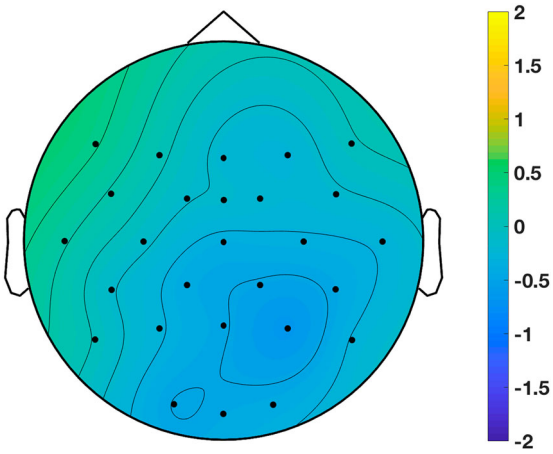

Figure 2. Topographic distribution of the difference between ERP amplitudes in the expected and unexpected conditions for target determiners, in the time-window 200–600 ms. The figure shows the ERP voltage of the unexpected determiner minus the expected determiner.

Null hypothesis significance testing

Repeating the analysis reported by O&vB, we computed an ANOVA with Expectedness, Hemisphere and Anteriority as factors and with mean ERP amplitudes in the time-window 200–600 ms after determiner onset as the dependent variable. Note that parallel to O&vB, this analysis also excluded the midline electrodes (since they were not part of any quadrant). As described above, because we already analysed part of this data collected during initial data collection, we adopt a Bonferroni-corrected significance level of $\alpha = 0.025$.

Contrary to the results of O&vB, we did not observe a significant main effect of Expectedness, F (1, 57) = 2.31, $p = .13$, $\eta_p^2 = 0.039$ 90% *CI* [0, 0.145]. Thus, across all quadrants, there is no difference in ERP amplitudes between the two conditions. Looking at the interaction between Expectedness and Hemisphere, the data show a non-significant effect F (1, 57) = 5.16, $p = .026$, $\eta_p^2 = 0.08$ 90% *CI* [0.005, 0.20] (note that the p-value here is just above our pre-determined α-level). There was also a significant main effect of Anteriority, F (1, 57) = 43.57, $p < .001$, $\eta_p^2 = 0.43$, which is not of interest for the present study. No other factors rendered significant results.

Detectability of the expectedness effect in Otten and Van Berkum (2009)

In order to compare our obtained effect size with the one observed by O&vB, we followed a recommendation in Simonsohn (2015) and asked whether the original study could adequately detect the effect of the size that we observed in our replication. In other words, with the sample size they had, was the original study sufficiently powered to be able to reliably detect that the effect that we observed is different from zero? If yes, then our effect size is consistent with their finding and conclusions. If not, then the effect we observe is incompatible with their conclusions, since, based on what we observe, their study would not have been adequate to draw such conclusions. If the latter is the case, we can say that our findings contradict their conclusion. We will follow the informal suggestion of Simonsohn (2015) to consider an effect size too small for the original study to be considered trustworthy, if the upper bound of the its confidence interval rendered the original study a power of less than 33%.

The effect size for Expectedness in the present study, $\eta_p^2 = 0.039$ 90% *CI* [0, 0.145], is considerably lower than the one reported by O&vB, ($\eta_p^2 = 0.15$, 90% *CI* [0.008, 0.336]). Nonetheless, given the upper bound of the effect size reported here, the original O&vB study would have power 58% to detect this effect. Following our adopted criterion, we conclude that O&vB was sufficiently powered to detect the effect observed in the present study. Thus, even though we do not observe a significant effect in the replication data, our effect size is compatible with the original study's result and conclusion.

Replication Bayes factor analysis

In order to find out whether the effect we observed is more likely under the null or the alternative hypothesis, in our third analysis we turned to Bayesian statistics. Since our starting position is the effect obtained in O&vB, we take that effect as our prior belief and calculate the Replication Bayes Factor (BF) as argued for by Verhagen and Wagenmakers (Verhagen & Wagenmakers, 2014; we also made use of the analysis script made available together with this paper). The reasoning behind this analysis is the following: we are comparing the sceptic's position, who does not believe in the results of the original study and claims that the effect does not exist (i.e. there is no difference in ERP amplitudes in the expected and unexpected condition; H_0), with a proponent's position who believes the effect exists and takes the information from the original study as a prior (i.e. there is a difference between expected and unexpected conditions of the size that O&vB observed; our H_1).

Due to unavailability of ready-made methods to do this for ANOVA, we performed this calculation based on the *t*-statistic for the Expectedness effect. We computed the *t*-value for the Expectedness effect in O&vB from the reported *F*-value. For our own study, we

computed the *t*-statistic of difference between expected and unexpected conditions across all channels. This means that whereas the ANOVA analysis of O&vB excluded the midline electrodes (not for theoretical reasons, but because of aiming for an even quadrant split), in our *t*-statistic value we did include them, since we are interested in a potential effect across all channels. Besides the Replication BF value, we also inspect the Meta-analysis BF value which pools the effect sizes from both the original study and the replication together. Following guidelines suggested by, e.g. Kass and Raftery (1995), we only consider BF values above 3 as conclusive.

Results of the Replication BF analysis are depicted in Figure 3. The Replication BF we obtain is 1.24 which means that our data are 1.24 times more likely under the alternative than under the null hypothesis. Pooling the effect sizes from both studies together, the Meta-analysis BF = 1.87. These BF values do not offer support to either the null or the alternative hypothesis.

Analysis of N400 effect on the target noun

For completeness, we also report the N400 effects that were observed on the noun.[6] In this analysis, two additional participants were excluded due to more than 50% trial loss. The average proportion of trials that was excluded during pre-processing of the data for the noun analysis was 19% (range 1–50%). Grand averaged ERP waveforms on target nouns are shown in Figure 4 and scalp topography of the difference between conditions is shown in Figure 5.

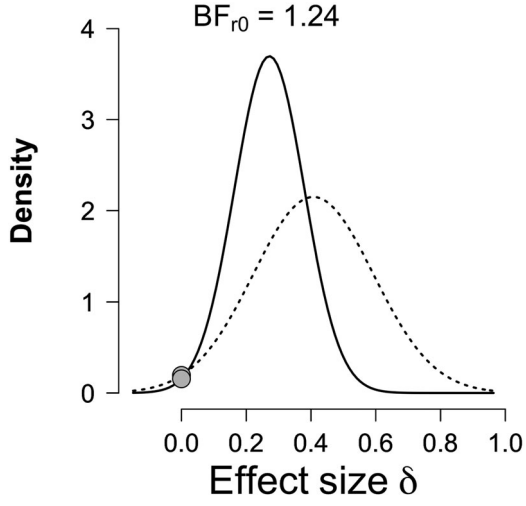

Figure 3. Results from the replication BF test. The dotted line represents the prior based on the effect size observed by O&vB, whereas the solid line represents the posterior belief in the alternative hypothesis after taking into account our observed effect. The grey dots represent believability of H_0 before and after our data is taken into account. BF_{r0} is the ratio of the believability of H_0 before and after our data is taken into account, it only slightly decreases with our data taken into account.

For nouns, mean amplitudes of ERPs between 300 and 500 ms after target noun onset (canonical N400 time-window) were computed. As with determiners, we divided the electrodes into four quadrants and Expectedness, Hemisphere and Anteriority were entered as predictors in the statistical analysis.

We observed a main effect of Expectedness, F (1, 55) = 29.19, $p < .001$, $\eta_p^2 = 0.34$ 90% *CI* [0.17, 0.47], a main effect of Anteriority, F (1, 55) = 122.75, $p < .001$, $\eta_p^2 = 0.69$ as well as a significant interaction of Condition and Anteriority, F (1, 55) = 8.08, $p = .006$, $\eta_p^2 = 0.12$. Thus, the unexpected nouns elicited more negative ERP waveforms in this time window and this effect was present across all electrodes, with a more pronounced effect on posterior electrodes.

Exploratory analysis

O&vB report an effect of Expectedness on the determiners which is bigger in the right than the left hemisphere (main effect of Expectedness, and interaction of Expectedness and Hemisphere). An inspection of raw condition means in the present dataset shows a larger numerical difference between expected and unexpected items in the right hemisphere, compared to the left hemisphere. In order to explore a potential right-lateralization of the Expectedness effect in the present study, we conducted two post-hoc *t*-tests, looking at the effect of Expectedness in each hemisphere separately. The effect of Expectedness does not survive correction of the *p*-values for multiple comparisons in either hemisphere, however (right hemisphere: $t(57) = 2.1$, $p = .020$, one-sided, uncorrected for multiple comparisons, $d_z = 0.27$ 90% *CI* [0.05, 0.49]; left hemisphere: $t(57) = 0.63$, $p = .26$, one-sided, uncorrected for multiple comparisons, $d_z = 0.08$ 90% *CI* [–0.13, 0.29]). We thus do not find a similarly strong hemispheric difference.

Discussion

We did not replicate the pattern of lexical prediction reported in O&vB in terms of a significant ERP modulation in relation to expectedness on gender-marked articles. In addition, we failed to obtain conclusive evidence regarding the existence of the expectedness effect in this replication attempt with Replication BF test. However, the patterns found are in the expected direction in terms of polarity and scalp distribution; moreover, the obtained effect size is consistent with the findings of O&vB. We take the relative similarity of

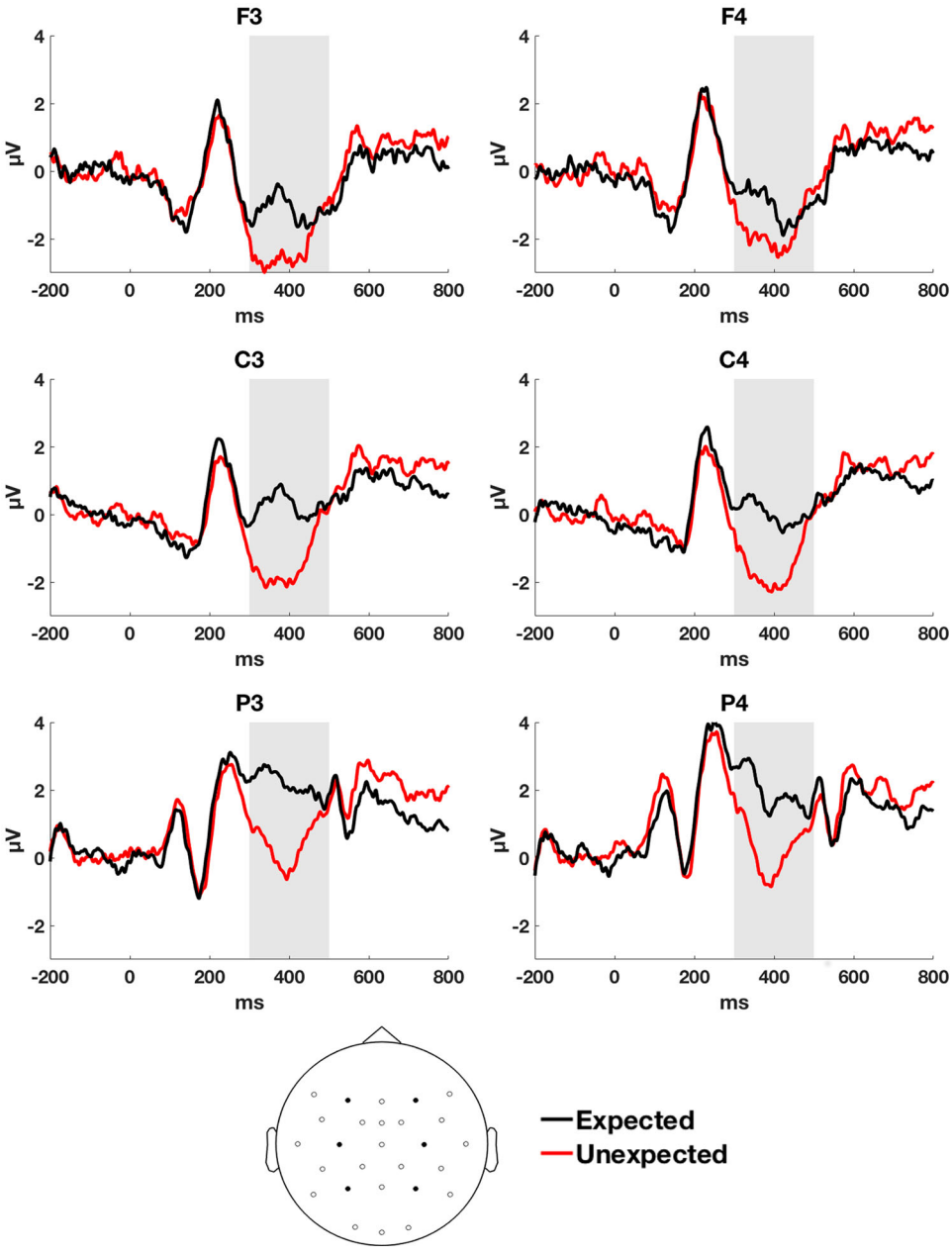

Figure 4. ERPs elicited by target nouns in the expected (black line) and unexpected (red line in online version, gray line in print version) conditions on individual channels. NB: negative polarity is plotted downwards.

our and the O&vB findings as encouragement for further investigation of lexical prediction in relation to grammatical gender manipulations. Nonetheless, at this point we have to conclude that there is no strong evidence for the pre-activation of nouns, measured on gender-marked articles, in Dutch. Moreover, visual inspection of the whole epoch ERPs shows that the selected time-window 200–600 ms may not be showing a pattern that is different from the rest of the epoch. The difference between conditions that is present in this particular time-window is there from the beginning of the epoch. Since our goal here was only to replicate the original analysis and draw conclusions based on that, we leave resolution of this issue for future research.

Importantly, the present study differs from O&vB in a number of aspects as it was not intended to be a *direct* replication. As such, the materials and certain design and procedural aspects were adapted from the original study. In addition, originally a different data pre-processing and analysis procedure was adopted (Kochari, 2015). Given the recent discussion on the Delong et al. effects (Nieuwland et al., 2018) and substantial differences in types of effects in previous grammatical gender studies (cf. Table 1), the present data were

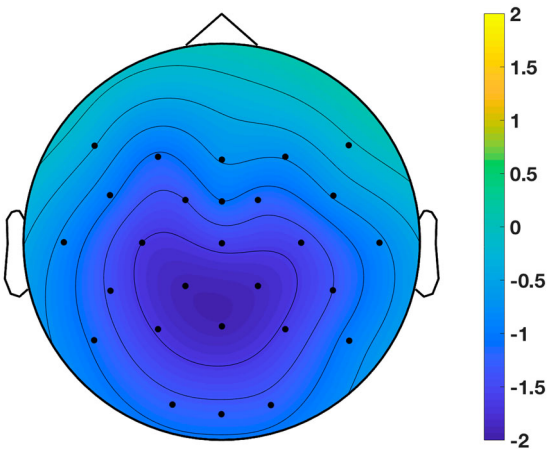

Figure 5. Topographic distribution of the difference between ERP amplitudes in the expected and unexpected conditions for target nouns, in the time-window 300–500 ms. The figure shows the ERP voltage of the unexpected noun minus the expected noun.

reanalysed following the precise steps reported in O&vB, providing a conceptual replication of the original study. Below, we first address the potential role of the methodological discrepancies between the original and the replication study. Then, we discuss some important issues in relation to studies looking at lexical prediction using grammatical gender manipulations.

The current dataset is publicly available at http://osf.io/ptqwm/ (note that we are also willing to share raw EEG data upon request). We encourage further analyses and hope it will be a valuable contribution to the current discussion on prediction in language comprehension.

Comparing the original and replication studies: methodological discrepancies

Even though we do not replicate the effects observed by O&vB in this study, we did see a difference between conditions in terms of raw means, and the topographic distribution of the effect is similar to the one reported by O&vB. Thus, one possibility is that pre-activation of nouns is observable on grammatical-gender marked determiners in Dutch, but in our study these effects were attenuated as compared to O&vB due to study design differences. In this section, we discuss potentially relevant methodological aspects of both studies.

One potentially relevant difference relates to stimulus presentation rate, which has been shown to affect comprehenders' degree of engagement in predictive processing (Ito et al., 2016; Wlotko & Federmeier, 2015). However, based on the results of previous studies, if anything, the direction of the difference should have enlarged the prediction effects in the present study. The present study had slower stimulus presentation rates than the original study (1.3 words per second, compared to 2.07 words per second in O&vB; difference arising due to IWI at present being 400 ms instead of 106 ms in O&vB), and previous studies report larger prediction effects on nouns at *slower* presentation rates (see Ito et al., 2016; Wlotko & Federmeier, 2015; see also N400 effects in Dambacher et al., 2012).

Another factor concerns the use of a smaller number of filler items in the present study, leading overall to a higher proportion of unexpected words, and thus disconfirmed predictions, in the present study, i.e. in almost half of the trials (exactly 46.1%). In O&vB, only 25% of trials involved unexpected words. Yan et al. (2017) propose that the absence of filler trials in Nieuwland et al. (2018) (in contrast to DeLong et al., 2005) created an experimental context with a larger proportion of unexpected words and thus an environment of greater uncertainty concerning upcoming sentence material. This may have led participants to adapt (specifically, *reduce*) their overall predictive behaviour in the experiment. It has indeed been shown that N400 effects on individual words can become smaller in environments that discourage the formation of predictions (Delaney-Busch, Morgan, Lau, & Kuperberg, 2017; Lau, Holcomb, & Kuperberg, 2013; Lau, Weber, Gramfort, Hämäläinen, & Kuperberg, 2016). Considering the Delong et al. materials, consisting of similarly structured short sentences with the target noun always in sentence-final position, participants' awareness of the high likelihood that these nouns were unexpected could be triggered, leading them to adapt their expectations. Looking at the present study, however, materials are more variable (trials consisted of two sentences with varying structures), so it is not clear to what extent participants could have adapted their predictive behaviour to the high number of unexpected words in the current experiment. In order to better quantify the difference between O&vB and the present study in this respect, we looked at the proportion of determiner phrases with expectation violations that participants saw in each experiment (as suggested by a reviewer). Specifically, we counted the total number of determiner phrases with a definite or indefinite determiner in the materials and calculated the percentage of prediction-inconsistent determiners out of this total. While in case of the present study 16% of all determiner phrases were expectation violations, in O&vB this was 9%. This shows overall that indeed the present study contained a larger number of prediction-inconsistent determiners, so adaptation to experiment materials remains a factor that could potentially be responsible for the lack of a lexical pre-activation effect in our replication.

Overall, we support previous suggestions (Brothers, Swaab, & Traxler, 2017; Kuperberg & Jaeger, 2016; Yan et al., 2017) of a need for more thorough investigation of the influence of specific contextual and experimental design aspects on the effect sizes obtained. Speculatively, one can also take the current set of findings as showing that generally, lexical prediction effects in relation to gender-marking are relatively small and not very robust in normal language comprehension, in the sense that they can easily be modulated by specific details of given experimental set-ups. It is thus important to investigate these issues further.

Prediction effects at the target noun

Another aspect of the present results that deserves attention is the effect that we observed at the noun position. While we did not obtain prediction effects at the determiner position, we did see a significant difference in ERPs elicited by prediction-consistent and prediction-inconsistent nouns. Observing an N400 effect at the noun in many of the other studies looking at pre-activation using pre-nominal marking has been taken as diagnostic for whether the experiment was able to successfully detect modulations of N400 amplitudes at all (DeLong et al., 2005; Nieuwland et al., 2018). Note, however, that whereas in those studies the noun directly followed the determiner, in O&vB and the present set of materials the determiner and target noun were always separated by 2–5 adjectives. Because of this large time-window between encountering prediction-inconsistent information and the target noun, it is possible that participants would drop or revise their prediction in the unexpected condition. In fact, there is emerging evidence for prediction updating upon encountering prediction-inconsistent grammatical information (see Chow & Chen, 2018; Szewczyk, 2018). However, we still observe a relatively large N400 effect in relation to Expectedness on the nouns. There are two possibilities for why we still see the N400 effect on the nouns. First, participants did not pre-activate (the grammatical gender of) the expected nouns and therefore were not surprised to see a determiner of a mismatching gender. This means they did not have to drop or revise their prediction. Only upon encountering the noun itself did their processing system notice that this noun does not match with the representations activated by prior context. The second possibility is that participants *did* pre-activate (the grammatical gender of) the nouns and at the encounter of the unexpected determiner they dropped or revised their prediction. However, because their prediction was confirmed in the expected condition, the N400 effect that is observed on the noun is the difference between an expected noun and no prediction at all/revised prediction in the unexpected condition. To shed light on this issue, it is interesting to examine ERPs on the noun in O&vB, since, according to their results, participants in their study did detect a prediction-mismatch already at the determiner. They also observed a significant N400 effect at the noun position (Otten, 2008, chapter 4 which reports the same study as O&vB). Given the fact that prediction effects were found in both locations, the latter possibility seems most likely: The N400 effect of Expectedness on the nouns in these materials reflects the processing difference between a highly expected noun and no prediction at all or a revised prediction for unexpected items.

Dutch grammatical gender: considerations for future studies

A number of studies investigating the pre-activation of specific content words during language comprehension were conducted in *Dutch*, exploiting its feature of grammatical gender agreement (Otten & Van Berkum, 2008, 2009; Otten, Nieuwland, & Van Berkum, 2007; Van Berkum et al., 2005). The line of reasoning is that each noun carries one of two possible grammatical genders which the definite determiner and the adjective have to be in agreement with. Thus, upon encountering a determiner or an inflected adjective incompatible with a predicted noun, predictions are disconfirmed at that moment in time. However, there are some factors associated with Dutch grammatical gender agreement that complicate its use as a test case for the pre-activation of lexical material.

As mentioned in the Introduction, in Dutch, the two determiners and adjectival inflections are not exclusively applied to singular, concrete nouns. The definite determiner DE is also used for any noun in plural form. The only nouns that cannot be marked with the determiner DE are mass nouns of neuter gender (*het water*, the water) which do not have plural forms. Similarly, the definite determiner HET is used for any noun in a diminutive form. Diminutives are prevalent in everyday spoken Dutch and practically any noun can be diminutivised (e.g. Shetter, 1959). This imperfect match between determiner and noun implies that, in case of an unexpected determiner (e.g. seeing DE when expecting HET [a neuter gender noun]), comprehenders may not need to revise their prediction at the semantic level, but only at the level of form (i.e. revising their expectation of a singular noun to a plural noun). Hitherto, this feature of the Dutch gender marking system has not been highlighted or discussed in predictive language processing literature. An important issue for future research is to take into account in a

systematic way to what extent the materials used allow for or favour anticipation of plural or diminutive forms.

Nevertheless, despite gender-marked articles and adjectival inflections being an imperfect predictive cue in Dutch, predictability effects have been shown consistently in previous studies. If these effects are real, what type of processing do they reflect in light of the non-exclusivity of predictions based on DE and HET? One possibility is that in those studies participants had strongly committed to the upcoming word being a singular non-diminutive form and, upon encountering the article, revised the *meaning* of the upcoming noun. An alternative possibility (as outlined above) is that the ERP effects observed on determiners do not reflect revision of the expected noun's meaning, but revision of the expected noun *form*. This would still mean, however, that predictions at the semantic level had been generated (e.g. a noun requiring DE), but the form encountered did not match the prediction (this is where potentially a revision has to take place, e.g. expect a diminutive of the same noun). Importantly, both explanations would still speak for pre-activation of lexical items and their grammatical gender, but ERP effects would reflect different revision processes. It is important to investigate this in more detail, as it sheds light on what these prediction effects entail (see also Kuperberg & Jaeger, 2016).

An important related issue is that the critical determiners were always presented in their definite form (the gender-distinction is only visible on definite determiners in Dutch), regardless of whether it would be pragmatically felicitous in the given context. The question thus emerges to what extent effects could be driven by disconfirmed predictions and revision processes related to *definiteness* of the articles (e.g. expecting an indefinite article, but encountering a definite form). Given that participants saw definite articles in both conditions, we may assume that any predictions in relation to definiteness should be equal across conditions, and thus not of influence on the patterns observed. However, potential influences of definiteness should be taken into account more carefully in future studies (also see Schlueter, Namyst, and Lau (2018) who find an N400 effect of (un)expected definiteness of articles in English).

Grammatical gender agreement as a test case for lexical prediction: open questions

Making use of grammatical gender agreement as a test-case for lexical prediction across different languages makes sense, given that gender systems have distinct pre-nominal forms for distinct genders. As highlighted in recent discussions (Huettig, 2015; Kuperberg & Jaeger, 2016; Yan et al., 2017), different levels of prediction (meaning, form, phonology) should be distinguished. It is not completely clear which level of prediction we are dealing with in case of grammatical gender-related effects. One question discussed above is whether we actually predict the specific *form* of a *noun* (such as, whether it is going to be a plural or diminutive) or only its meaning (see the above section also). Another question is whether we pre-activate a specific noun or also a specific *form* of the preceding determiner/adjective inflection. In other words, do we see a mismatch effect at the point of encounter with the determiner, because it is difficult to integrate it with an already activated noun? Or do we find these effects because the predicted noun in fact pre-activated the specific form of the determiner/adjectival inflection, and the predicted form is not encountered at that point in time? Systematic investigation of the time-course of the ERP effects associated with different predicted information will be of relevance to answer this question.

An important question is whether we predict only one particular lexical item or multiple lexical items with different weights assigned to them. In other words, is prediction a gradient or an all-or-nothing phenomenon? One of the important findings in the original DeLong study was the correlation between article cloze values and the size of the ERP effect obtained on the article. It was interpreted as the first demonstration of the graded nature of prediction in language comprehension (however, see Van Petten & Luka, 2012 for criticism of this view). Since this effect has not yet been replicated (Nieuwland et al., 2018), this question is still open. None of the studies exploiting grammatical gender have tried to correlate predictability and the observed ERP effect yet; this is an important issue to explore in future grammatical gender-related prediction research.

Finally, an interesting point that can be useful for future research is that in all studies using grammatical gender to look at predictive processing so far, cloze values of the noun were used as a measure of expectedness of the determiner. This is different from DeLong et al. (2005) who used the cloze value of the article itself as a predictor of the ERP amplitudes on predicted vs. unpredicted articles. By looking at the cloze value of the article itself, DeLong et al. operationalised predictability as the sum of cloze values of all predicted nouns compatible with that article.

Conclusion

This paper reports a conceptual replication attempt of Otten and Van Berkum (2009), one of the studies reporting evidence for lexical prediction, using pre-nominal material (Dutch gender-marked articles). We did not replicate this effect. The original effect was used to

argue for pre-activation of nouns in Dutch. Our results thus cast doubt onto the validity of yet another strong piece of evidence for lexical prediction in general (see Nieuwland et al., 2018). Although we followed the original set-up closely, there are some important methodological discrepancies which could have played a role for the weaker effect reported in the present study. These discrepancies call for more research to conclusively argue for the presence or absence of prediction effects at gender-marked determiners in Dutch. In relation to this replication attempt, we also raised some points concerning the use of (Dutch) grammatical gender marking in answering questions about predictive language processing which could be taken into account in future research.

Results of multiple previous studies using grammatical gender discussed in this paper support the idea that we *do* routinely pre-activate lexical items along with their grammatical features in language comprehension, one of these being grammatical gender. However, even though these studies show substantial effects, they are inconsistent as to what they look like. Moreover, there has been no theoretical explanation for what these effects reflect in terms of language processing: what content is predicted and what happens when we encounter an unexpected grammatical gender input? Substantial progress can be made in our understanding of the contents of and the exact signature underlying the process of lexical prediction, in further research targeting grammatical gender manipulations.

Notes

1. Moreover, none of the studies explicitly refer to their effect as N400; Wicha, Moreno et al., (2003) describe the effect they observe as an "N400-*like* component". Otten et al. (2007) state that their observed effect *possibly* has similar neural generators as N400.
2. Other studies in Dutch use gender-marked adjectives as their measuring point; we consider this a difficult test case as the relevant inflection is marked at adjective offset, making it difficult to pinpoint at what point in time the reader/listener notices the unexpected input.
3. The 50% critical trial loss criterion was applied separately for the analysis of determiner trials and for the analysis of noun trials. The number reported here applies to the analysis on the determiners. In addition, two participants were excluded from the noun analysis.
4. Originally, it was planned to collect data with this design from non-native (second language) speakers of Dutch and compare the two groups. In order to make processing easier for the non-native speakers, we opted for a slower presentation rate.
5. A reviewer raised a concern in relation to our procedure for ocular artefact rejection; we departed from O&vB by manually rejecting trials contaminated with eye-movements, rather than using an ICA procedure. In order to address this concern, we performed an additional analysis of the data using ICA for eye-movement artefact rejection instead of manual rejection. This additional analysis thus fully followed the pre-processing steps described in O&vB. The results of these analyses were not different from the results reported in the present version of the manuscript. Please find a detailed description of this additional analysis in the supplemental online materials of this manuscript.
6. Although the analysis of ERPs on the nouns was not reported in O&vB, it was reported in the section devoted to the same experiment in the PhD thesis of Marte Otten (Otten, 2008; chapter 4). They look at the N400 in the time-window 300–500ms after noun onset and observe a similar result as reported here.

Acknowledgments

The authors are immensely grateful to Mante Nieuwland for his helpful advice and fruitful discussions during the data analysis and manuscript preparation stages of this research. We would also like to thank Joost Rommers for commenting on a draft of this paper. Finally, we would like to thank Birgit Knudsen for her help with data collection.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi:10.1016/S0010-0277(99)00059-1

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. doi:10.1016/j.jml.2016.10.002

Chow, W.-Y., & Chen, D. (2018). *Listeners rapidly use unexpected information to update their predictions: Evidence from eye-movements*. Poster presented at the 31st annual CUNY sentence processing conference, Davis, CA, USA.

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852–1870. doi:10.1016/j.neuropsychologia.2012.04.011

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. (2017). Comprehenders rationally adapt semantic predictions to the statistics of the local environment: A Bayesian model of trial-by-trial N400 amplitudes. In *Proceedings of the 39th annual conference of the cognitive science society,* London, UK.

DeLong, K. A., Urbach, P. T., & Kutas, M. (2017). *Concerns with Nieuwland et al. 2017*. Retrieved from http://kutaslab.ucsd.edu/FinalDUK17Comment9LabStudy.pdf

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. doi:10.1038/nn1504

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand: Processing in L2 Spanish. *Studies in Second Language Acquisition*, *35*(2), 353–387. doi:10.1017/S0272263112000915

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. doi:10.1006/jmla.1999.2660

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146. doi:10.1111/1469-8986.3920133

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

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. doi:10.1016/j.jml.2015.10.007

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. doi: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. doi:10.1080/23273798.2017.1323112

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.1080/01621459.1995.10476572

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112. doi:10.1162/jocn_a_00148

Kochari, A. (2015). *Mediating factors in predictive language processing: An EEG study on the effects of working memory, inhibitory control and processing speed* (Master's Thesis). Utrecht University. Retrieved from https://dspace.library.uu.nl/handle/1874/318532

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. doi:10.1080/23273798.2015.1102299

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342. doi:10.1016/j.cognition.2014.05.017

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. doi: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. doi:10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. doi:10.1038/307161a0

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. doi:10.1016/j.jml.2009.06.004

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. doi:10.1162/jocn_a_00328

Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2016). Spatiotemporal signatures of lexical–semantic prediction. *Cerebral Cortex*, *26*(4), 1377–1387. doi:10.1093/cercor/bhu219

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. doi:10.1111/psyp.12639

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

Nieuwland, M. (2015). The truth before and after: Brain potentials reveal automatic activation of event knowledge during sentence comprehension. *Journal of Cognitive Neuroscience*, *27*(11), 2215–2228. doi:10.1162/jocn_a_00856

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

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*. doi:10.1155/2011/156869

Otten, M. (2008). *Discourse-based lexical anticipation: The nature and contextual basis of predictions in language comprehension* (PhD Thesis). University of Amsterdam.

Otten, M., Nieuwland, M. S., & Van Berkum, J. J. (2007). Great expectations: Specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*(1), 89. doi: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. doi:10.1080/01638530802356463

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. doi:10.1016/j.brainres.2009.07.042

R Development Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria*. doi:10.1038/sj.hdy.6800737

Rommers, J., & Federmeier, K. D. (2018). Electrophysiological methods. In A. M. B. de Groot, & P. Hagoort (Eds.), *Research methods in psycholinguistics and the neurobiology of language: A practical guide* (pp. 247–265). Hoboken: John Wiley & Sons.

Schlueter, Z., Namyst, A., & Lau, E. (2018). *Predicting discourse status: N400 effects of determiner expectation*. Poster

presented at the 31st annual CUNY sentence processing conference, Davis, CA, USA.

Shetter, W. Z. (1959). The Dutch diminutive. *The Journal of English and Germanic Philology*, *58*(1), 75–90. doi:10.2307/27707226

Simonsohn, U. (2015). Small telescopes. *Psychological Science*, *26*(5), 559–569. doi:10.1177/0956797614567341

Szewczyk, J. M. (2018). P*rediction-inconsistent information leads to prediction updating – an ERP study on sentence comprehension*. Talk given at the 31st annual CUNY sentence processing conference, Davis, CA, USA.

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. doi:10.1016/j.jml.2012.12.002

Tanenhaus, M. K., Spivey-knowlton, M. J., Eberhard, K. M., & Sedivy, C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268* (5217), 1632–1634. doi:10.1126/science.7777863

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. doi:10.1037/0278-7393.31.3.443

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

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. doi:10.1037/a0036731

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. doi: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. doi:10.1016/S0010-9452(08)70260-0

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. doi:10.1162/0898929041920487

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. doi:10.1016/j.cortex.2015.03.014

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (Or Not) during language processing: A commentary On Nieuwland et al. (2017) and Delong et al. (2005). *bioRxiv*, 143750. doi:10.1101/143750