Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (in press). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition.*

Definitely saw it coming? The dual nature of the pre-nominal prediction effect

Damien S. Fleur[1], Monique Flecken[1,2], Joost Rommers[2,3] & Mante S. Nieuwland[1,2,4]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[2] Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

[3] School of Psychology, University of Aberdeen, UK

[4] Heinrich-Heine-University, Düsseldorf, Germany

Correspondence details:
Mante S. Nieuwland
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: mante.nieuwland@mpi.nl
Phone: +31-24-3521911

Author Note: In accordance with the Peer Reviewers' Openness Initiative (https://opennessinitiative.org, Morey, Chambers, Etchells, Harris, Hoekstra, Lakens, et al., 2016), all materials and scripts associated with this manuscript are available on OSF project "Article Gender & Definiteness" at https://osf.io/6drcy/.

ABSTRACT

In well-known demonstrations of lexical prediction during language comprehension, pre-nominal articles that mismatch a likely upcoming noun's gender elicit different neural activity than matching articles. However, theories differ on what this pre-nominal prediction effect means and on what is being predicted. Does it reflect mismatch with a predicted article, or 'merely' revision of the noun prediction? We contrasted the 'article prediction mismatch' hypothesis and the 'noun prediction revision' hypothesis in two ERP experiments on Dutch mini-story comprehension, with pre-registered data collection and analyses. We capitalized on the Dutch gender system, which marks gender on definite articles ('de/het') but not on indefinite articles ('een'). If articles themselves are predicted, mismatching gender should have little effect when readers expected an indefinite article without gender marking. Participants read contexts that strongly suggested either a definite or indefinite noun phrase as its best continuation, followed by a definite noun phrase with the expected noun or an unexpected, different gender noun phrase ('het boek/de roman', the book/the novel). Experiment 1 (N=48) showed a pre-nominal prediction effect, but evidence for the article prediction mismatch hypothesis was inconclusive. Informed by exploratory analyses and power analyses, direct replication Experiment 2 (N=80) yielded evidence for article prediction mismatch at a newly pre-registered occipital region-of-interest. However, at frontal and posterior channels, unexpectedly definite articles also elicited a gender-mismatch effect, and this support for the noun prediction revision hypothesis was further strengthened by exploratory analyses: ERPs elicited by gender-mismatching articles correlated with incurred constraint towards a new noun (next-word entropy), and N400s for initially unpredictable nouns decreased when articles made them more

predictable. By demonstrating its dual nature, our results reconcile two prevalent explanations of the pre-nominal prediction effect.

INTRODUCTION

Language comprehenders sometimes anticipate upcoming words based on the meaning of a story or conversation. Particularly informative in tracking the relevant anticipatory processes are event-related brain potentials (ERPs) recorded from the scalp. The ERP signal in response to words consists of various components including the N400 reflecting semantic processing (Kutas & Hillyard, 1980, 1984; for review, see Kutas & Federmeier, 2011) and post-N400 positivities in response to unexpected words that disconfirm likely expectations (for review, see Van Petten & Luka, 2012). Arguably the strongest evidence for word anticipation comes from studies using pre-nominal manipulations, which measured behavioral or neural responses to an article or adjective appearing before a noun (for review, see Kutas, DeLong & Smith, 2011; Van Berkum, 2009). Most studies of this type use gender-marking of pre-nominal articles, such as in Spanish and Dutch, and report differential event-related potential (ERP) responses to articles that mismatch the gender of a highly predictable noun, compared with gender-matching articles (e.g., for Dutch, Otten and Van Berkum, 2009; Van Berkum et al., 2005; for Spanish, Foucart, Martin, Moreno & Costa, 2014; Gianelli & Molinaro, 2018; Martin, Branzi & Bar, 2018; Molinaro, Gianelle, Caffarra & Martin, 2017; Wicha, Bates, Moreno & Kutas, 2003; Wicha, Moreno & Kutas, 2003, 2004). Of particular relevance is a study by Otten and Van Berkum (2009), wherein participants read two-sentence mini-stories that contained an article-adjectives-noun combination of which the noun was either predictable (e.g., "de verfijnde maar toch opvallende ketting", the_{com} sophisticated yet striking $\text{necklace}_{\text{com}}$) or not predictable and of a different gender than the predictable noun (e.g., "het verfijnde maar toch opvallende collier", the_{neu} sophisticated yet striking $\text{collar}_{\text{neu}}$). The gender-mismatching articles elicited an N400-like differential ERP effect,

which was not observed for the same article-adjective-noun combinations in non-constraining contexts featuring the same content words. Given that this was a comparison between words that were grammatical and did not differ in meaning, the observed effect must be ascribed to the grammatical relation between the presented pre-nominal article and the predicted - but not yet presented - noun.

Although the available literature supports noun prediction, the precise functional significance of 'pre-nominal prediction effects' remains unclear. A minimal interpretation, which we dub the 'noun prediction revision hypothesis', is that people predict the noun (with or without its gender) and then use article gender, once available, to revise the noun prediction (e.g., Van Berkum et al., 2005 Otten & Van Berkum, 2009; see also Otten, Nieuwland & Van Berkum, 2007; Otten & Van Berkum, 2008). However, a stronger claim has been made, namely that people predict a specific article-noun combination including the gender-marked form of the article itself (Kutas, DeLong & Smith, 2011; Wicha et al., 2003b, 2004; DeLong et al., 2005; see also Dell & Chang, 2014; Van Petten & Luka, 2012). In what we dub the 'article prediction mismatch hypothesis', the pre-nominal prediction effect reflects processing of the mismatch between the predicted and encountered article. We contrasted these hypotheses in two ERP studies on Dutch mini-story comprehension, with pre-registered data collection and analyses. We capitalized on the Dutch gender system, which marks gender on definite articles ('de' for common gender, 'het' for neuter gender) but not on indefinite articles ('een'). Our rationale was that if articles themselves are predicted, as assumed by the article prediction mismatch hypothesis, then the gender manipulation should have little effect when readers expected an indefinite article without gender marking.

Along with different interpretations of pre-nominal prediction effects, there is also inconsistency in the type of effect that has been observed empirically. In the first published study with a pre-nominal gender manipulation (Wicha et al., 2003a), Spanish speakers listened to sentence pairs in which a predictable noun or an incongruent noun was replaced with a drawing. The authors observed greater N400 amplitude for gender-marked pre-nominal articles that mismatched the gender of the predictable nouns, compared to articles with matching gender. A follow-up study with written materials and line drawings (Wicha et al., 2003b) also obtained an N400 effect of gender-mismatch. In a follow-up with fully written sentences and no line drawings (Wicha et al., 2004), gender-mismatching articles now elicited a P600 effect[1], which was interpreted as indicating an article-noun agreement violation.

The first Dutch study with a pre-nominal manipulation did not use articles but adjectives (Van Berkum et al., 2005). The participants listened to mini-stories that contained either a highly predictable noun or a different-gender, unpredictable noun. The nouns were preceded by adjectives that were gender-marked (using the adjectival suffix rule that adds '-e' to neuter nouns) in agreement with the upcoming noun. Time-locked to inflection-onset, gender-mismatches elicited an early positivity between 50 and 250 ms compared to gender-matches (however, see Nieuwland, Arkhipova, & Rodríguez-Gómez, 2020, for a failure to replicate this positivity in a large-scale, pre-registered study). Two follow-up studies with the same

[1] This P600 effect appears to be a unique observation, however, as recent studies with written Spanish sentences show predominantly N400-like effects in relation to gender-mismatching articles, i.e. enhanced negativities in the typical N400 time window (Foucart et al., 2014; Martin et al., 2018; Molinaro et al., 2014). The reported N400-like effects do seem to differ, at least visually, from typical N400 effects elicited with predictable vs unpredictable *nouns* with respect to latency and scalp distribution.

manipulation (Otten et al., 2007; Otten & Van Berkum, 2008) reported different ERP effects. In a study with spoken stories (Otten et al., 2007), gender-mismatching adjectives elicited a negative, right-frontal ERP effect between 300 and 600 ms after adjective-onset. In a study with written materials (Otten & Van Berkum, 2008), gender-mismatches elicited a late negativity ERP effect at 900-1200 ms after adjective-onset. In the Otten and Van Berkum (2009) study discussed previously, a negativity was observed in the 200-600 ms time window at right-frontal electrodes, which grew in size over time. Finally, Kochari & Flecken (2018), using the Otten and Van Berkum (2009) materials but omitting the non-constraining contexts, did not obtain a statistically significant effect of gender-mismatch. Mismatching articles did elicit a slowly developing negative shift compared to matching articles, over posterior electrodes instead of frontal electrodes. The observed pattern was consistent with that in the original data in terms of effect size (leaving aside differences in scalp-distribution), but a Bayesian analysis supported neither the null-hypothesis (no prediction effect) nor the alternative hypothesis (the effect size reported by Otten and Van Berkum).

Two open questions

The current study aimed to answer the following questions. First, do people only predict a noun and then use the information that the article provides to revise their prediction, or do they predict the specific article *itself* (which is marked for gender and for definiteness), along with the meaning and form of the noun? Following the noun prediction revision hypothesis (Van Berkum et al., 2005), the initial prediction could be limited to a specific noun meaning (with or without

activation of gender information[2]). Once the article is presented, the available gender information can be used to revise the prediction[3]. This revision minimally involves registering the prediction as being no longer viable and needing reconsideration. It could also involve an actual change to the noun prediction (e.g., a suppression or reduction of the original prediction, or a switch to a different noun prediction), although we refrain from strong claims at this point. Some of our exploratory analyses do speak to this issue.

The article prediction mismatch hypothesis instead assumes prediction of the article itself (see DeLong et al., 2005; Kutas et al., 2011; Wicha et al. 2003b, 2004; for discussion, see Ito et al., 2017c). Gender and definiteness information becomes activated before the article appears, for example as part of a lexical pre-activation process where people access syntactic and semantic information associated with a specific word form. The main difference between these hypotheses, therefore, is whether or not people predict a specific article word *form* (i.e., a lexical prediction).

The observation of gender-mismatch effects on adjectives (Van Berkum et al., 2005), which are less predictable than articles, suggests that prediction of a pre-nominal word-form is not required to elicit an effect (consistent with the noun prediction revision hypothesis).

---

[2] A parallel can be drawn to the literature on activation of gender information during word production. Some models of production argue that gender information is activated when people access a specific word meaning, a lemma (lexical access), but other models argue that gender information is only activated when people access a phonological form, the lexeme (since that form may depend on gender; for discussion, see Caramazza, 1997; Roelofs, Meyer & Levelt, 1998; Schiller & Caramazza, 2006; Schriefers & Jescheniak, 1999).

[3] This hypothesis, originally coined as a potential explanation of pre-nominal prediction effects reported by Van Berkum et al. (2005), is similar in spirit to recent 'prediction updating' proposals about the functional significance of the N400 component (Rabovsky, 2020; Szewczyk, & Wodniecka, 2020). However, these proposals take N400 amplitude to index change in a semantic feature-based probabilistic representation of sentence meaning, and do not assume prediction of word form.

However, those effects are not consistent across studies and do not seem to involve modulation of the N400 (for discussion, Ito et al., 2017a,c), which leaves open the possibility that pre-nominal N400 effects do reflect mismatch with a lexical prediction. We emphasize that previously reported effects of article gender-mismatch are in principle consistent with both prediction of articles (e.g., Wicha et al., 2004, 2004; see also DeLong et al., 2005) or only prediction of nouns (e.g., Otten et al., 2007; Otten & Van Berkum, 2008, 2009; Van Berkum et al., 2005). In the current study, we tried to tease apart these hypotheses by testing for gender-mismatch effects on articles that *themselves* were either expected or unexpected in terms of another feature: definiteness.

In addition to investigating what is predicted, our second question asks what the role of definiteness is in the pre-nominal prediction effect. In languages that mark both gender and definiteness on the article (e.g., Dutch and Spanish), the article contains grammatical information and semantic/referential information that is relevant to interpretation (e.g., Abbott, 2004, 2006; Frazier, 2006; Heim, 1982). Previous experiments on Spanish have compared gender-matching and -mismatching articles that are both either definite or indefinite. In Dutch, however, definite articles are gender-marked while indefinite articles are not, which is why Otten and Van Berkum (2009) and Kochari and Flecken (2018) only used definite articles. Both the Spanish and the Dutch studies used a sentence completion procedure to establish predictability, in which participants completed sentences truncated before the article, but scored cloze values in different ways: cloze values in the Spanish studies directly reflected the obtained article-noun responses, whereas cloze values in the Dutch studies discounted the articles and only reflected the noun responses. In the Dutch cloze values, the gender-manipulation with definite articles was

implemented for sentence contexts where most completions involve an indefinite article, at least for some items[4]. Therefore, some of the articles in Otten & Van Berkum (2009) and Kochari & Flecken (2018) were probably unexpected or infelicitous because of their definiteness, regardless of gender. The contexts that license the introduction of a novel definite referent are more limited or restricted than those that license the introduction of novel indefinite reference (for discussion, see Abbott, 2004, 2006; Clifton, 2013; Fraurud, 1990, Frazier, 2006; Heim, 1982; Singh, Fedorenko, Mahowald, & Gibson, 2016), and definite reference is more commonly used for previously mentioned referents than for new referents. Unexpected or infelicitous definiteness of the article may itself increase N400 amplitude (e.g., Kirsten, Tiemann, Seibold, Hertrich, Beck & Rolke, 2014; Schlueter, Namyst & Lau, 2018; see also Anderson & Holcomb, 2005; Schumacher, 2009). This could indicate that (in)definiteness conveys meaning and therefore results in additional semantic processing, or even that people predict the definiteness of upcoming referents or perhaps have difficulty integrating the article into an event-based representation of the discourse context (a 'situation model'; Zwaan et al., 1995). The results of Otten and Van Berkum (2009) and Kochari and Flecken (2018) thus reflect an unknown mix of effects associated with gender and definiteness, meaning that it is unclear whether people in fact predict pre-nominal lexical material. The current study therefore manipulated discourse contexts

[4] We were unable to obtain the raw cloze responses from Otten & Van Berkum (2009). In Kochari & Flecken (2018), which used materials based on the Otten and Van Berkum study, 45% of all cloze responses contained an indefinite article (97 out of the total number of 112 items contained at least one response with an indefinite article, and in 62 out of all items, more than half the responses contained the indefinite 'een').

to be constraining towards a definite or indefinite referent, and tested for gender-mismatch effects on definite articles that were either expectedly or unexpectedly definite.

Examining these issues in Dutch could therefore provide insights into the consistency of article gender-mismatch effects across languages. Qualitatively different ERP effects of gender mismatch have been observed. This variability may signal something meaningful like cross-linguistic differences or differences associated with specific methodological choices, it may signal random fluctuations (noise), or an unknown mix of the above (for discussion, see Ito et al., 2017c). As discussed previously, almost all the studies with Romance languages such as Spanish, Catalan or Italian report N400 effects (e.g., Wicha et al., 2003b; Foucart et al., 2014; Martin et al., 2018; Molinaro et al., 2014), and one observed a P600 effect (Wicha et al., 2004). In addition, there are two Dutch studies reporting 'N400-like' effects with different scalp distributions (Kochari & Flecken, 2018; Otten & Van Berkum, 2009), which are most relevant to the current study. We believe there is reason to doubt that the patterns observed in these two Dutch studies are truly generated by the article[5]. It is currently unclear why Kochari and Flecken (2018) and Otten and Van Berkum (2009) report different ERP results. However, it should be noted that the Dutch definite articles 'de' and 'het', besides signalling a singular noun of

[5] In both studies, the gender-match and -mismatch conditions start to diverge as early as 0 ms after article onset, and continue to diverge into the later time windows. Given that an effect as early as that is physiologically implausible, an alternative explanation is that these effects reflect a slow signal drift associated with voltage differences in the baseline period. In other words, it is not clear to what extent the obtained effects are truly generated by the article (see also Ito et al., 2017a,c; Nieuwland et al., 2018), and whether the obtained effects would hold when a countermeasure is performed to deal with the potential baseline problem (e.g., applying a 0.1 Hz filter instead of the 0.03 Hz filter, or applying a post-onset baseline correction; see also Ito et al., 2017b).

common or neuter gender, can also signal plurals and diminutives, irrespective of gender. As such, it is possible that the effect of gender-mismatch in Dutch was diluted by items where the article itself was unexpectedly definite. This would not have occurred in the studies with languages like Spanish or Italian, which have gender-marking on definite and indefinite articles and separate marking for plurality. The present study explicitly manipulated the expected definiteness of the articles.

The current study

In the current ERP study, we investigated lexical prediction during Dutch mini-story comprehension in order to address two outstanding questions on ERP effects associated with a gender-mismatch between a pre-nominal article and a predictable noun. We asked (1) whether such effects reflect article prediction mismatch or noun prediction revision, and (2) whether gender-mismatch in Dutch elicits enhanced N400 amplitude once definiteness is controlled for.

Our participants read two-sentence mini-stories in four different conditions (see Table 1 for an example item), with the critical noun phrase embedded in the second sentence. Each participant read one of two contexts that suggested a specific noun as its best continuation as part of either a definite noun phrase or an indefinite noun phrase (as established in a cloze task, see Methods). Each context was followed by a definite noun phrase, which either contained the predictable noun or an unpredictable, different-gender noun. Because half of the stories contained unexpectedly definite articles, we included filler stories with predictable indefinite noun phrases, such that a mismatch of the expected definiteness was as common as a mismatch of the expected gender.

Table 1. Dutch example mini-story in each of the four conditions, plus approximate translation. The entire set of materials is available on osf.io/6drcy

Article	**Context**	**Critical noun phrase**		**ending**
		Gender-match	**Gender-mismatch**	
Expectedly definite	Het is zondagochtend. De gehele gelovige familie gaat zoals altijd naar *It is Sunday morning. The whole religious family goes, as always, to*	de$_{com}$ kerk$_{com}$ *the$_{com}$ church$_{com}$*	het$_{neu}$ gebedshuis$_{neu}$ *the$_{neu}$ worship place$_{neu}$*	in het dorp. *in the village.*
Unexpectedly definite	Mijn moeder is erg gelovig. Op vakantie gaan we altijd direct op zoek naar *My mother is strongly religious. When on vacation, we always look directly for*			in de stad. *in the city.*

Our study was not a direct replication attempt of Otten and Van Berkum (2009), nor of Kochari and Flecken (2018), as our experimental design and analyses were different. Our primary dependent variable was N400 amplitude, defined as the average voltage value in the 300-500 ms time window after word onset at a centroparietal electrode selection. We defined additional dependent variables for anterior electrodes and for the subsequent time-window (500-700 ms) to capture later activity like extended N400 effects, the Post-N400 Positivity (PNP) or P600 (DeLong, Quante, & Kutas, 2014; Nieuwland et al., 2019; Van Petten & Luka, 2012). We

predicted that article gender-mismatch would elicit enhanced N400 amplitude compared to gender-match, like the patterns observed in Spanish and Italian (Wicha et al., 2003b; Foucart et al., 2014; Martin et al., 2018; Molinaro et al). This would be consistent with pre-activation of the noun, but would not suffice to conclude participants predicted article form. In addition, we predicted that unexpectedly definite articles would elicit enhanced N400 amplitude compared to expectedly definite articles (e.g., Schlueter et al., 2018).

Our central question was whether or not we would observe an interaction pattern. If people predict the articles themselves (DeLong et al., 2005; Wicha et al., 2004), we should observe an interaction effect: a gender-mismatch effect for expectedly definite articles but not for unexpectedly definite articles. If people do not predict the articles themselves, but merely use them to incrementally revise their prediction of the nouns, we would observe no interaction. We considered a third, hybrid option wherein people predict specific articles but also use gender information on unpredicted articles to inform their prediction, which would be supported by a gender-mismatch effect that was obtained for both expectedly and unexpectedly definite articles but that was larger for expectedly definite articles. Finally, we considered a fourth possibility, that the effect of gender-mismatch is qualitatively different for expectedly and unexpectedly definite articles (e.g., a P600 effect for expectedly definite articles and a N400 effect for unexpectedly definite articles, or vice versa), which would support a distinction between the processing of an article form prediction mismatch and the incremental use of gender information during predictive processing.

Experiment 1

METHODS

Participants

We recruited 48 participants[6] (17 males; mean age of 24 years, range 19-33) from the participant pool of the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. We did not perform an a priori power analysis to determine the required sample size, but we decided on a sample size that was a multiple of 4 and larger than previous studies at the time. All participants were native Dutch speakers, right-handed, with normal or corrected-to-normal vision and without a history of language impairment. After receiving information about the experimental procedures, participants gave informed written consent to take part in the experiment, which was approved by the Ethics Committee for Behavioural Research of the Social Sciences Faculty at Radboud University Nijmegen in compliance with the Declaration of Helsinki. Participants were paid for their participation (18 €). Participant data were excluded from further analysis based on criteria about the number of artefact-free trials (fewer than 25 trials in any of the conditions, or fewer than 30 trials on average across conditions) and the accuracy with which they answered the comprehension questions (<80% correct). We excluded and replaced 3 additional participants to achieve our sample size.

Materials

The final set of materials for this study was selected from a larger set based on specific constraints. We initially created a set of 280 items, of which each item contained two different versions of a Dutch mini-story. The two-sentence stories were written such that one version

[6] Our pre-registration was submitted after having tested five participants, and included our target sample size (N=48), experimental design and stimulus list assignment, data pre-processing, data exclusion and statistical analysis on AsPredicted.org via OSF (https://osf.io/mv4hw/). Analyses that were not pre-registered are labelled as exploratory.

presumably led people to expect a specific definite noun phrase (the ‘definite context’), and the other version presumably led people to expect that same noun as part of an indefinite noun phrase (the ‘indefinite context’). The definite and indefinite contexts sometimes differed in the number of words in the first sentence, but always contained the same number of words in the second sentence (i.e., the sentence position of the target words was matched between versions, but not between items). To establish whether the stories indeed were sufficiently constraining towards these noun phrases, we performed a cloze probability test in the form of an online questionnaire. All mini-stories were truncated before the target article. We created two lists of 280 stories such that each participant saw only one version of each item. Within each list, we randomized definite and indefinite contexts. We recruited 40 participants (20 per list) from the pool of participants of the Max Planck Institute of Psycholinguistics in Nijmegen who received a financial compensation (10€). They were instructed to read each mini-story in the order they were given, and complete each item with the continuation they would have expected. Participants were given an example of a mini-story with a possible ending that matched the structure of the test items. They received no specific instruction regarding the number of words to use but were asked to avoid repeating words over multiple stories, to not think too long about a specific story and to use whatever completion came to mind first.

From the obtained responses, we counted how often the expected article and/or noun was used. We also counted certain answers towards the target noun when the response had lexical overlap with and the same gender as the target noun (e.g. ‘de pc’ for ‘de computer’), when the response was a misspelling or differently-spelled version of the target noun, when the response was the diminutive version of a neuter-gender target noun (e.g. ‘het spelletje’ for ‘het spel’).

Cloze probability was calculated as the percentage of responses containing the target article or target noun. We selected the items in which each version had a cloze probability of at least 75% for the definite and indefinite target article and the noun, and where an unexpected article of the wrong gender or definiteness was never higher than 15%. For 89 items that did not make this selection, we rewrote one or both versions and performed a second cloze test with 20 participants who had not participated in the first cloze test, and we computed new cloze probability scores and again selected items that made the 75% cloze probability cut-off for both articles and the noun.

The final selection contained 160 items, with an average cloze value for the expected target article of 94% (SD = 7, range = 75-100) and 92% for the expected target noun (SD = 8, range = 75-100). Of note, gender was not fully balanced across items, because we had 99 target nouns of common gender (de-words) and only 61 of neuter gender (het-words). Cloze values for these article types are presented in Table 2. This disparity matches the relatively high frequency of de-words compared to het-words (Deutsch & Wijnen, 1985; Tuinman, 1996, quoted in Geerts, 1975; Van Berkum, 1997). We controlled for a potential effect of the article form in our main statistical analyses (and we report tests for potential processing differences between ‘de’ and ‘het’ in the appendix). On average, the target article was the 8th word in the second sentence (SD = 1.9, range = 3-13) and the target noun followed right after. Sentence position of the target article and noun was matched for the definite and indefinite context of each item.

Table 2. Cloze values (mean *M* and standard deviation *sd*) for 'de' and 'het'				
Definiteness	Gender	Article	*M (%)*	*sd*
Expected	Match	de	94.1	6.7
		het	91.4	6.9
	Mismatch	de	1.2	3.2
		het	1.4	3.5
Unexpected	Match	de	1.1	2.8
		het	0.5	1.8
	Mismatch	de	0.9	2.1
		het	1.4	3.0

For the ERP experiment, we created the gender-mismatch condition by replacing the target article-noun combination with an unexpected, different-gender article-noun combination[7] (Article: Mean = 1%, SD = 3, range = 0-15; Noun: Mean = 0%, SD = 2, range = 0-20). We selected mismatching nouns that we considered relevant and at least somewhat plausible or non-anomalous given the story context. Only after the second experiment did we obtain plausibility norms, which showed that on a scale from 1 to 5 from very implausible to very plausible, the

[7] There were a few exceptions where we used a diminutive noun (12 items, always preceded by the neuter gender article 'het', e.g., 'het bonnetje') or plural noun (1 item, 'de scherven'), and where the gender of the associated lemma was sometimes the same as that of the predictable noun. Importantly, these unexpected nouns were never the diminutive or plural form of the predictable noun.

average plausibility of mismatching nouns was 3.12, SD = .88, range 1.14-4.93 (this is further discussed in the section 'Exploratory tests for noun prediction revision'). The mismatching nouns were, on average, longer and less frequent than the matching nouns (Keuleers, Brysbaert & New, 2010), details can be found on our OSF page. To create the unexpectedly definite condition, we then replaced the expected indefinite article ('een') of each indefinite-context with a definite article of the correct gender. In addition, we added at least one and at most three words after the target noun, and this sentence-ending was identical for the definite and indefinite context.

In the experimental stories, expectations of an indefinite noun phrase were never met. To avoid that participants would pick up on this regularity (and therefore, possibly, would stop predicting indefinite noun phrases), we included 80 filler stories with a high-cloze indefinite noun phrase (Article: Mean = 94%, SD = 3, range = 75-100%; Noun: Mean = 91%, SD = 8, range = 75-100). The fillers were generated from the set of materials that did not make it into the experimental materials, and were of the same two-sentence form as our experimental materials (e.g., "*Lisa's dochter lijkt koorts te hebben. Om de temperatuur te meten leent ze bij de buurvrouw EEN thermometer voor kinderen",* approximate translation: 'Lisa's daughter seems to have a fever. To measure the temperature, she borrows from the neighbour A thermometer for children', critical article capitalized for demonstration purpose only). Due to the fillers, ERP participants saw the same ratio of unexpectedly definite articles and articles with an unexpected gender compared to expected articles, namely in a third of all stories.

In our experimental materials, we manipulated the two variables 'expected article definiteness' and 'article gender-match' in a 2 (Definiteness: expected, unexpected) by 2

(Gender: mach, mismatch) factorial design. We created 4 stimulus lists such that each participant saw 40 items from each of the 4 conditions, and each participant saw only one condition of an item, but across the lists each item was seen in each condition equally often. For each stimulus list, we generated two randomizations, to a total of 8 lists.

To encourage participants to pay attention to the meaning of the stories, they were asked to answer yes/no comprehension questions on 60 trials (i.e., 25% of all trials were followed by a question). These comprehension questions were roughly evenly spread across the experiment and separated from each other by at least two trials.

Procedure

Participants were seated before a monitor in a soundproof, electrically shielded room. Using a button box, participants could start each trial, which started with a fixation cross displayed at the centre of the screen, followed by the first sentence of a story shown in its entirety. Participants could press a button to start the second sentence, which was presented one word at a time at the centre of the screen. Word duration was 300 ms and was followed by a blank screen for 300 ms until the next appeared. If the story was followed by a comprehension question, participants were required to respond yes or no with the button box before the next trial started.

A brief practice session with five trials preceded the actual experiment, so that participants could get used to the procedure. The experiment was divided in six blocks with brief breaks in between.

EEG recording and data-processing

We recorded continuous EEG signal from 27 active scalp electrodes mounted in an elastic cap (ActiCap), placed according to the 10-20 convention and each referenced online to the left mastoid. An additional reference electrode was placed at the right mastoid. Furthermore, we recorded voltage at 4 EOG electrodes (above and under the left eye for the vertical dimension, next to the left and right eye for the horizontal dimension). The signal was amplified using BrainAmps amplifiers and recorded with Brain Vision Recorder (Brain Products, München) at 500 Hz with a band-pass filter at 0.016-150 Hz (time constant 10s).

We used BrainVision Analyzer for offline data processing. Following the pre-registration, we visually screened the data for bad channels (due to drifting, spiking, excessive line noise) and interpolated bad channels through spline interpolation. We then filtered the continuous data with a 0.1-100 Hz (24 dB/octave roll-off) band-pass filter, and we re-referenced all channels to the average of the left and right mastoid. We then epoched the data into segments from -500 to 1000 ms relative to target article or noun onset. We subsequently removed artefact-containing segments (i.e., containing large movement-related artefacts, large bursts of muscle activity, or amplifier blocking) after visual inspection. We then performed an ICA-based correction for blinks, eye movements, and steady muscle activity. After this, we applied a 30 Hz low-pass filter (24 dB), followed by a baseline correction to 200 ms before each critical word. Finally, we automatically rejected segments with values that exceed ±75 μV at any channel. In total, 4.4% of the epoched data was removed.

Statistical analyses

We performed linear mixed-effects analyses in R, with the two-level factors ‘definiteness’ (expected/unexpected) and ‘gender’ (match/mismatch). Definiteness refers to the

match between the definite article with the story context, as the articles were either expectedly definite or unexpectedly definite (high cloze values for definite or indefinite articles, respectively). Gender refers to whether the article matched the gender of the expected noun. We included an additional factor 'article' (de/het) to account for potential effects associated with the specific articles, which was important given the lexical differences between 'de' and 'het' ('de' is more frequent, and may elicit smaller N400s overall; Kutas & Federmeier, 2011), and given that most of our items had 'de' as the expectedly definite article. All three categorical variables were deviation-coded.

Using a spatiotemporal region-of-interest (ROI) approach, our main dependent measure (N400 amplitude) was the average voltage across six centro-parietal channels (Cz, CP1, CP2, P3, Pz, P4) in the 300–500 ms window after word onset for each trial. To evaluate effects at anterior electrodes, we also computed average voltage across six anterior electrodes (F3, Fz, F4, FCz, FC1, FC2). For both ROIs, we also computed average voltage in the 500-700 ms (post-N400) time window. For the articles, we performed analyses at both time windows in both ROIs, resulting in four analyses[8]. For the nouns, we only performed two analyses, namely on voltage in the 300-500 ms time window at the posterior ROI and the 500-700 ms time window at the anterior ROI. We evaluated the effect of 'definiteness' and 'gender' by performing model-comparison using chi-square goodness-of-fit tests.

[8] We also pre-registered secondary, distributional analyses of the article-elicited ERPs involving 4 electrode quadrants. Because the results of these analyses did not impact our main conclusions regarding the interaction between definiteness and gender, we do not report them in this paper but refer the interested reader to our OSF materials.

Following the recommendations of Barr, Levy, Scheepers, & Tily (2013), we first tried to fit the maximal random effect structure as justified by the design but simplified the random effect structure to deal with non-convergence. For the article and noun analyses, we included random intercepts for subjects and items and by-subject and by-item random slopes for 'gender'.

RESULTS

Pre-registered article-analyses

In accordance with our predictions, our experimental manipulations were associated with modulations of N400 activity, visible at posterior electrodes within the 300-500 ms time window after article onset (Figure 1; see also Supplementary Figure 1 and 2, where we plot ERPs at all individual channels). Our analyses yielded the following patterns (see Table 3, for details): Gender-mismatching articles elicited reliably more negative voltage (enhanced N400 activity) compared to gender-matching articles at the posterior ROI. This effect extended into the 500-700 ms time window[9], as also observed in previous studies with Spanish sentences (Martin et al., 2013, 2018; Foucart et al., 2014). Unexpectedly definite articles elicited more negative ERPs than expectedly definite articles at both ROIs and in both time windows, although this effect was strongest at the posterior ROI in the N400 time window, thus consistent with an N400 effect. The gender-mismatch effect was numerically larger for expectedly definite articles (-0.74 µV, *SE* = 0.27, Z = 2.69, p = 0.007[10]) than for unexpectedly definite articles (-0.35 µV, *SE* = 0.27, Z =

[9] Our analyses in the 500-700 ms time windows converged but revealed random effect correlations of ±1, indicating overfitting. Re-running these analyses after removing the relevant random slope did not meaningfully change the observed pattern of effects.

[10] These pairwise tests were not pre-registered but added upon reviewer request.

1.29, p = 0.20), but the results did not allow us to reject the hypothesis that these effects are similar. In the 500-700 ms time window, where we also obtained effects of gender and definiteness, there was no hint of an interaction pattern because the estimate for the interaction term was close to zero, and we observed gender-mismatch effects both for expectedly definites (-0.60 μV, SE = 0.29, Z = 2.06, p = 0.039) and for unexpectedly definites (-0.59 μV, SE = 0.29, Z = 2.04, p = 0.041). Figure 2 shows the scalp distribution of the article effects.

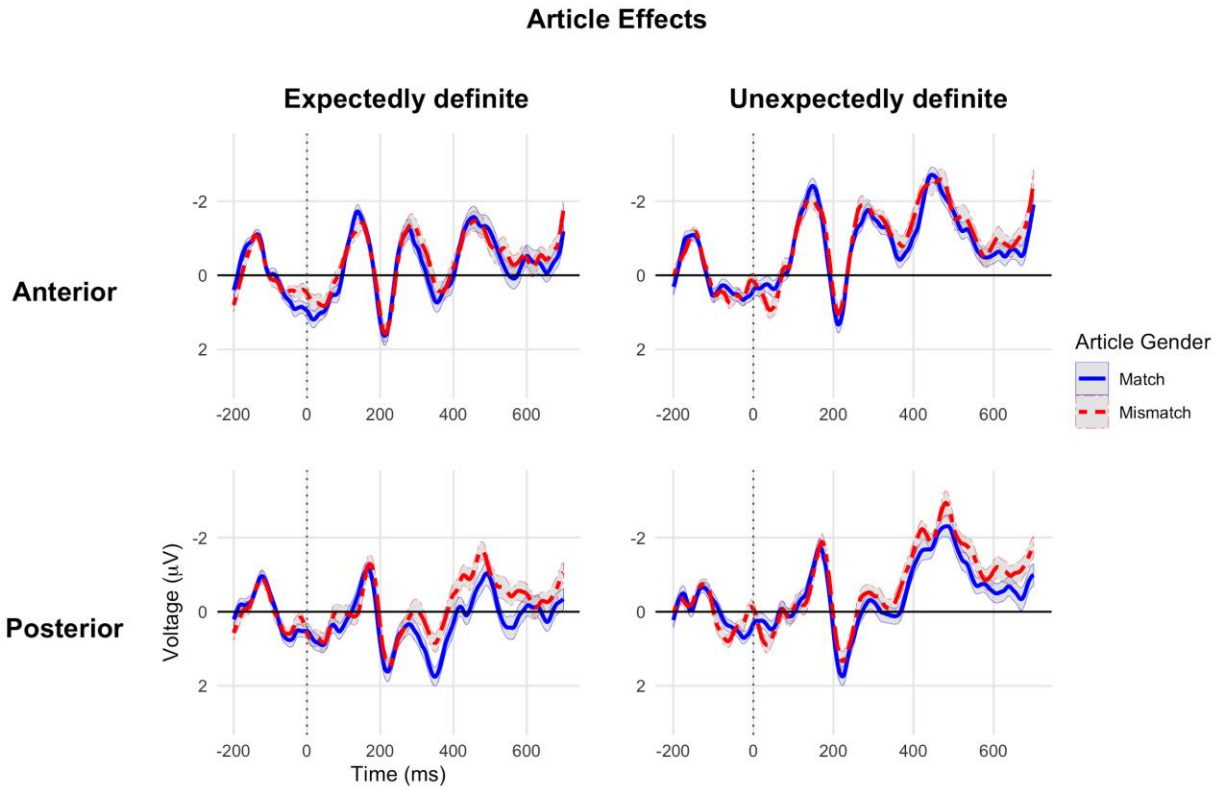


Figure 1. Article effects in Experiment 1. The graphs show the grand-average ERPs elicited by gender-matching articles (solid blue lines) and gender-mismatching articles (dotted red lines) at the pre-registered anterior and posterior ROIs (top and bottom graphs, respectively), when articles were expectedly and unexpectedly definite (left and right graphs, respectively). Grey-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 used linear mixed-effects to account for variance associated with different items and the two article forms ('de/het').

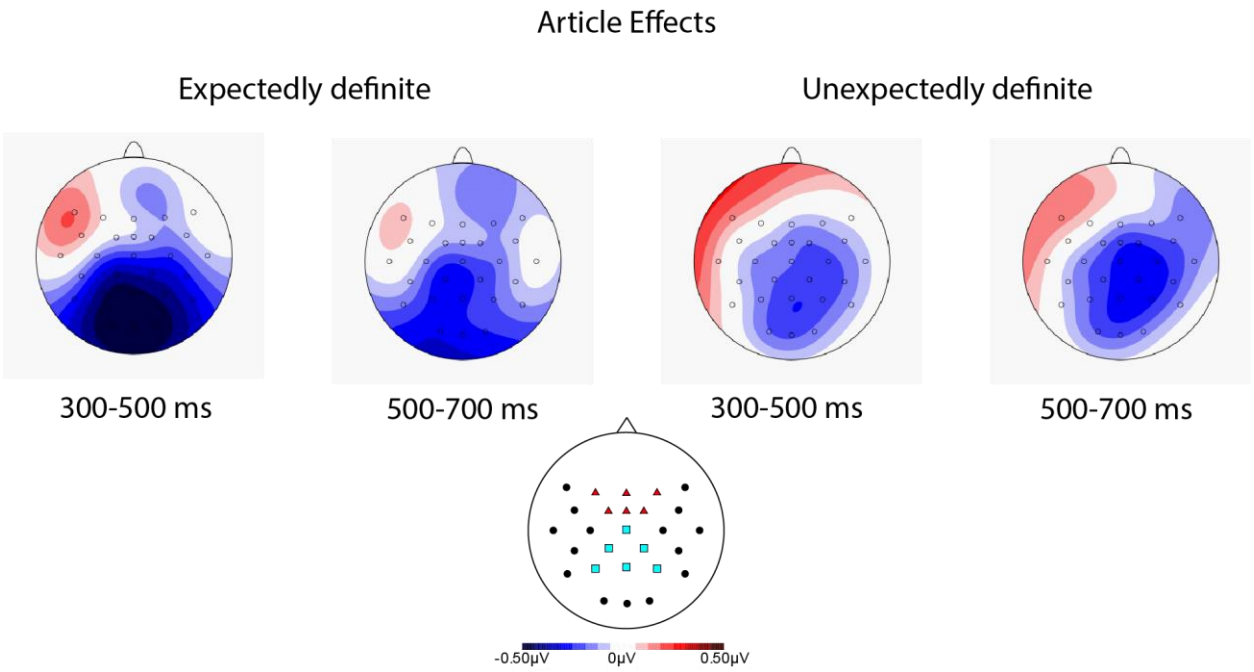


Figure 2. Scalp plots of the gender effects (mismatch minus match) for expectedly and unexpectedly definite articles in both time windows of analysis in Experiment 1. Blue squares and red triangles in the center head plot show the positions of the electrodes contained in the posterior and anterior ROI, respectively.

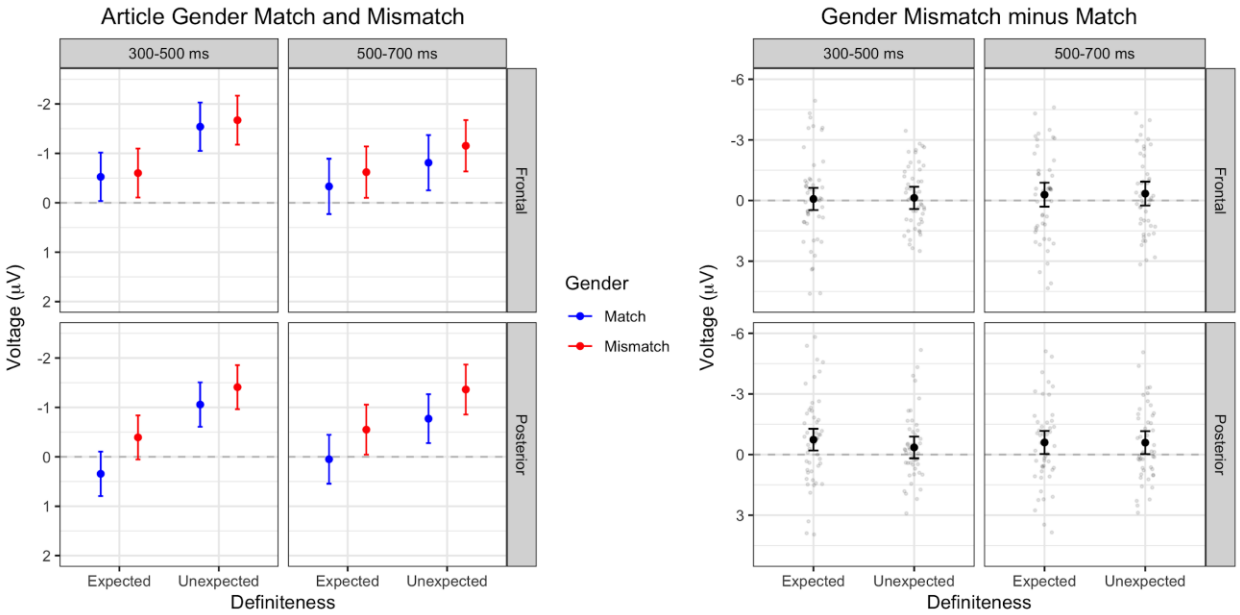


Figure 3. Article effects in Experiment 1. The left graphs show the estimated marginal means per condition from the mixed-effects model output for each ROI (large dots), along with the 95% confidence interval (vertical whisker). The right graphs show the corresponding estimated marginal means for the gender-mismatch effect (mismatch minus match), along with 95% confidence interval and subject-level mean effects.

Table 3. Results from the article-analyses in Experiment 1. For each spatial and temporal region-of-interest, the tables shows the estimated difference between the expected and unexpected conditions (unexpected minus expected), the associated 95% confidence interval, the $\square^2$ test-result and associated p-value (for details, see analysis files on https://osf.io/6drcy).

		Time window					
Factor	ROI	300-500 ms (N400)			500-700 ms (post-N400)		
		ß, CI	$\square^2$	*p*	*ß, CI*	$\square^2$	*p*
Gender	Anterior	-0.10, [-0.52,0.31]	0.56	0.45	-0.32, [-0.75,0.12]	1.96	0.16
	Posterior	-0.55, [-0.96,-0.14]	6.64	<0.01	-0.59, [-1.02,-0.17]	7.18	<0.01
Definiteness	Anterior	-1.04, [-1.42,-0.67]	30.18	<0.001	-0.51, [-0.90,-1.11]	6.22	0.01
	Posterior	-1.21, [-1.57,-0.85]	43.89	<0.001	-0.82, [-1.20,-0.43]	17.26	<0.001
Gender :Definiteness	Anterior	-0.06, [-0.80,0.69]	0.02	0.88	-0.05, [-0.85,0.74]	0.02	0.89
	Posterior	0.38, [-0.33,1.10]	1.11	0.29	0.01, [-0.76,0.78]	0.00	0.98

Pre-registered noun-analyses

As expected, prediction-mismatching nouns elicited more negative ERPs in the posterior ROI at 300-500 ms after noun onset, i.e., an N400 effect, compared to matching nouns (Figure 3; Supplementary Figures 3 and 4; Table 4), and more positive ERPs in the anterior ROI at 500-700 ms, although this later positive ERP effect appeared much weaker than the earlier N400 effect. In the posterior ROI at 300-500 ms, nouns following expectedly definite articles elicit more negative ERPs compared to nouns following unexpectedly definite articles. Finally, ERPs in the posterior ROI at 300-500 ms showed an interaction effect: the N400 effect of prediction mismatch was more pronounced for unexpectedly definite nouns (-1.58 μV, SE = 0.32, Z = 4.99, $p < 0.001$) than for expectedly definite nouns, (-0.72 μV, SE = 0.32, Z = 2.30, p = 0.02). The average voltages were less positive overall for nouns after expectedly definite articles (match, mean = 2.00 μV, *SE* = 0.39; mismatch, mean = 1.28 μV, *SE* = 0.41) than after unexpectedly definite articles (match, mean = 3.05 μV, *SE* = 0.39; mismatch, mean = 1.47 μV, *SE* = 0.41). The interaction pattern thus mostly resulted from the effect of definiteness on the matching nouns. Figure 4 shows that the N400 effect of prediction mismatch after expectedly definite articles had a slightly unusual frontal distribution.

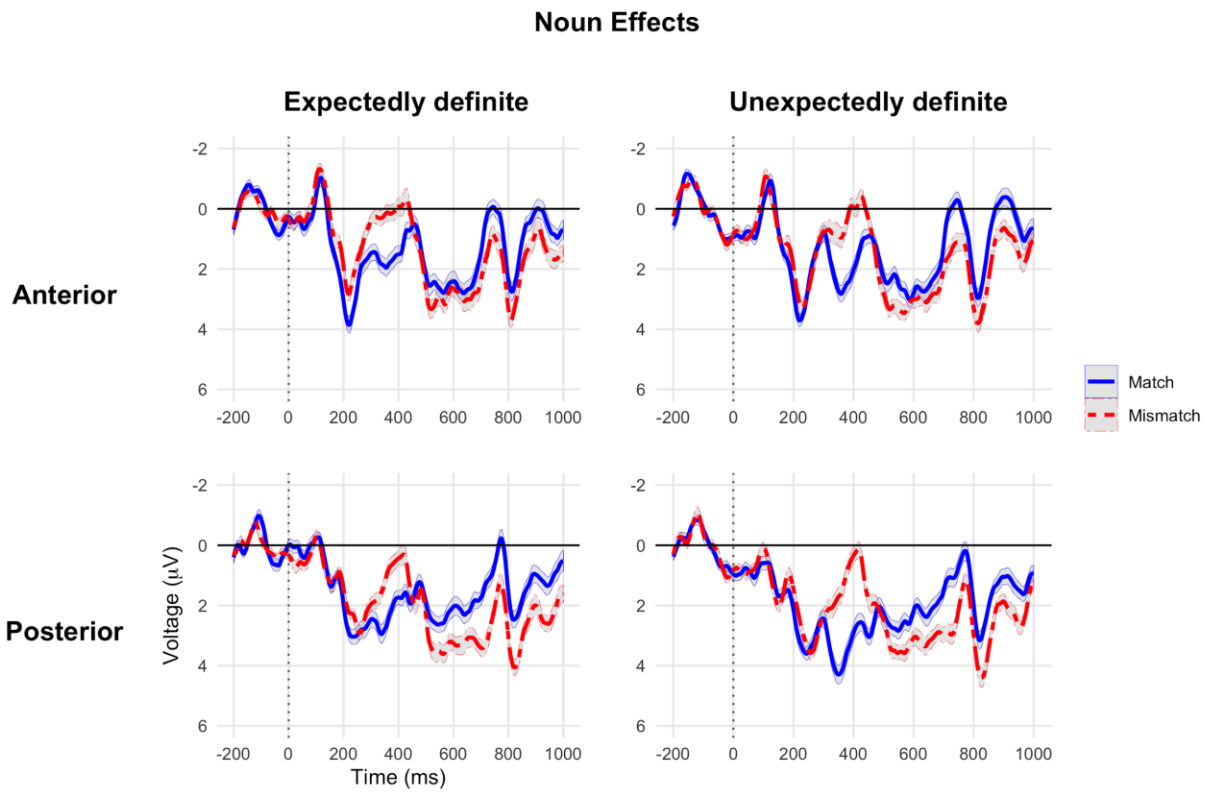


Figure 3. Noun effects in Experiment 1. The graphs show the grand-average ERPs elicited by prediction-matching nouns (solid blue lines) and prediction-mismatching nouns (dotted red lines) at the anterior and posterior ROIs (top and bottom graphs, respectively), following articles that were expectedly and unexpectedly definite (left and right graphs, respectively).

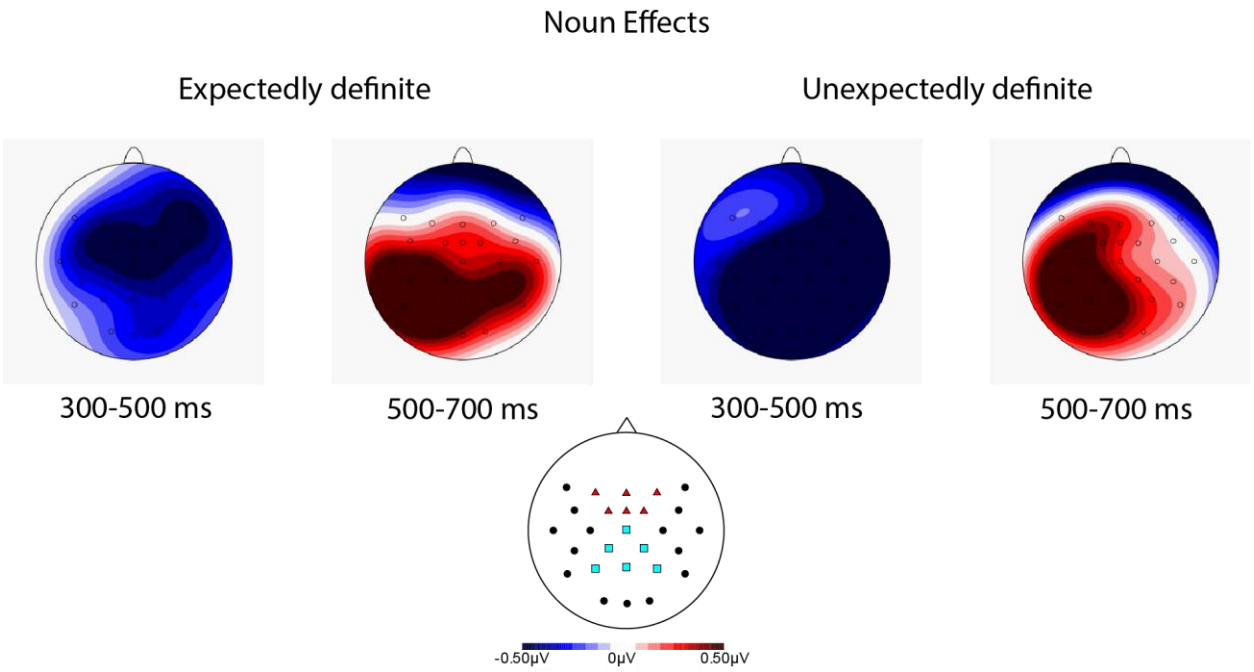


Figure 4. Scalp plots of the noun effects (prediction mismatch minus match) in Experiment 1.

Table 4. Results from the noun-analyses in Experiment 1.						
	ROI					
Factor	Posterior, 300-500 ms			Anterior, 500-700 ms		
	β, CI	□^{2}	*p*	*β, CI*	□^{2}	*p*
Mismatch	-1.15, [-1.65,-0.66]	18.39	<0.001	0.55, [0.02,1.08]	4.06	0.04
Definiteness	0.26, [0.24,1.01]	10.13	<0.01	0.02, [-0.38,0.43]	0.01	0.90
Mismatch:Definiteness	-0.85, [-1.62,-0.09]	4.78	0.03	0.09, [-0.73,0.90]	0.04	0.84

Exploratory article-analyses

Our pre-registered article-analyses yielded significant effects of gender-mismatch and definiteness, but non-significant *p*-values for the interaction between these factors. Thus, there was no evidence that expected definiteness modulated gender mismatch effects. However, this could reflect a lack of statistical power. In addition, we used an ROI analysis approach that, although justifiable a priori, may have missed relevant effects outside the ROI. Visual inspection of the gender-mismatch effect for expectedly definite articles showed strongest effects at occipital channels (Figure 2; Supplementary Figure 1), and perhaps a somewhat earlier onset than the gender-mismatch effect for unexpectedly definite articles (the latter effect seemed more salient in the second half of the N400 ROI than in the first half).

To address these concerns, we performed a mass mixed-effect regression analysis to determine where and when the interaction effect was strongest, and we used the results to pre-register a direct replication study with an additional ROI (Experiment 2). First, we downsampled the pre-processed, segmented data to 100 Hz to speed up the analysis. Then, for each sample between -200 to 1000 ms relative to article onset, and for each channel, we performed a mixed-effects model analysis using the 'lme4' package (Bates et al., 2014) as implemented in R (R Core Team, 2018). We used the same fixed and random effects as in the pre-registered analysis, but to speed up the analysis we did not include any random slopes. Full code for the entire analysis is available on our OSF page. For each model, we extracted a coefficient estimate with a standard error, a t-value and p-value associated with 'gender', 'definiteness', the interaction term 'gender:definiteness', and for the simple effects of gender mismatch within expectedly and unexpectedly definite articles. The results for the interaction term are plotted in Figure 5, which depicts where the mismatch effect (mismatch minus match) is bigger (yields more negative voltage) for expected definites than for unexpected definites. Of note, although Figure 5 marks the samples where the interaction term is statistically significant at alpha = 0.05, none of these samples survived correction for multiple comparisons (using Benjamini and Hochberg method as implemented in R's p.adjust, applied to p-values from samples in the 200-500 ms window, either across all channels or only posterior channels where N400 modulations are strongest). Nevertheless, the results did support the observation from the scalp distributions that the interaction effect (i.e., stronger gender-mismatch effect for expectedly definite articles than for unexpectedly definite articles) was stronger towards the back of the head (e.g., occipital channels) and most pronounced in the time window before 400 ms after article-onset.

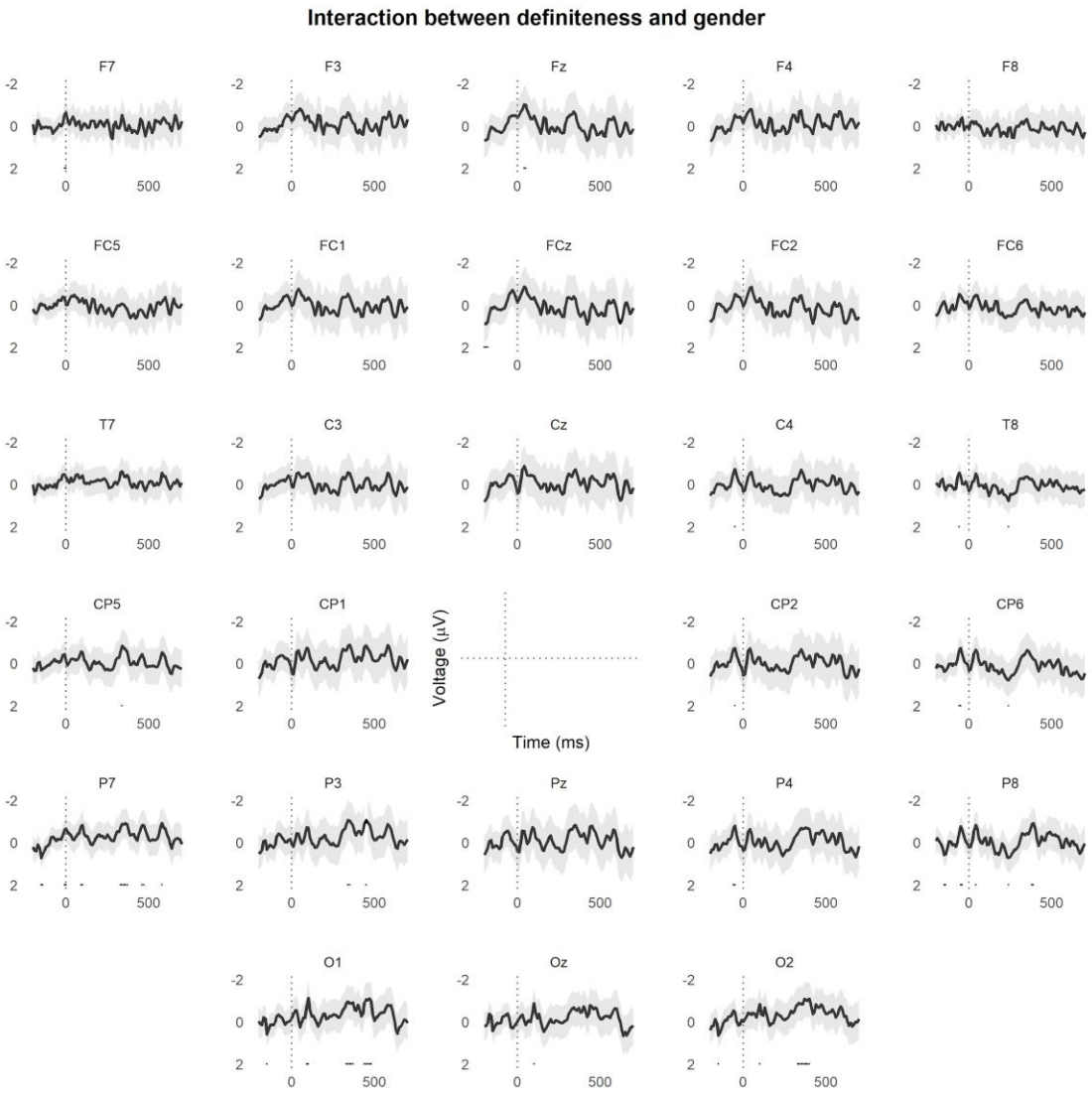


Figure 5. Results from the mass mixed-effects regression analysis for Experiment 1. Black lines represent the voltage associated with the interaction term. More negative voltage means that the mismatch effect (mismatch minus match) was larger (more negative) for expectedly definite than for unexpectedly definite articles. The grey area represents the 95% confidence interval, and the black dots underneath mark statistically significant samples (uncorrected; these samples were not statistically significant after correction). Our new ROI was selected based on the occurrence of significant samples at occipital channels in the 300-400 ms time window, as seen in this figure.

Discussion

Gender-mismatching articles elicited enhanced negativity in the 300-500 ms time window (i.e., increased N400 amplitude) compared to matching articles, consistent with several previous demonstrations of prediction of specific upcoming words (e.g., Foucart et al., 2014; Martin et al., 2018; Molinaro et al., 2018; Wicha et al., 2003a,b). This effect extended into the subsequent 500-700 ms time window.

Furthermore, unexpectedly definite articles elicited enhanced N400s compared to expectedly definite articles, consistent with a previous report by Schlueter et al. (2018; see also Kirsten et al. 2014). This effect also extended into the subsequent 500-700 ms time window. Unexpected definiteness thus seems to have repercussions for semantic processing. For example, it may cause enhanced semantic processing because it requires a change to the event-based representation of the discourse context (e.g., Clifton, 2013; Frazier, 2006; Zwaan & Radvansky, 1998).

Crucially, the gender-mismatch effect in the 300-500 ms time window was numerically larger for expectedly definite articles than for unexpectedly definite articles, consistent with the article prediction mismatch hypothesis, but statistically the evidence for this interaction effect in the 300-500 ms time window was inconclusive. In the 500-700 ms time window, however, the results suggested that expectedly and unexpectedly definite articles both elicited a gender-mismatch effect. To address possible concerns about our sample size and about the sub-optimal ROI for detecting the interaction effect, we performed a direct replication study.

Experiment 2

For Experiment 2, we pre-registered one additional ROI and a larger sample size, based on the results of the exploratory analysis of Experiment 1. The new ROI was based on where the interaction effect had seemed strongest, namely average voltage across occipital channels (O1/Oz/O2) in the 300-400 ms time window after article onset. In this ROI, a mixed-effect regression analysis on data from Experiment 1 showed a statistically significant interaction effect[11] (ß = 0.74 μV, CI = [0.14, 1.34], t = 2.48, p = 0.013). Because the obtained estimate is likely an overestimation of the true effect (e.g., Gelman & Carlin, 2014), we then performed a power analysis simulation with the SIMR package (Green & MacLeod, 2016) to estimate the required sample size to achieve 80% power for an effect of only 0.65 μV (1/8 smaller than the original effect of 0.74 μV; this specific value was chosen somewhat arbitrarily; the associated scripts are available on our OSF page). Based on the outcome of this analysis, we pre-registered a sample size of 80 participants (https://osf.io/9xm4g)

METHODS

We recruited 84 participants (24 males; mean age = 27 years, range 19-68) from the same participant pool and using the same criteria as used in Experiment 1, with the additional criterion that participants had not participated in Experiment 1. Four participants were excluded from the

[11] We emphasize that this exact value cannot be reproduced from our online materials, because it was performed before we noticed an error in the pre-processing of 3 participants from Experiment 1 (the right mastoid and right VEOG channel had been swapped during recording but not swapped back during pre-processing), which we have corrected in the available data. The corrected data thus also gave different output for a prior power at the pre-registered sample size of 80, namely 82.1% power to detect an effect of 0.65 μV, and 76.6% power to detect an effect that is ⅛ smaller (0.61 μV) than the effect obtained in Experiment 1. Because the change in a priori power was small, we decided to maintain the pre-registered sample size.

analysis based on the number of trials after artefact rejection, and were replaced by new participants to reach the pre-registered sample size of 80 participants.

Materials, procedure, data collection, pre-processing and statistical analysis were identical to Experiment 1. We pre-registered an additional, occipital ROI (average voltage across occipital channels O1/Oz/O2 in the 300-400 ms time window) for the analysis.

RESULTS

Pre-registered article-analyses

As in Experiment 1, our experimental manipulations were associated with modulations of activity in the 300-500 (N400) and 500-700 ms time window after article onset (Figure 6-8; see also Supplementary Figure 5-6, for ERPs at all individual channels). Our analyses yielded the following patterns (see Table 5 for details): ERPs at the posterior ROI showed the same patterns observed in Experiment 1. Gender-mismatching articles elicited reliably more negative voltage (enhanced N400 activity) compared to gender-matching articles in the 300-500 ms time window and this effect extended into the 500-700 ms time window. This was also the case for unexpectedly definite articles relative to expectedly definite articles. At both time windows, although the gender-mismatch effect was somewhat larger for expectedly definite articles (300-500 ms: -0.54 μV, SE = 0.19, Z = 2.80, p = 0.005; 500-700 ms: -0.62 μV, SE = 0.22, Z = 2.81, p = 0.005) than for unexpectedly definite articles (300-500 ms: -0.32 μV, SE = 0.19, Z = 1.63, p = 0.10; 500-700 ms: -0.50 μV, SE = 0.22, Z = 2.27, p = 0.023), we did not obtain convincing evidence for an interaction pattern.

At the anterior ROI in the 300-500 ms time window, we observed enhanced negativity for unexpected definiteness and not for gender-mismatch, like in Experiment 1. However, unlike

in Experiment 1, we observed an additional interaction pattern, with a statistically significant gender-mismatch effect for unexpectedly definite articles (-0.51 μV, SE = 0.22, Z = 2.28, p = 0.022) and not for expectedly definite articles (0.10 μV, SE = 0.22, Z = 0.47, p = 0.64). In the 500-700 ms time window, the results patterned with Experiment 1, with enhanced negativity for unexpected definiteness, but no clear effect of gender-mismatch or interaction.

Crucially, ERPs at the occipital ROI confirmed the interaction pattern we observed in Experiment 1, with a clear gender-mismatch effect for expectedly definite articles (-0.78 μV, SE = 0.16, Z = 4.89, p < 0.001) but not for unexpectedly definite articles (-0.05 μV, SE = 0.16, Z = 0.28, p = 0.77). The observed estimate for the interaction term (0.73 μV, SE = 0.22) was highly similar to that observed in Experiment 1 (0.69 μV, SE = 0.29)[12].

[12] Exploratory Bayesian mixed-effects model analyses showed that the credible interval for the occipital effect in Experiment 2 (b = 0.69, CrI = [0.20, 1.18]) fell entirely within the credible interval for this effect in Experiment 1 (b=0.63, CrI= [0.04, 1.23]), consistent with a practically equivalent effect (i.e. successful replication) from a Bayesian estimation perspective (Kruschke & Lidell, 2018a,b).

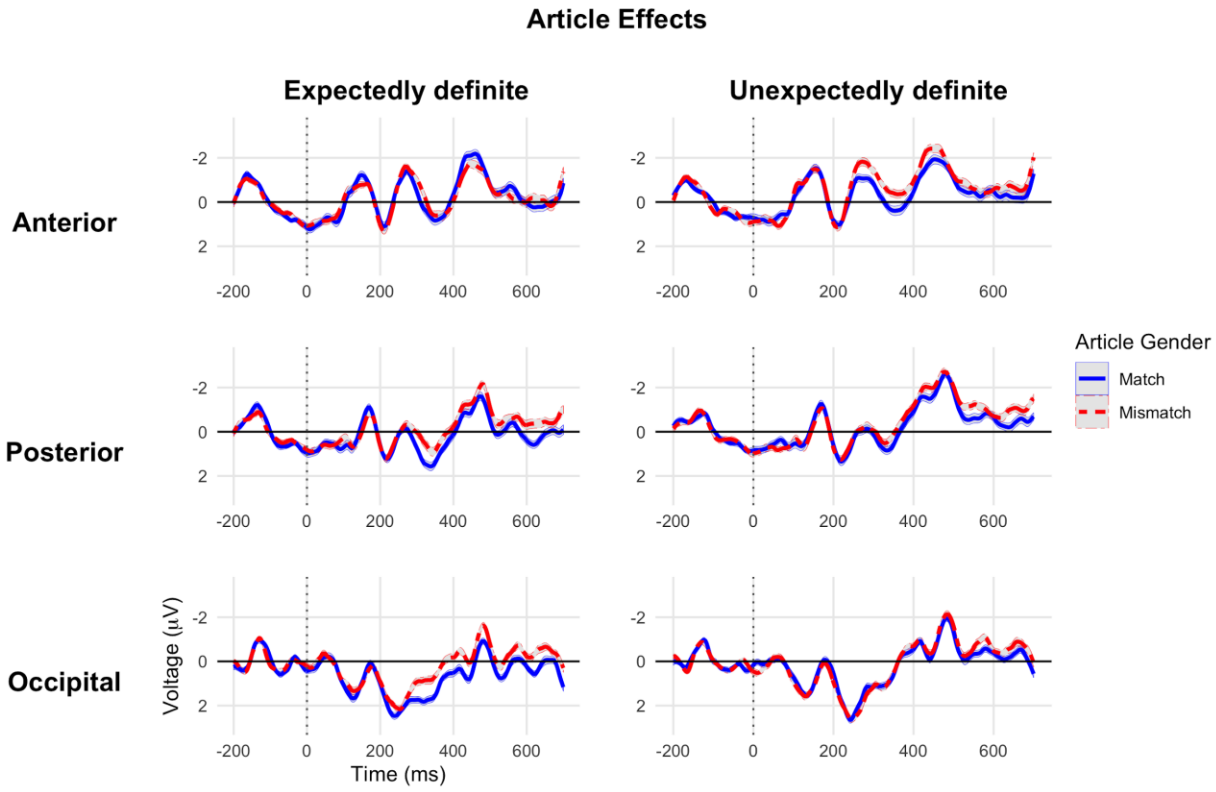


Figure 6. Article effects in Experiment 2. The graphs show the grand-average ERPs elicited by gender-matching articles (solid blue lines) and gender-mismatching articles (dotted red lines) at the pre-registered anterior, posterior and occipital ROIs (top, middle and bottom graphs, respectively), when articles were expectedly and unexpectedly definite (left and right graphs, respectively).

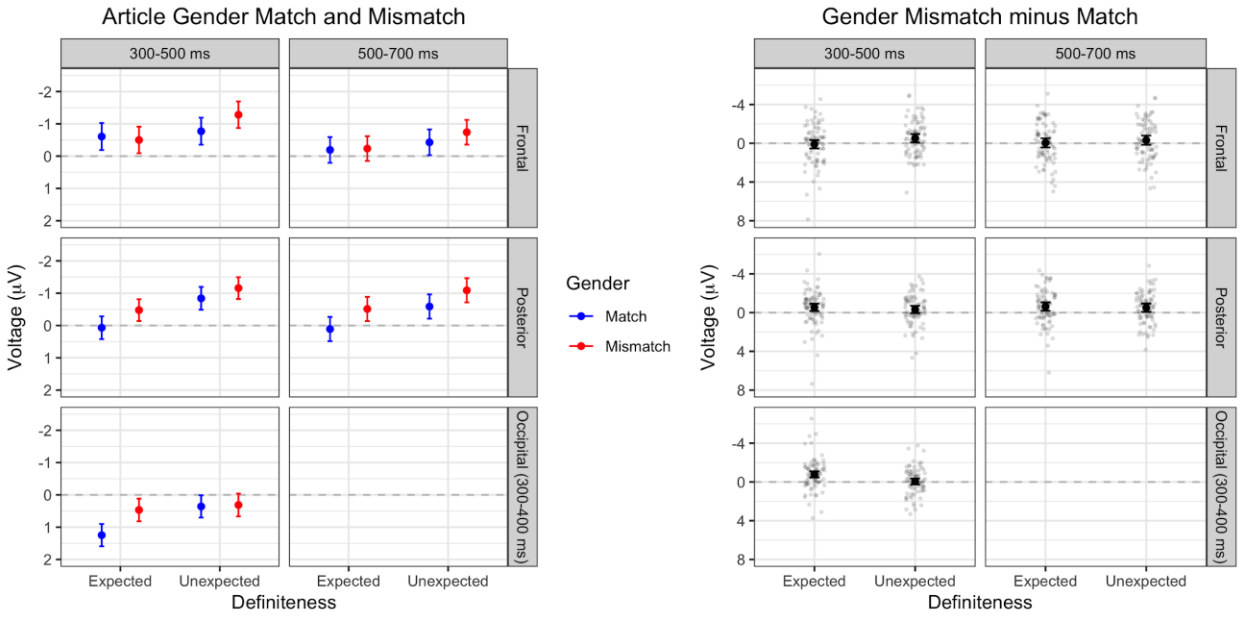


Figure 7. Article effects in Experiment 2. The left graphs show the estimated marginal means per condition from the mixed-effects model output for each ROI (large dots), along with the 95% confidence interval (vertical whisker). The right graphs show the corresponding estimated marginal means for the gender-mismatch effect (mismatch minus match), along with 95% confidence interval and subject-level mean effects.

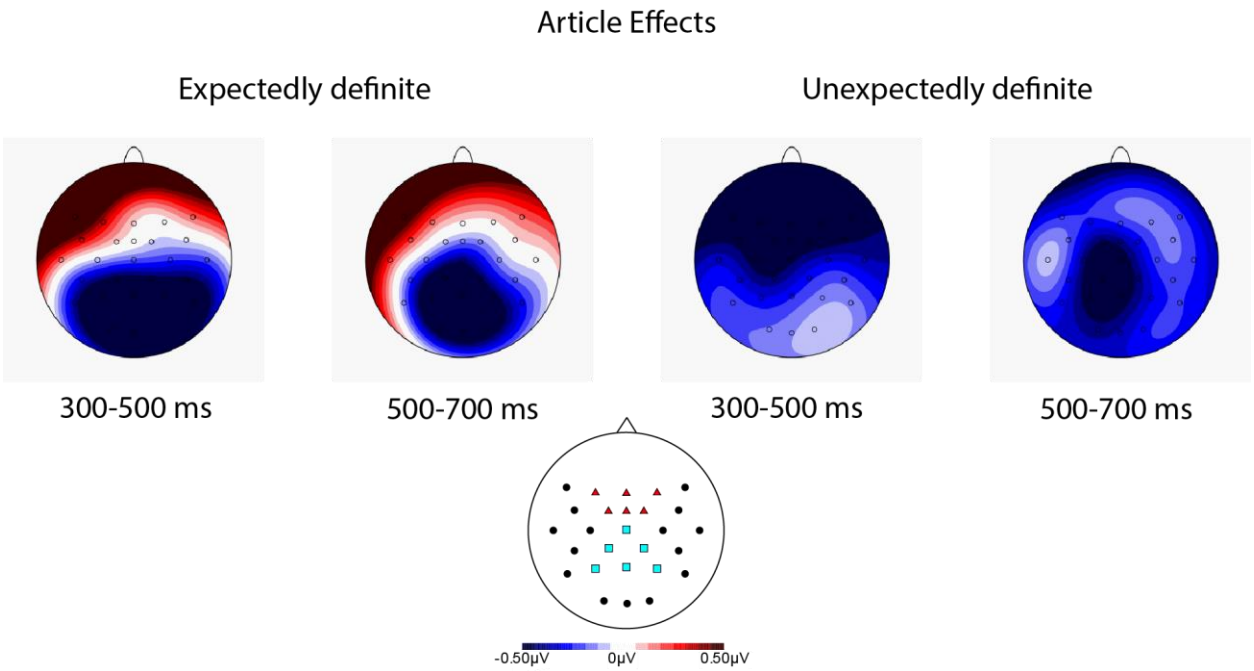


Figure 8. Scalp plots of the article effects (gender mismatch minus match) in Experiment 2.

Table 5. Results from the pre-registered article-analyses in Experiment 2.							
		Time window					
Factor	ROI	300-500 ms (N400)			500-700 ms (post-N400)		
		ß, CI	□^{2}	*p*	*ß, CI*	□^{2}	*p*
Gender	Anterior	-0.20, [-0.54,0.14]	1.37	0.24	-0.18, [-0.54, 0.19]	0.90	0.34
	Posterior	-0.43, [-0,71, -0.15]	8.88	<0.01	-0.56, [-0.89, -0.23]	10.69	<0.01
	Occipital (300-400 ms)	-0.41, [-0.64, -0.19]	13.09	<0.001			
Definiteness	Anterior	-0.47, [-0.75, -0.19]	10.78	<0.001	-0.37, [-0.67, -0.07]	5.79	0.02
	Posterior	-0.80, [-1.06, -0.54]	35.87	<0.001	-0.64, [-0.92, -0.36]	19.92	<0.001
	Occipital (300-400 ms)	-0.52, [-0.67, -0.07]	22.53	<0.001			
Gender: Definiteness	Anterior	-0.61, [-1.18 ,-0.05]	4.58	0.03	-0.27, [-0.87, 0.33]	0.77	0.37
	Posterior	0.23, [-0.29, 0.75]	0.74	0.39	0.12, [-0.44, 0.68]	0.17	0.67
	Occipital (300-400 ms)	0.73, [0.30, 1.17]	10.61	<0.001			

Pre-registered noun analyses

The patterns we observed for the nouns (Figure 9-10; Supplementary Figures 7-8, Table 6) were highly similar to those from Experiment 1. Prediction-mismatching nouns elicited more negative ERPs in the posterior ROI at 300-500 ms after noun onset, i.e., an N400 effect, compared to matching nouns, $\beta = -1.22$, CI = [-1.66, -0.78], $\chi^2(1) = 26.86$, $p < 0.001$, and more positive ERPs in the anterior ROI at 500-700 ms, $\beta = 0.59$, CI = [0.12, 1.05], $\chi^2(1) = 6.05$, $p = 0.01$. In the posterior ROI at 300-500 ms, nouns following expectedly definite articles elicited more negative ERPs compared to nouns following unexpectedly definite articles, $\beta = 0.85$, CI = [0.56, 1.13], $\chi^2(1) = 33.67$, $p < 0.001$. Finally, we found evidence for the same type of interaction observed in Experiment 1 in the posterior ROI at 300-500 ms, reflecting a more pronounced N400 effect of gender mismatch for unexpectedly definite nouns(-1.97 μV, SE = 0.27, Z =7.38 , $p < 0.001$) than for expectedly definite nouns, (-0.48 μV, SE = 0.27, Z = 1.80, $p = 0.07$). Overall, the average voltages were more positive for unexpectedly definite nouns match, mean = 3.65 μV, SE = 0.0.29; mismatch, mean = 1.68 μV, SE = 0.33) than for unexpectedly definite nouns (match, mean = 2.05 μV, SE =0.29; mismatch, mean = 1.57 μV, SE = 0.33). The interaction pattern thus mostly resulted from the effect of definiteness on the matching nouns.

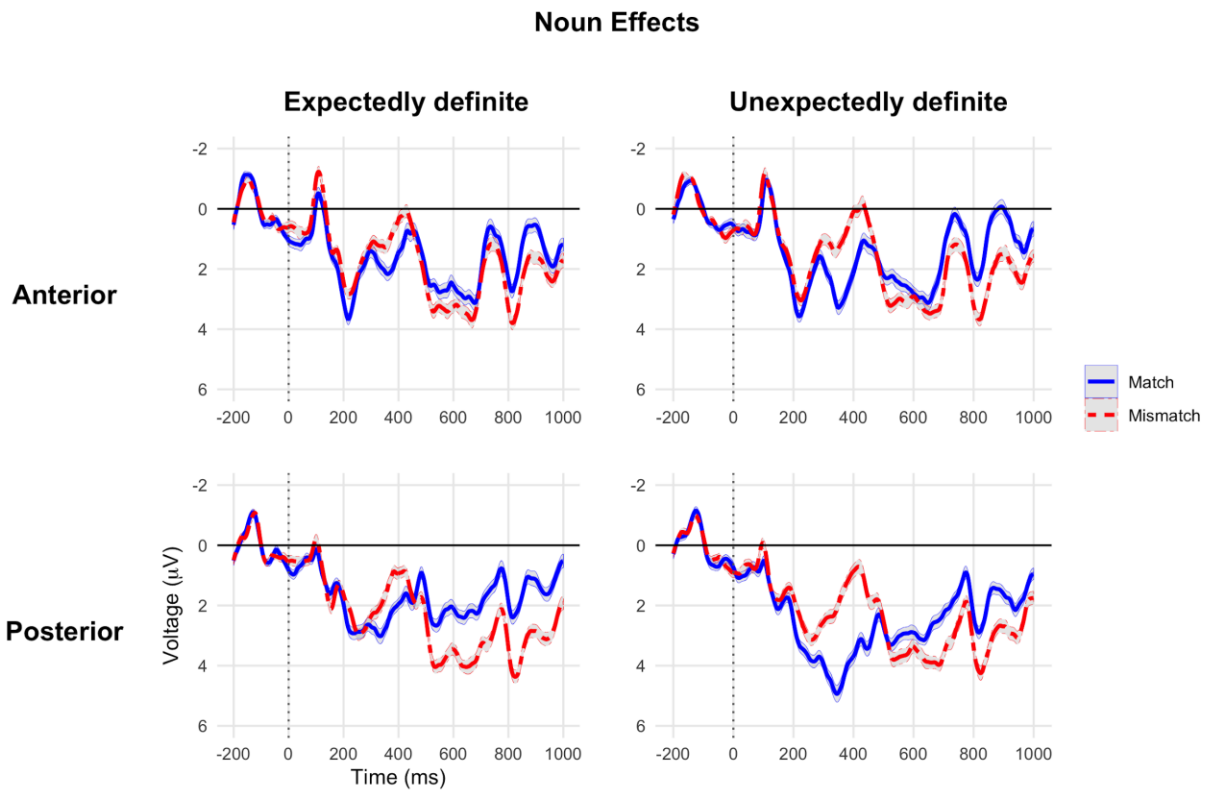


Figure 9. Noun effects in Experiment 2. The graphs show the grand-average ERPs elicited by prediction-matching nouns (solid blue lines) and prediction-mismatching nouns (dotted red lines) at the anterior and posterior ROIs (top and bottom graphs, respectively), following articles that were expectedly and unexpectedly definite (left and right graphs, respectively).

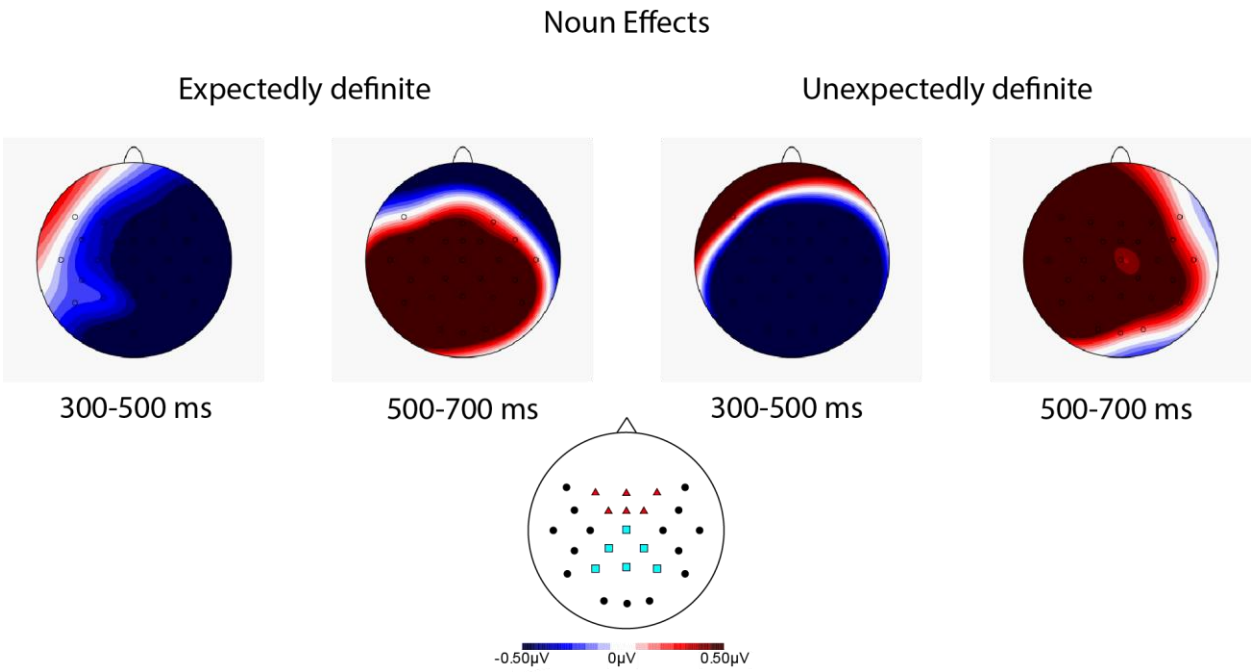


Figure 10. Scalp plots of the noun effects (prediction mismatch minus match) in Experiment 2.

Table 6. Results from the noun-analyses in Experiment 2.						
	ROI					
Factor	Posterior, 300-500 ms			Anterior, 500-700 ms		
	β, CI	□^{2}	*p*	*β, CI*	□^{2}	*p*
Mismatch	-1.22, [-1.66,-0.78]	26.86	<0.001	0.59, [0.12,1.05]	6.05	0.01
Definiteness	0.85, [0.56, 1.13]	33.67	<0.001	-0.14, [-0.45,0.17]	0.80	0.37
Mismatch:Definiteness	-1.49, [-2.06,-0.92]	25.95	<0.001	0.02, [-0.61,0.64]	0.00	0.96

Exploratory tests: definiteness versus gender

In both experiments, the N400 effect of unexpected definiteness was numerically stronger than that of unexpected gender, which suggests that unexpected definiteness has a bigger impact on semantic processing than unexpected gender. We therefore followed up with a pairwise comparison between the conditions that mismatched with the expected condition only in gender or definiteness; that is, expectedly definite, gender-mismatching articles vs. unexpectedly definite, gender-matching articles. These analyses used the combined data from Experiment 1 and 2 (N=128), and focused on the earlier time windows (300-500 ms for anterior and posterior ROIs, 300-400 ms for the occipital ROI). Unexpectedly definite, gender-matching articles elicited larger (more negative) N400s than expectedly definite, gender-mismatching articles in the anterior ROI (β = -0.52 , SE = 0.17 , z = -3.05 , p =.002) and the posterior ROI (β = -0.48 , SE = 0.15 , z = -3.09 , p =.002) but not the occipital ROI (β = -0.053 , SE = 0.13 , z = -0.41 , p =.68). So, while the occipital ROI does not appear sensitive to the type of prediction mismatch, the typical posterior (N400) ROI and the anterior ROI showed greater sensitivity to definiteness mismatch than to gender mismatch.

For exploratory analyses on the combined data sets that examine processing differences between ‘de’ and ‘het’ (see also e.g., Brouwer, Sprenger & Unsworth, 2017; Loerts, Wieling & Schmid, 2013), we refer interested readers to the Appendix.

Exploratory tests: revision of the noun prediction

An important remaining question is whether gender-mismatching articles caused our participants to revise their prediction to a different noun (e.g., Van Berkum et al., 2005), rather than to merely drop or dampen the original noun prediction. If our participants revised their prediction, such a process could correlate with the contextual constraint towards one alternative continuation (e.g., if 'de' disconfirmed the initial prediction for 'het boek'(the book) participants may have revised their prediction to 'roman' (novel) instead). An effect of noun prediction revision might then be detectable in the neural response to gender-mismatching articles, still before the noun is encountered. Moreover, a successfully revised prediction should facilitate access to the meaning of an alternative noun, one that was not the most predictable given the context, which we refer to as 'initially unpredictable'[13]. Mismatching nouns in our experiment should then elicit smaller N400s if they became highly predictable *upon* reading the mismatching article. We addressed this question with exploratory tests on the combined datasets from Experiment 1 and 2 that, for simplicity's sake, focused on the expectedly definite, gender-mismatching condition.

First, we performed two additional cloze completion tests to determine the most predictable nouns following the gender-mismatching articles. One 'restricted' cloze test instructed participants (N=25) to generate plausible continuations without using plural or

[13] By 'initially unpredictable' we mean a near zero cloze probability in the initial cloze test. We do not claim that their meaning was entirely unpredictable, because often they were related in meaning to the prediction-matching nouns. However, their meaning was probably less predictable than that of the matching nouns. Importantly, our analyses control for the possibility that nouns that became predictable after the mismatching article are more similar in meaning to the matching nouns than nouns that remained unpredictable.

diminutive nouns, and one 'unrestricted' cloze test (N=30) did not impose this restriction[14]. Participants completed only one of these two tests, and had not participated in the previous cloze test or the EEG experiments. We excluded participants who mostly gave ungrammatical responses with the originally highly-predictable nouns or responses that did not match the instructions. From the remaining 20 and 27 participants from the restricted and unrestricted test, respectively, we counted different spellings and words with partial lexical overlap (e.g., 'beeldscherm/scherm') towards the same response (as in the previous cloze test), but did not count ungrammatical responses with the originally predictable nouns.

As a measure of revised contextual constraint towards a specific continuation, we then computed Shannon's next-word entropy ($-\Sigma p_i \log_2(p_i)$, wherein p_i is the cloze probability of each unique response; Shannon, 1948; Taylor, 1954; see also Aurnhammer & Frank, 2019; Corps, Pickering & Gambi, 2019). Lower entropy values, i.e. a lower number of unique responses, correspond to stronger constraint[15]. Average entropy for the restricted test was 3.06 (SD = 0.73,

[14] We ran these two tests simultaneously because we were unsure which instruction would yield the most informative responses regarding the presumed revision processes. If our EEG participants often revised their prediction to a diminutive or plural version of the predictable noun (e.g., from 'de kerk' to 'het kerkje', and from 'het boek' to 'de boekjes') then the unrestricted responses could be informative but the restricted responses would not be. However, it is not evident that our EEG participants would revise their predictions to diminutive or plural forms of the predictable noun, because these never appeared in the experiment. Moreover, we worried that an unrestricted instruction would lead to the use of diminutives and plurals as a strategy to complete the test more quickly without paying much attention to meaning and sentence plausibility. The restricted instructions could therefore yield more informative responses if participants revised their predictions to another meaning (lemma).

[15] Next-word entropy is conceptually related to the traditional measure of contextual constraint (cloze probability of the most frequent response; e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007), but it can distinguish distributions that the traditional constraint measure cannot (e.g., 'book' and 'novel' with a 50% cloze probability each, versus 50% for 'book' and 10% for five different responses each). The average, traditional constraint for our restricted test was 32%

range 0.29-4.22). Entropy for the unrestricted test was slightly lower at 2.83 (SD = 0.79, range 0.72-4.20, paired t-test, $p < 0.001$), probably because the unrestricted test elicited many diminutive or plural forms of the predictable noun, which were the most frequent completion for 51 of the 160 items.

We used entropy as a z-transformed continuous predictor, together with the categorical predictor 'article type' ('de', 'het') and the z-transformed continuous predictor 'position' (word position in the sentence), for article-elicited EEG activity in the pre-registered ROIs (all models had by-subject random slopes for entropy and article type). Interestingly, higher entropy in the restricted test was associated with more positive voltage (see Figure 11, left graph), especially at the anterior ROIs (300-500 ms, $\beta = 0.33$, SE = 0.16 , t= 2.07 , $p = 0.041$; 500-700 ms, $\beta = 0.41$, SE = 0.18, t= 2.31 , $p = 0.023$) and the posterior ROI in the 500-700 ms time window ($\beta = 0.33$, SE = 0.16 , t= 2.12 , $p = 0.036$), but less so at the N400 ROI ($\beta = 0.19$, SE = 0.14 , t= 1.37 , $p = 0.17$) and the occipital ROI ($\beta = 0.10$, SE = 0.10, t= 0.97 , $p = 0.332$). Entropy from the unrestricted test elicited weaker, not statistically significant effects (anterior ROI 300-500, $\beta = -.02$, SE = 0.14 , t= 0.14 , $p = 0.89$; 500-700 ms, $\beta = 0.09$, SE = 0.17 , t= 0.54 , $p = 0.59$; posterior ROI 300-500 ms, $\beta = 0.05$, SE = 0.13 , t= 0.35 , $p = 0.73$; 500-700 ms, $\beta = 0.19$, SE = 0.16 , t= 1.27 , $p = 0.21$; occipital ROI, $\beta = 0.08$, SE = 0.10 , t= 0.88 , $p = 0.38$).

(SD = 18, range 10-95), and for the unrestricted test 38% (SD = 19, range 9-88). Analyses with this traditional measure showed similar, albeit somewhat weaker effects compared to entropy.

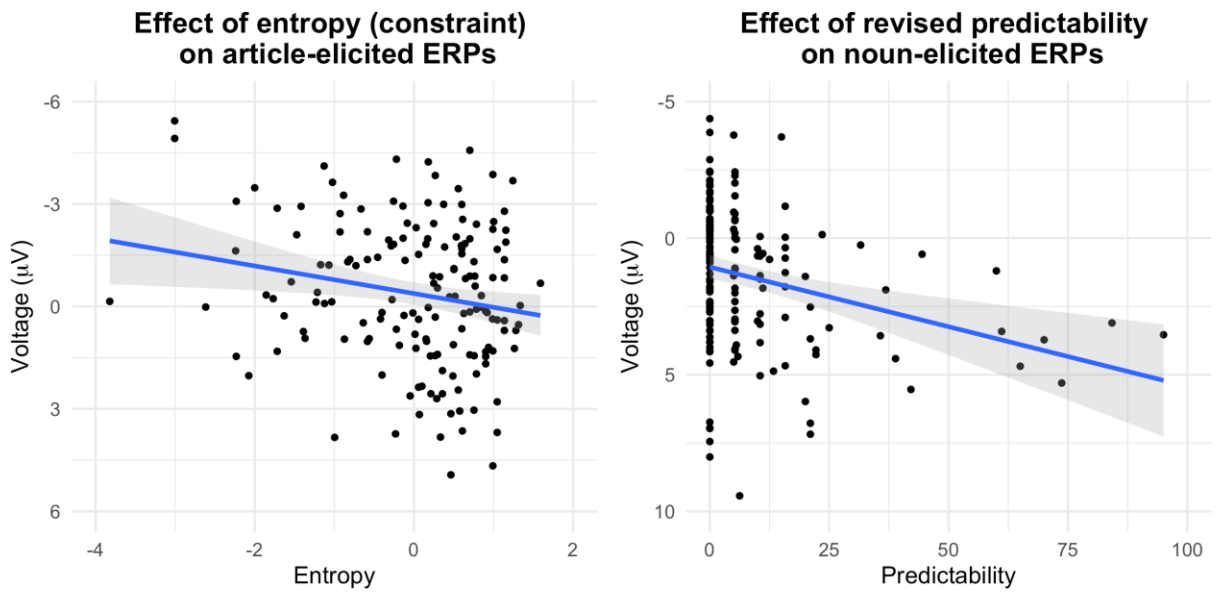


Figure 11. Results from the exploratory tests for noun prediction revision, using values from the restricted cloze test. The left graph shows the effect of next-word entropy (z-transformed) on article-elicited ERPs (anterior ROI in the 500-700 ms time window), with dots showing the mean voltage of each item. Greater entropy (weaker constraint) was associated with more positive ERPs. The right graph shows the effect of revised predictability (cloze probability of the prediction- mismatching nouns given a gender-mismatching article) on N400 activity (posterior ROI, 300-500 ms time window). More predictable nouns elicited smaller (less negative) N400s.

Subsequently, we computed 'revised' predictability (cloze probability given the gender-mismatching article) of the prediction-mismatching nouns presented in the experiment. While these nouns still did not have high cloze values (restricted test, mean = 7%, SD = 13, range 0-84; unrestricted test, mean = 8%, SD = 17, range 0-86), the new values were higher and more variable than the original cloze values. We tested whether revised predictability correlated with noun-elicited N400 activity (posterior ROI, 300-500 ms time window), while controlling for a range of relevant variables that differed between items: word length (number of characters), frequency (Keuleers et al., 2010), word position, semantic similarity to the initially predictable noun (Mandera, Keuleers & Brysbaert, 2017) and plausibility. Plausibility was obtained with an additional online rating test, wherein participants (N=28) rated how well the noun fitted the context on a 5-point scale from 1 (very poorly) to 5 (very well). Simultaneously modelling these sources of variance gives greater confidence that obtained effects of predictability result from prediction, rather than, for example, ease of integration (see Nieuwland et al., 2019). We performed mixed-effects model analyses with N400 amplitude as dependent variable and with fixed effects of the continuous measures predictability, plausibility, and their interaction (see Nieuwland et al., 2019), as well as word length, frequency, and semantic similarity, with by-subject random slopes for all fixed effects. We performed these analyses with predictability either as raw cloze probability or as log-transformed probability (log-transform gives greater weight to differences between low cloze values than between high cloze values, and has been argued to better capture the quantitative relationship between prediction and online processing measures, e.g., Smith & Levy, 2013). All predictors were z-transformed.

As depicted in Figure 11, our analyses revealed a general pattern of effects wherein more predictable words elicited smaller (less negative) N400s than less predictable words (restricted cloze, $\beta = 0.82$, SE = 0.21 , t= 3.97 , $p < 0.001$; restricted log-transformed cloze , $\beta = 0.61$, SE = 0.22 , t= 2.81 , $p = 0.006$; unrestricted cloze $\beta = 0.58$, SE = 0.20 , t= 2.83 , $p = 0.005$; unrestricted log-transformed cloze, $\beta = 0.55$, SE = 0.21 , t= 2.64, $p = 0.009$). Moreover, the effects strengthened, rather than weakened when excluding the 7 unpredictable nouns that nevertheless had non-zero cloze values in the original cloze test. In addition to these effects of predictability, we found overall smaller (less negative) N400s to be associated with increases in plausibility (all *t*-values > 2.6), increases in word position (all *t*-values > 3.1) and decreases in word length (all *t*-values > 3.2). Because our primary interest was in predictability, we do not report all details here. We also obtained some evidence for an interaction pattern wherein the effect of plausibility was smaller with increasing predictability (all *t*-values > 1.3; interactions were strongest when cloze was not log-transformed), which further strengthens the conclusion that nouns elicited reduced N400s because they became more predictable when participants encountered the article, not just because these nouns rendered the sentence meaning more plausible[16].

[16] We also performed the same analyses for the posterior ROI in the 500-700 ms time window, where prediction-mismatching nouns elicited a positivity compared to matching nouns. The post-N400 parietal positivity is sometimes taken as a measure of integration difficulty because it has been observed for implausible nouns (e.g., Brouwer, Fitz & Hoeks, 2012; for a review, see Van Petten & Luka, 2012). However, in our study, less plausible nouns were associated with enhanced negativity (for models with restricted/unrestricted, raw/log-transformed cloze, all *t*-values > 2.5), not positivity, along with effects of word frequency and sentence position. This pattern, also observed by Nieuwland et al. (2019), could index an ‘extended N400 effect' associated with continued semantic processing of less plausible words.

In sum, ERP activity elicited by the articles and the nouns correlated with the ease with which participants may have revised a disconfirmed prediction to a new one, yielding additional support for the noun prediction revision hypothesis.

GENERAL DISCUSSION

In two ERP studies on Dutch mini-story comprehension, we investigated the functional significance of 'pre-nominal prediction effects', the differential neural activity elicited by pre-nominal articles that mismatch the gender of a likely upcoming noun (Kutas et al., 2011; Van Berkum, 2009), when compared to gender-matching articles. We contrasted two hypotheses from the extant literature. According to what we dub the article prediction mismatch hypothesis, people predict the article along with the noun (e.g., DeLong et al., 2005; Kutas et al., 2011; Wicha et al 2003, 2004; see also Dell & Chang, 2014) and the effect reflects processing of the mismatch with the predicted article. According to the alternative, noun prediction revision hypothesis (e.g., Van Berkum et al. 2005), the effect merely reflects use of the article to inform and revise the noun prediction, and no article form prediction is assumed. We contrasted these hypotheses, capitalizing on the fact that Dutch definite articles are gender-marked ('de/het') whereas indefinite articles are not ('een'). If the pre-nominal prediction effect reflects mismatch with a predicted article, then the effect should occur when participants expected a gender-marked definite article, but not when they expected an indefinite article without gender-marking. Alternatively, if the effect reflects use of gender-marked input to revise a noun prediction, then the effect should occur regardless of expected definiteness.

In Experiment 1, our pre-registered analyses revealed increased N400 amplitude for gender-mismatching articles compared to matching articles, demonstrating that readers made predictions, and an effect of definiteness (unexpectedly definite articles compared to expectedly definite articles), with both effects extending into the 500-700 ms time window. Crucially, although the gender mismatch effect was numerically larger for expectedly definite articles than for unexpectedly definite articles, consistent with the article prediction mismatch hypothesis, evidence for this interaction was inconclusive. Supporting the noun prediction revision hypothesis, however, both expectedly and unexpectedly definite articles elicited a gender mismatch effect at the posterior ROI in the 500-700 ms time window.

Exploratory mass regression analyses and a power analysis suggested a sub-optimal choice of ROI and insufficient sample size for detecting an interaction pattern in Experiment 1. We therefore performed direct replication Experiment 2 (N=80), which confirmed the interaction pattern at a newly pre-registered, occipital ROI (300-400 ms), where only expectedly definite articles elicited a gender-mismatch effect. The interaction effects in Experiment 1 and 2 were practically equivalent in terms of effect size (e.g., Kruschke & Lidell, 2018a,b). Furthermore, like in Experiment 1, unexpected definiteness yielded enhanced negativity at the posterior (parietal) ROI in the 300-500 (N400) and 500-700 ms time window. Supporting the noun prediction revision hypothesis, and replicating Experiment 1, expectedly and unexpectedly definite articles elicited a similar gender mismatch effect at the posterior ROI in the 500-700 ms time window. Moreover, unlike in Experiment 1, an additional interaction effect occurred at anterior channels in the 300-500 ms time window, where only the unexpectedly definite articles elicited a negativity associated with gender mismatch. With exploratory analyses for the

combined datasets, we showed that unexpected definiteness yielded a larger amplitude N400 than unexpected gender, and we provided further evidence that participants used gender marking on the article to revise their noun prediction.

In sum, our results support both the article prediction mismatch hypothesis and the noun prediction revision hypothesis. As already briefly foreshadowed in our introduction, these hypotheses are not mutually exclusive, and the pre-nominal prediction effect may be a multi-faceted phenomenon encompassing at least two distinct effects. Broadly speaking, these effects differ on two dimensions: the effect associated with prediction mismatch occurs relatively early and has a strongly posterior distribution, while effects associated with prediction revision occur later and have a more anterior distribution, at least in our data.

In the section below, we unpack our conclusions regarding the article prediction mismatch hypothesis and the noun prediction revision hypothesis, respectively.

Processing article prediction mismatch

The article prediction mismatch hypothesis assumes that people predict not just the meaning of an upcoming referent, but also the word form of the noun plus the corresponding article (DeLong et al., 2005; Kutas et al., 2011; Wicha et al 2003, 2004). People presumably first predict a specific noun including its gender, and then also predict the specific form of the article (which depends on definiteness and gender, at least in Dutch). Processing the mismatch between the predicted and encountered article form then gives rise to a pre-nominal ERP effect.

Support for this hypothesis came from the gender mismatch effect that was unique to the expectedly definite articles. This effect had an occipital (or, more accurately, occipital-parietal)

maximum that was consistent across our two experiments. This scalp distribution appears consistent with those reported in Spanish language studies (e.g., Martin et al., 2018; Molinaro et al. 2017; Wicha et al., 2003b). This distribution is noteworthy, because although the effect was highly reminiscent of an N400 effect in terms of timing and waveform morphology, the strongly posterior maximum deviates from the typical centroparietal distribution of a noun-elicited N400 effect. This deviation may be related to semantic processing differences between articles and nouns that elicit different N400s to begin with, irrespective of prediction. Speculatively, it could also be due to increased contributions from occipital or occipital-temporal neural generators that process visual and word-form information, respectively, if article predictions are implemented as perceptual predictions of visual word form (e.g., Dambacher, Rolfs, Gollner, Kliegl, & Jacobs, 2009; but see Nieuwland, 2019, for a critical review). Although intriguing, the onset of the mismatch effect (starting at about 250-300 ms) does not support an explanation in terms of early visual word-form processing, and this remains an open question for follow-up research.

Revising a noun prediction

The noun prediction revision hypothesis does not assume article form prediction, but only prediction of the noun. Once people encounter the article, they use its form to inform their prediction of the noun. This hypothesis was first suggested by Van Berkum and colleagues (2005) to explain prediction effects on pre-nominal adjectives, and has recently been adopted by some authors as a more general explanatory mechanism indexed by N400 amplitude (‘semantic prediction updating’, Rabovsky, 2020; Rabovsky, Hansen, & McClelland, 2018; Szewczyk & Wodniecka, 2020).

Our rationale was that a gender-mismatch effect between two equally unpredictable articles cannot be explained by the article prediction mismatch hypothesis. If an effect does not index the mismatch with a predicted article, then it arguably indexes something less controversial, namely how the article is used to inform or revise the widely assumed noun prediction. Our results clearly demonstrated such an effect, perhaps in two forms: a later posterior negativity (Experiment 1 and 2) and an earlier frontal negativity (only observed in Experiment 2). The former, late posterior negativity was elicited in both experiments and by both expectedly and unexpectedly definite articles. We note that the late and extended nature of this effect is not that unusual in light of other results with gender-based manipulations (e.g., Ito, Gambi, Pickering, Fuellenbach, & Husband, 2020; Martin et al., 2013, 2018; Foucart et al., 2014). As we discuss below, this effect could be a reflection of the processes by which participants revised their prediction.

Regarding the latter, early frontal effect, we should raise two caveats. Although Experiment 1 showed a numerical effect in the same direction, this frontal effect does not appear as strong or replicable as the later posterior ERP effect. Moreover, the lack of a corresponding effect for expectedly definite articles suggests that this effect may not be related to prediction revision, which would presumably be elicited by both expectedly and unexpectedly definite articles.

These two effects are not accounted for by the article prediction mismatch hypothesis, but they provide only indirect evidence for actual revision of a prediction. We therefore explored the possibility that some combinations of context and article may have allowed participants to revise their prediction ahead of the noun more easily than other combinations. We quantified this

'prediction revision' as the constraint that gender-mismatching articles resulted in towards an alternative noun (measured as next-word entropy). Yielding support for prediction revision, ERPs at both frontal ROIs and the later posterior ROIs became more negative with increased constraint (lower entropy). Interestingly, given that the gender-mismatch effect for expectedly definite articles was also a negativity, this suggests that revision of a prediction may incur a processing cost compared to when no revision takes place (see also Szewczyk & Wodniecka, 2020, for a similar suggestion).

Second, we hypothesized that if participants successfully revised their noun prediction to initially unpredictable (or just less predictable) nouns that were presented during the experiment, then access to their meaning should be facilitated relative to nouns that remained unpredictable, as indexed by reduced N400 amplitude. Yielding further support for prediction revision, we observed that N400 amplitude gradually decreased with the revised predictability of the noun after the gender-mismatching article (see also Szewczyk & Wodniecka, 2020). Importantly, this analysis controlled for other relevant influences on N400 amplitude (word length and frequency, word position, semantic similarity to the predicted noun, and plausibility). This allowed confidence that initially unpredictable nouns elicited smaller N400s because they became predictable after the article, not because they were semantically similar to the initially predicted noun or rendered the sentence more plausible (see Nieuwland et al., 2019, for discussion).

Naturally, while these patterns suggest that gender-mismatching articles caused participants to revise their initial noun prediction, at least during some of the trials, we emphasize the exploratory nature of the analyses and need for further confirmation. In particular the ERP effect associated with revised constraint (next-word entropy) was not very strong, and

only reached the traditional level of statistical significance in a subset of the analyses we performed.

The role of definiteness

Our exploratory results also suggested that unexpected definiteness elicits a larger amplitude N400 than unexpected gender (even if only considering gender for expectedly definite articles), to the extent that these conditions can be directly compared. A possible explanation is that compared to unexpected gender, unexpected definiteness is more meaningful and leads to intensified semantic retrieval (e.g., Kutas & Federmeier, 2000, 2011; Van Berkum, 2009) or incurs a greater change to the semantic representation of sentence meaning (Bornkessel-Schlesewsky & Schlesewsky, 2019; Rabovsky, Hansen & McClelland, 2018; Nieuwland et al., 2018). In our experiment, unexpected gender may have signalled a (possibly very small) change in upcoming meaning, for example, instead of 'church' participants could revise their prediction to a less specific conceptual representation (e.g., some type of building for religious congregation) or some plausible, lexically specific alternative, as suggested by our revised predictability norms. However, unexpected definiteness has stronger repercussions for the situation model (which also entails the discourse information structure), because it is typically reserved for uniquely identifiable referents (ones that are already given, readily accessible or anticipated; Abbott, 2004, 2006; Almor & Nair, 2007; Ariel, 1988; Arnold et al., 2013; Frazier, 2006; Sanford & Garrod, 1988; Schumacher, 2009; Roberts, 2003). Unexpected definiteness therefore violates the presupposition of a uniquely identifiable referent (Karttunen, 1974; Krahmer, 1998; Levinson, 1983; Stalnaker, 1977; Von Fintel, 2004). This could act as a

'relevance signal' that triggers more detailed semantic processing or it might lead to a change in how the meaning of the sentence is represented (by accommodation of a unique referent into the discourse representation, e.g., one specific church; see Beaver, 1999; Von Fintel, 2008). Such changes in semantic processing, and the potential meanings they afforded, may be reflected in N400 activity (Bornkessel-Schlesewsky & Schlesewsky, 2019; Kutas & Federmeier, 2011; Rabovsky et al., 2018; Van Berkum, 2009).

The strong effect of unexpected definiteness might be related to a surprising result from our study, namely that predictable nouns elicited smaller N400 amplitude when they followed unexpectedly definite articles compared to when they followed expectedly definite articles. We can speculate that the semantic processing changes afforded by the additional information from unexpectedly definite articles (e.g., intensified semantic retrieval or updating of sentence meaning) could have boosted the semantic pre-activation of the predictable noun.

Implications for N400 prediction effects on pre-nominal articles

Our article results highlight that different effects of the context play out in what seem like different types of effects. For example, unexpected definiteness elicited what is often considered a typical N400 effect (i.e. an effect with a centroparietal maximum in the 300-500 ms time window). The effects of prediction mismatch, in contrast, had a more posterior maximum, whereas the effect of prediction revision had a frontal maximum and was most evident in the later 500-700 ms time window. It remains to be seen whether models of the effects of prediction error or prediction revision on ERPs can explain this, since they currently tend to focus on N400 amplitude (e.g., Fitz & Chang, 2019; Rabovsky et al., 2018; Rabovsky, 2020).

Our results also shed new light on previous failures to find clear or consistent N400 prediction effects on pre-nominal articles. For example, Otten and Van Berkum (2009) reported a frontally distributed N400-like effect, which differed from the effects reported by Wicha et al. (2003a,b; 2004), and Kochari and Flecken recently reported a subsequent failure to obtain statistically significant effects with materials similar to those of Otten and Van Berkum. Our study suggests that these differences may be traced back to the fact that these two Dutch studies predominantly involved unexpectedly definite articles, which yield qualitatively and quantitatively different effects from expectedly definite/indefinite articles as used in the Spanish studies (e.g., Wicha et al., 2003a, 2003b; Martin et al., 2018).

Different pre-nominal manipulations may elicit distinct prediction effects, and so may different languages (see also Bornkessel-Schlesewsky & Schlesewsky, 2009; Kamide, Scheepers, & Altmann, 2003; Van Bergen & Flecken, 2017). For example, Dutch does not mark gender on indefinite articles and its form for definite articles is not a perfectly reliable cue to noun gender. This is different in Spanish, which uses a unique article for each possible combination of gender, definiteness and number, and also has gender-marking on the nouns themselves. For these reasons, it is possible that Spanish is a more suitable language for eliciting pre-nominal prediction effects than Dutch. Yet different patterns of results may occur in other languages with rich case and gender systems such as German (e.g., Nicenboim, Vasishth & Rösler, 2020; Schoknecht, Roehm, Schlesewsky, & Bornkessel-Schlesewsky, 2019) or Polish (e.g., Szewczyk & Wodniecka, 2020).

Conclusions

Our results add to the growing body of evidence for linguistic prediction using a pre-nominal gender manipulation in highly constraining sentences. Our results support the article prediction mismatch hypothesis, in which an ERP effect is elicited by the mismatch between a predicted and encountered article (Wicha et al 2003, 2004). But our results also support the noun prediction revision hypothesis (e.g., van Berkum et al., 2005), by demonstrating a gender-mismatch effect even when article gender marking was unexpected. Crucially, these two effects have a distinct time course and scalp distribution: the prediction mismatch effect had a strongly posterior scalp distribution and was maximal around 300-400 ms, while the prediction revision effect was strongest in the 500-700 ms time window. Exploratory analyses yielded further support for prediction revision: ERPs elicited by gender-mismatching articles correlated with incurred constraint towards a new noun (next-word entropy), and N400s for initially unpredictable nouns decreased when articles made them more predictable. These results demonstrate the dual nature of pre-nominal prediction effects, reconciling two prevalent explanations from the extant literature.

References

Abbott, B. (2004). Definiteness and indefiniteness. *The Handbook of Pragmatics*, 122-149.

Abbott, B. (2006). Definite and indefinite. *Encyclopedia of Language and Linguistics*, *3*, 392-399.

Anderson, J. E., & Holcomb, P. J. (2005). An electrophysiological investigation of the effects of coreference on word repetition and synonymy. *Brain and Language*, 94(2), 200-216.

Almor, A., & Nair, V. A. (2007). The form of referential expressions in discourse. *Language and Linguistics Compass*, *1*(1-2), 84-99.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24(1), 65-87.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134, 107198.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Beaver, D. (1999). Presupposition accommodation: A plea for common sense. *Logic, Language and Computation*, 2, 21-44.

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)*, 289-300.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Towards a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology*, 10, 298.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446, 127-143.

Brouwer, S., Sprenger, S., & Unsworth, S. (2017). Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology*, 159, 50-65.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14(1), 177-208.

Clifton Jr, C. (2013). Situational context affects definiteness preferences: Accommodation of presuppositions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 487.

Corps, R. E., Pickering, M. J., & Gambi, C. (2019). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, 34(5), 615-627.

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.

Dambacher, M., Rolfs, M., Göllner, K., Kliegl, R., & Jacobs, A. M. (2009). Event-related potentials reveal rapid verification of predicted visual input. *PloS one*, *4*(3).

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634), 20120394.

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

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia, 61*, 150-162.

Deutsch, W., & Wijnen, F. (1985). The article's noun and the noun's article: Explorations into the representation and access of linguistic gender in Dutch. *Linguistics*, 23(5), 793-810.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75-84.

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.

Fraurud, K. (1990). Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4), 395-433.

Frazier, L. (2006). The big fish in a small pond: Accommodation and the processing of novel definites. Paper presented at the Ohio State University Presupposition Accommodation Workshop, Columbus, Ohio.

Geerts, G. (1975). Het genus van Engelse leenwoorden in het Duits en in het Nederlands. *Spel van zinnen*, 115-123.

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.

Giannelli, F., & Molinaro, N. (2018). Reanalyzing language expectations: Native language knowledge modulates the sensitivity to intervening cues during anticipatory processing. *Psychophysiology*, *55*(10), e13196.

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. doi: 10.1111/2041-210X.12504

Heim, I. (1982). *The semantics of definite and indefinite noun phrases.* Amherst, MA:University of Massachusetts doctoral dissertation.

Ito, A., Gambi, C., Pickering, M. J., Fuellenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, 136, 107291.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954-965.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(4), 635-652. doi:10.1037/xlm0000315.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017c). 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.

Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32(1), 37-55.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics*, 1(1-3), 181-194.

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Kirsten, M., Tiemann, S., Seibold, V. C., Hertrich, I., Beck, S., & Rolke, B. (2014). When the polar bear encounters many polar bears: event-related potential context effects evoked by uniqueness failure. *Language, Cognition and Neuroscience*, 29(9), 1147-1162.

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.

Krahmer, E. (1998). Presupposition and anaphora. Stanford: CSLI Publications.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), 155-177.

Kruschke, J. K., & Liddell, T. M. (2018b). 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., A look around at what lies ahead: Prediction and predictability in language processing, in *Predictions in the Brain: Using Our Past to Generate a Future*, Editor M. Bar, Oxford University Press, 2011, pp. 190-207.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463-470.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621-647.

Kutas, M., & Hillyard, S. A. (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.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press, Cambridge UK.

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.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57-78.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, 8(1), 1079.

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.

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, 107427.

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews*, 96, 367-400. doi:10.1016/j.neubiorev.2018.11.019.

Nieuwland, M.S., Arkhipova, Y., & Rodríguez-Gómez, P. (2020). *Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials.* Spoken presentation at CUNY 2020, the 33rd Annual CUNY Human Sentence Processing Conference, Amherst (MA), USA.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., Rueschemeyer, S.-A., Segaert, K., Tuomainen, J., & Von Grebmer Zu Wolfsthurn, S. (2019). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 375.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E.,

Husband, E. M., Donaldson, D. I., Kohút, Z., Rueschemeyer, S.-A., & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, 7: e33468. doi:10.7554/eLife.33468

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.

Otten, M., & Van Berkum, J. J. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research*, 1153, 166-177.

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

Otten, M., & Van Berkum, J. J. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse?. *Brain Research*, 1291, 92-101.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10), 1002.

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, 107466.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9), 693.

Roberts, C. (2003). Uniqueness in definite noun phrases. *Linguistics and Philosophy*, 26(3), 287-350.

Roelofs, A., Meyer, A. S., & Levelt, W. J. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition*, 69(2), 219-230.

Romero-Rivas, C., Corey, J. D., Garcia, X., Thierry, G., Martin, C. D., & Costa, A. (2017). World knowledge and novel information integration during L2 speech comprehension. *Bilingualism: Language and cognition*, 20(3), 576-587.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, 26(2-3), 159-190.

Schiller, N. O., & Caramazza, A. (2006). Grammatical gender selection and the representation of morphemes: The production of Dutch diminutives. *Language and Cognitive Processes*, 21(7-8), 945-973.

Schoknecht, P., Roehm, D., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2019). Looking forward does not mean forgetting about the past: ERP evidence for the interplay of predictive coding and interference during language processing. *BioRxiv*, 567560.

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.

Schriefers, H., & Jescheniak, J. D. (1999). Representation and processing of grammatical gender in language production: A review. *Journal of Psycholinguistic Research*, 28(6), 575-600.

Schumacher, P. B. (2009). Definiteness marking shows late effects during discourse processing: Evidence from ERPs. In *Discourse Anaphora and Anaphor Resolution Colloquium* (pp. 91-106). Springer, Berlin, Heidelberg.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3), 379-423.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302-319.

Singh, R., Fedorenko, E., Mahowald, K., & Gibson, E. (2016). Accommodating presuppositions is inappropriate in implausible contexts. *Cognitive Science*, *40*(3), 607-634.

Stalnaker, R. (1977). Pragmatic presuppositions. In *Proceedings of the Texas conference on per~ formatives, presuppositions, and implicatures.* Arlington, VA: Center for Applied Linguistics (pp. 135-148).

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, 120(2), 135-162.

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*

Taylor, W. L. (1954). *Application of 'cloze' and entropy measures to the study of contextual constraint in samples of continuous prose.* Doctoral dissertation, University of Illinois at Urbana-Champaign.

Van Bergen, G., & Flecken, M. (2017). Putting things in new places: Linguistic experience modulates the predictive power of placement verb semantics. *Journal of Memory and Language*, 92, 26-42.

Van Berkum, J. J. (1997). Syntactic processes in speech production: The retrieval of grammatical gender. *Cognition*, *64*(2), 115-152.

Van Berkum, J. J. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science*, 17(6), 376-380.

Van Berkum, J. J. (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., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 443.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, 83(2), 176-190.

Von Fintel, K. (2004). Would you believe it? The King of France is back! Presuppositions and truth-value intuitions. In *Descriptions and Beyond (pp. 315-341)*. Oxford University Press, Oxford.

Von Fintel, K. (2008). What is presupposition accommodation, again?. *Philosophical Perspectives*, 22, 137-170.

Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003a). Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, 346(3), 165-168.

Wicha, N. Y., Moreno, E. M., & Kutas, M. (2003b). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex*, 39(3), 483-508.

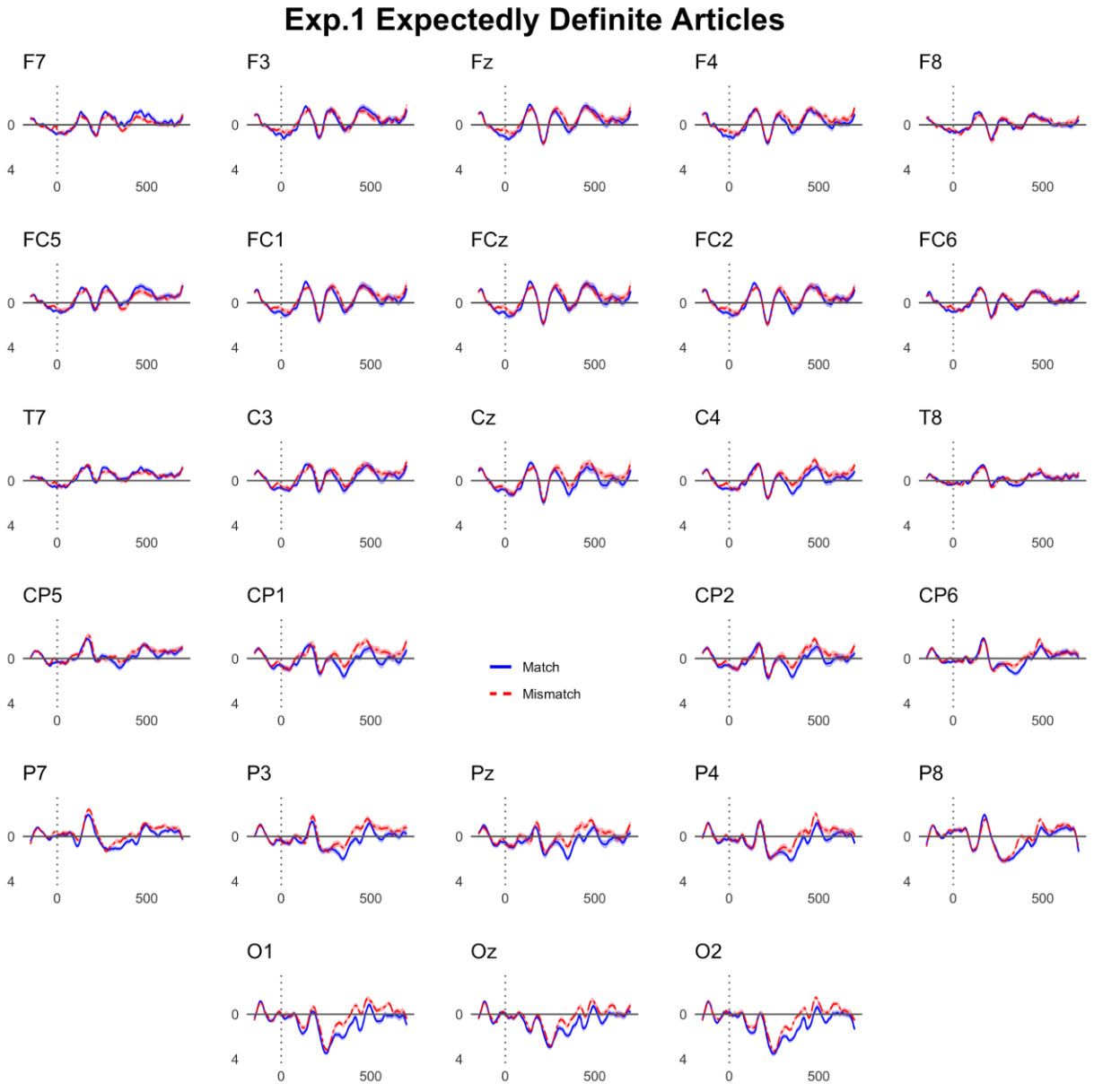
Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, 16(7), 1272-1288.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162.

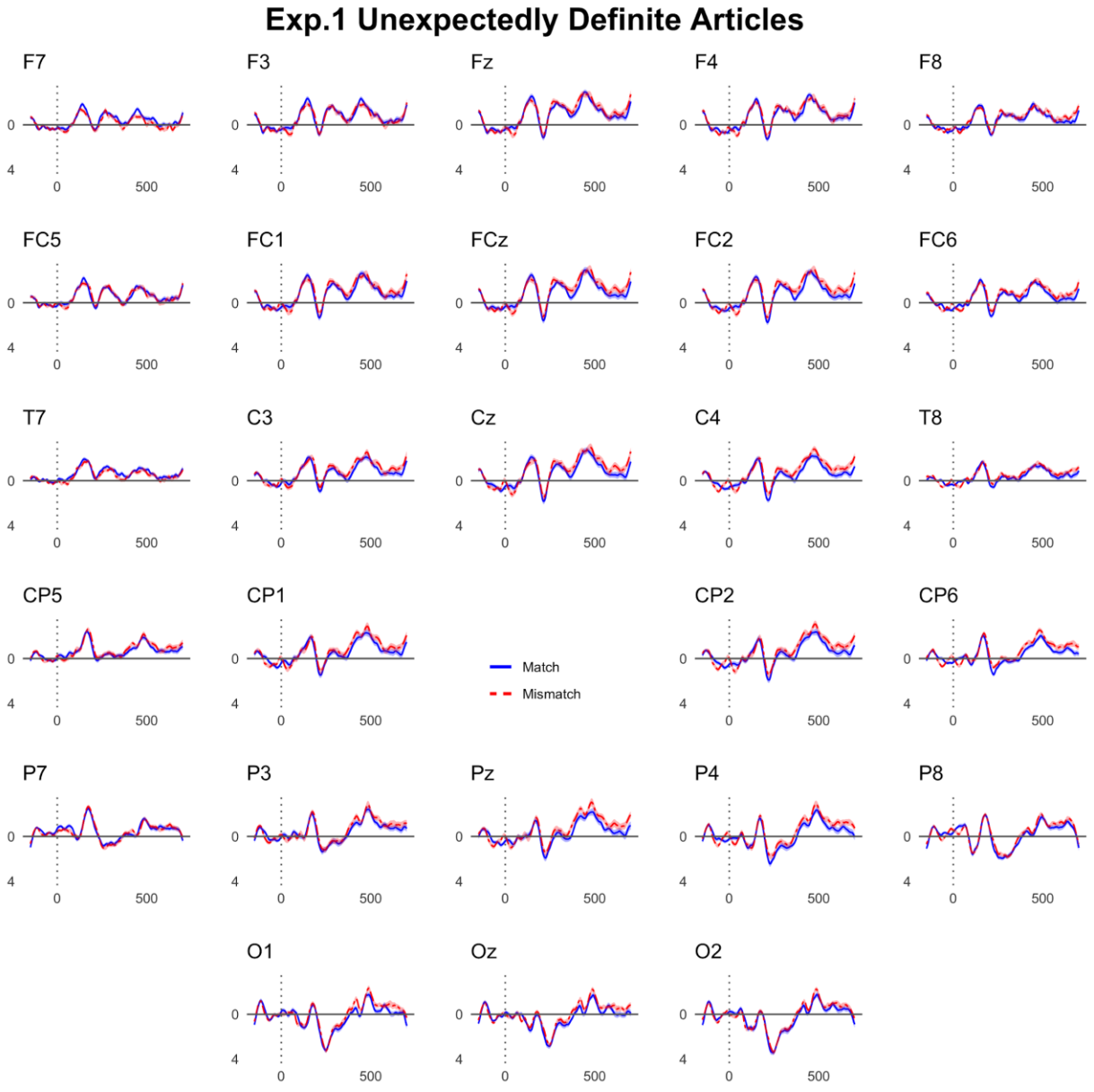
Acknowledgements

We thank Birgit Knudsen for help with EEG data collection and Elise van Wonderen for help with norming data collection. JR was partially supported by NWO Veni grant 275-89-032. We thank Tamara Swaab and two anonymous reviewers for providing helpful feedback on a previous draft of this manuscript. All materials associated with the current article are available on OSF:https://osf.io/6drcy/. For the analyses and plots, we used the following packages for R (R Core Team, 2018): "brms" (Bürkner, 2017), "lme4" (Bates, Maechler, Bolker & Walker, 2015), "simr" (Green & MacLeod, 2016), (Fox & Weisberg, 2019), "ggplot2" (Wickham, 2016), "dplyr" (Wickham, François, Henry & Müller, 2019), "patchwork" (Pederson, 2020), "emmeans" (Lenth, 2019).

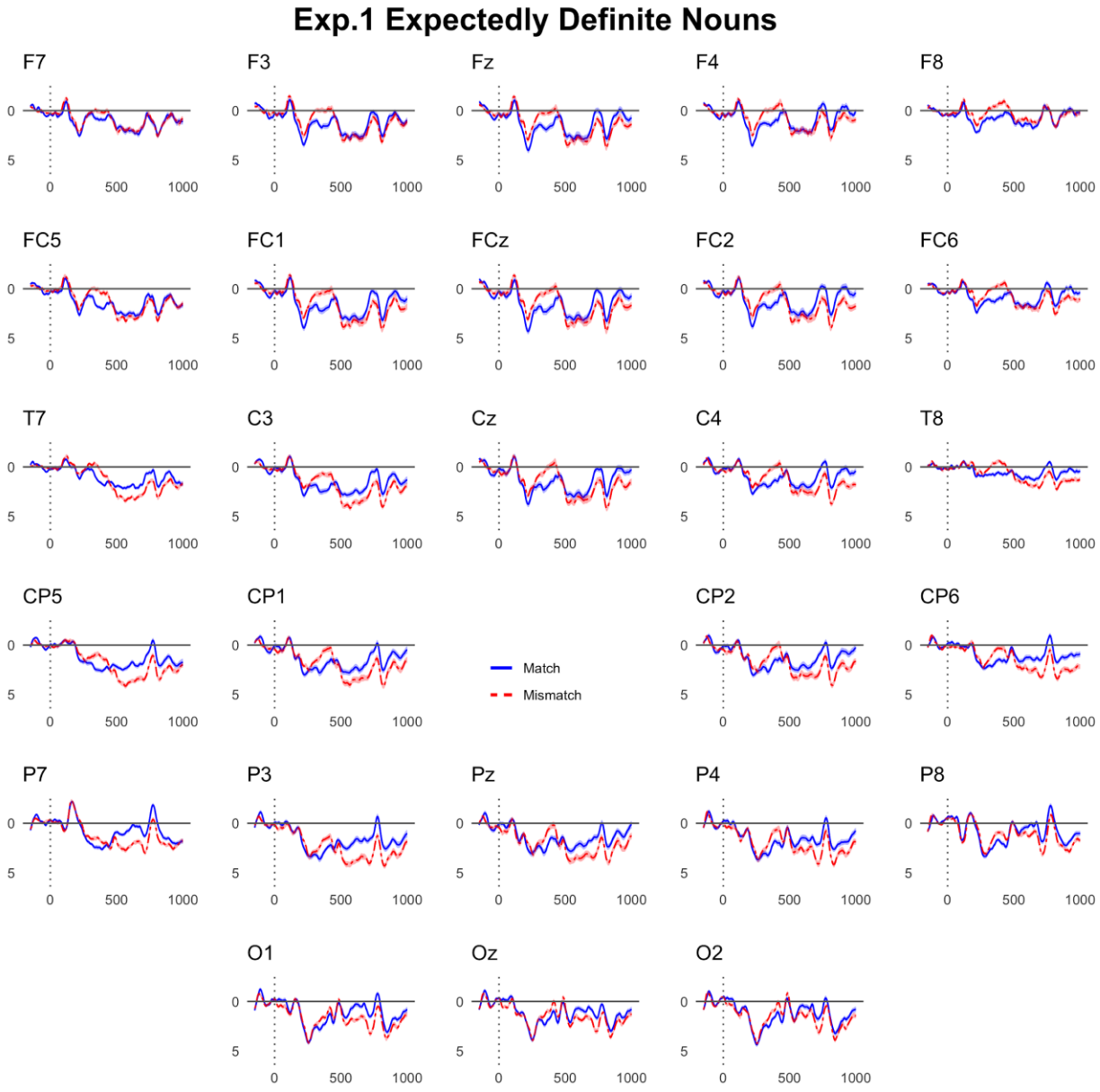
Supplementary Figure 1. Grand-average ERPs for expectedly definite articles at all channels in Experiment 1.



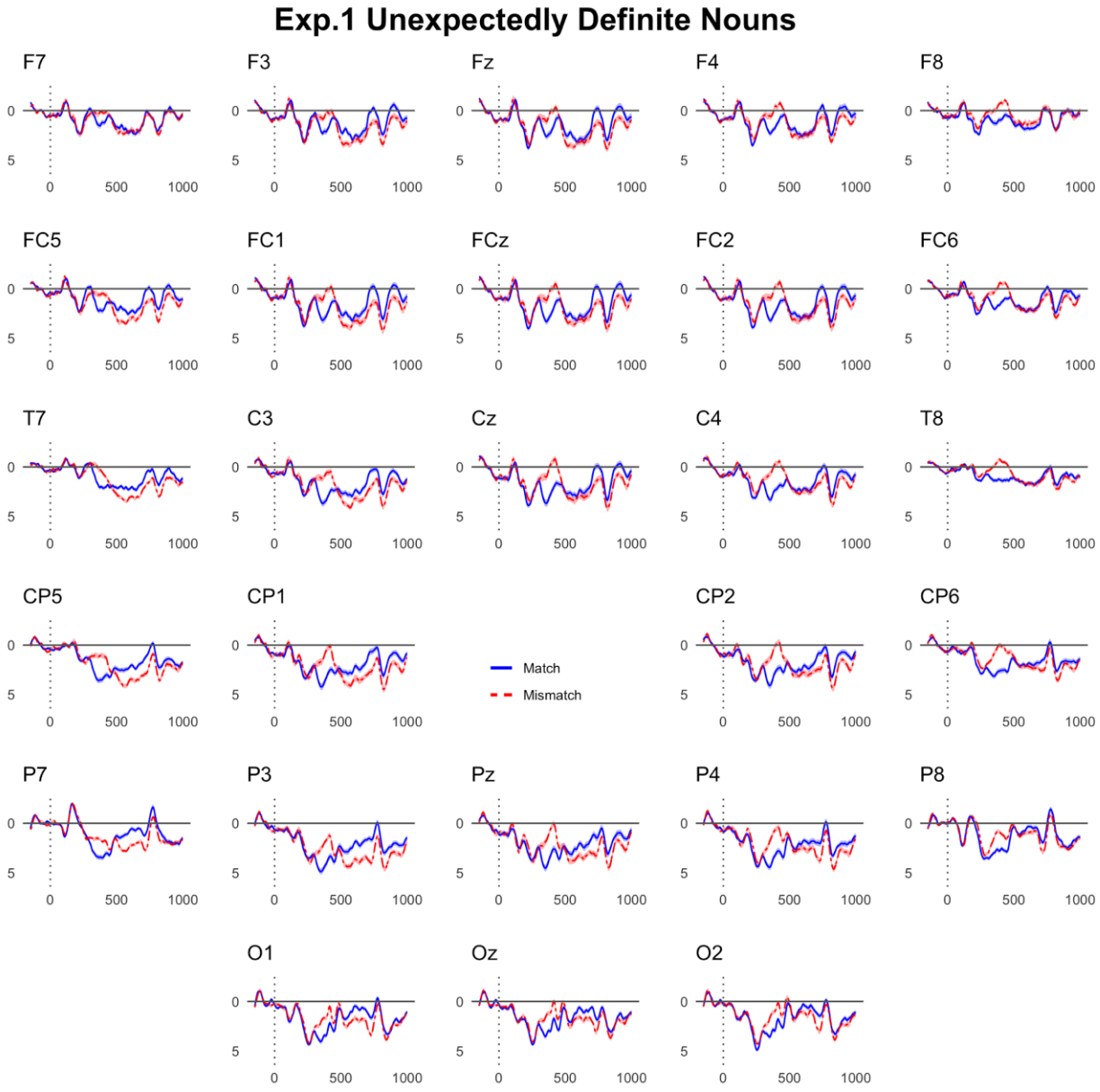
Supplementary Figure 2. Grand-average ERPs for unexpectedly definite articles at all channels in Experiment 1.



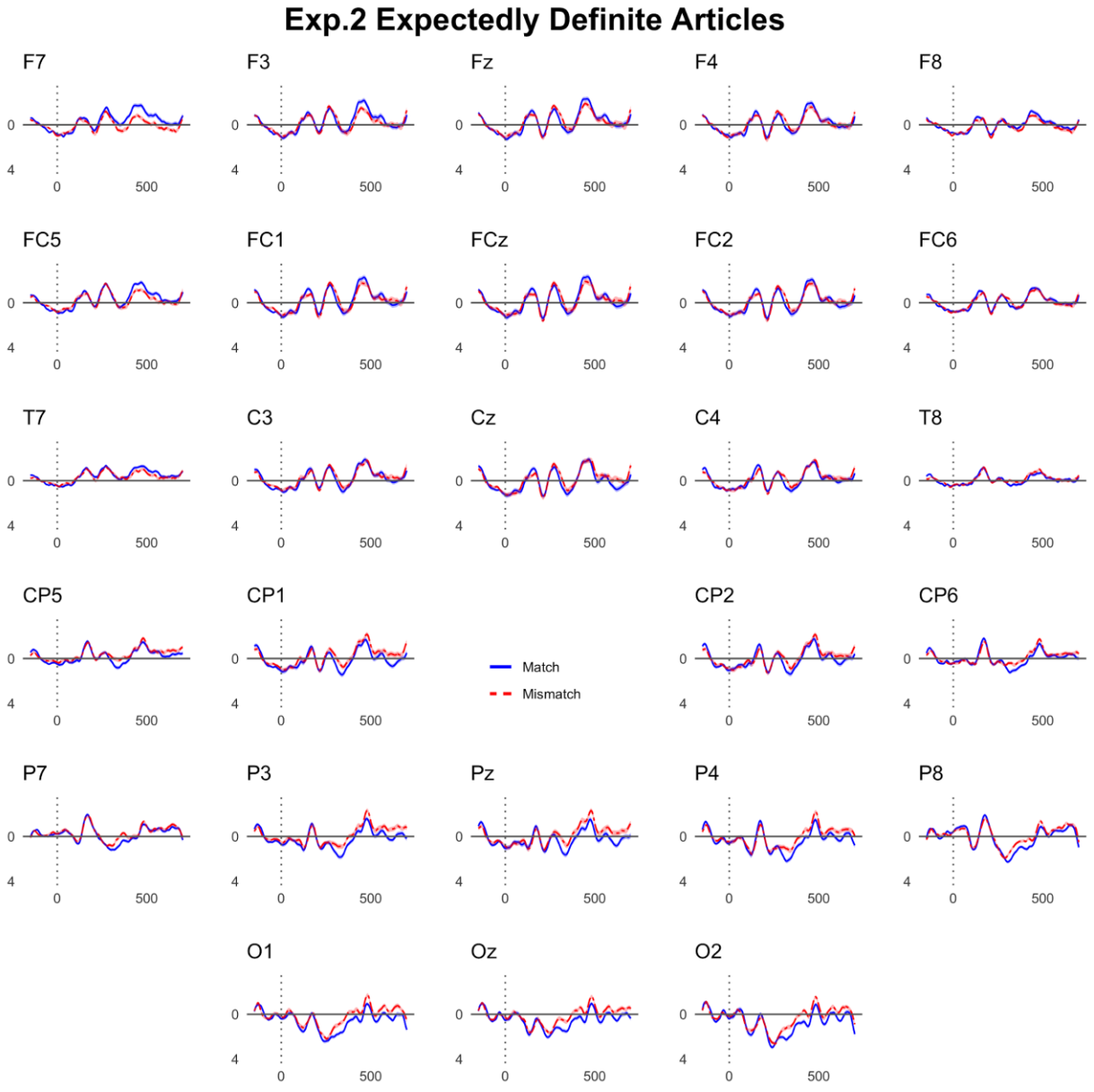
Supplementary Figure 3. Grand-average ERPs for expectedly definite nouns at all channels in Experiment 1.



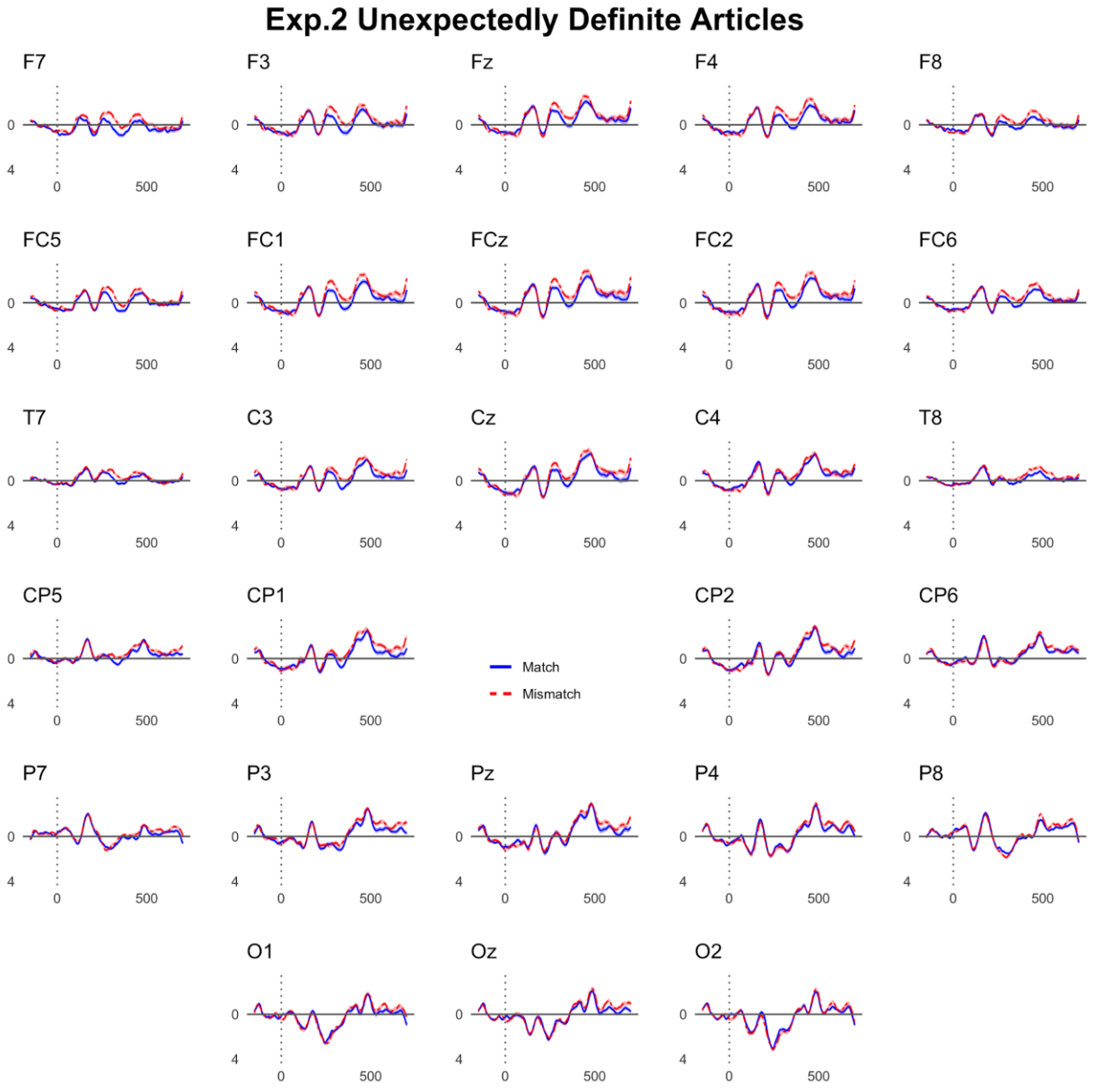
Supplementary Figure 4. Grand-average ERPs for unexpectedly definite nouns at all channels in Experiment 1.



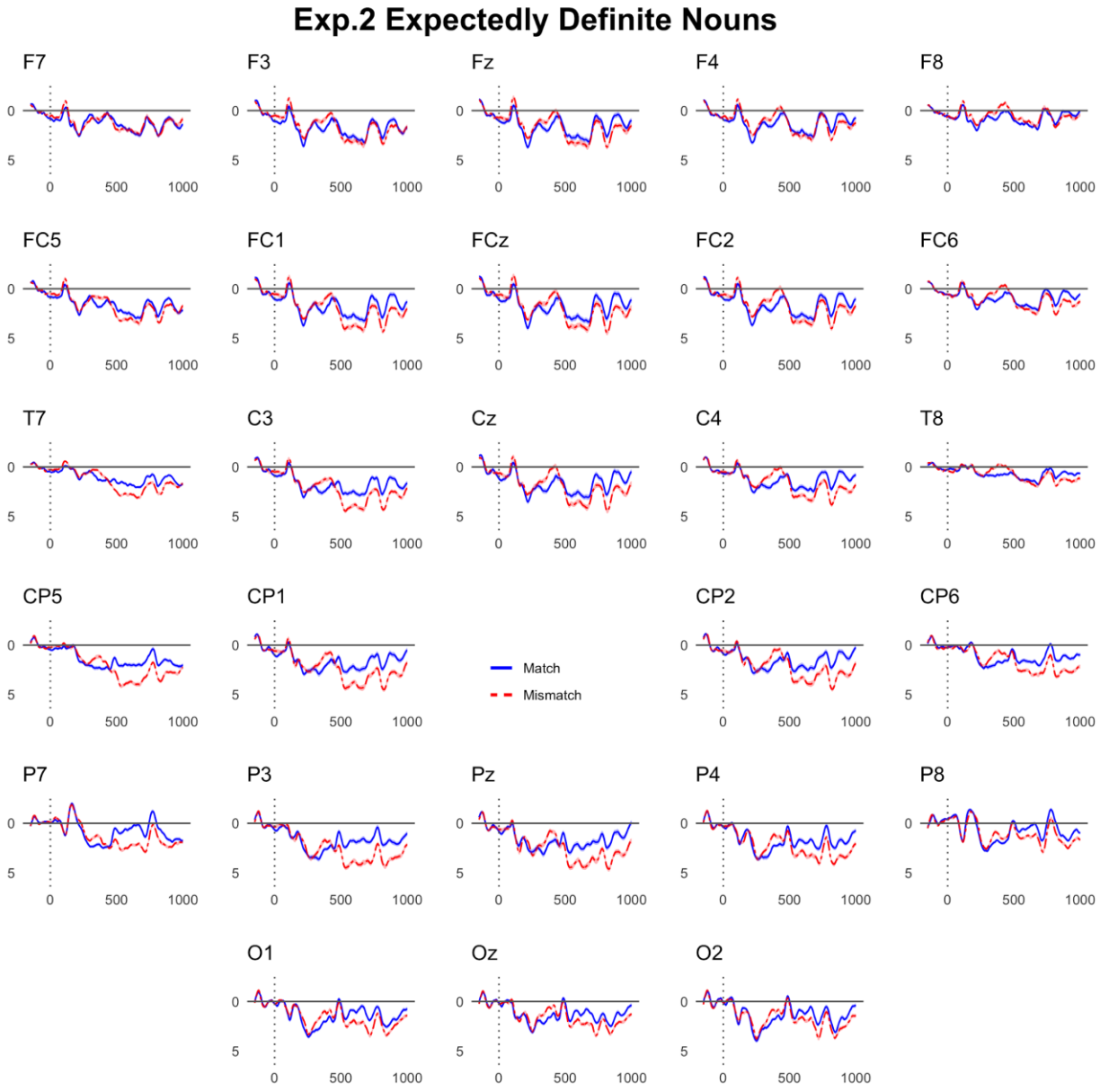
Supplementary Figure 5. Grand-average ERPs for expectedly definite articles at all channels in Experiment 2.



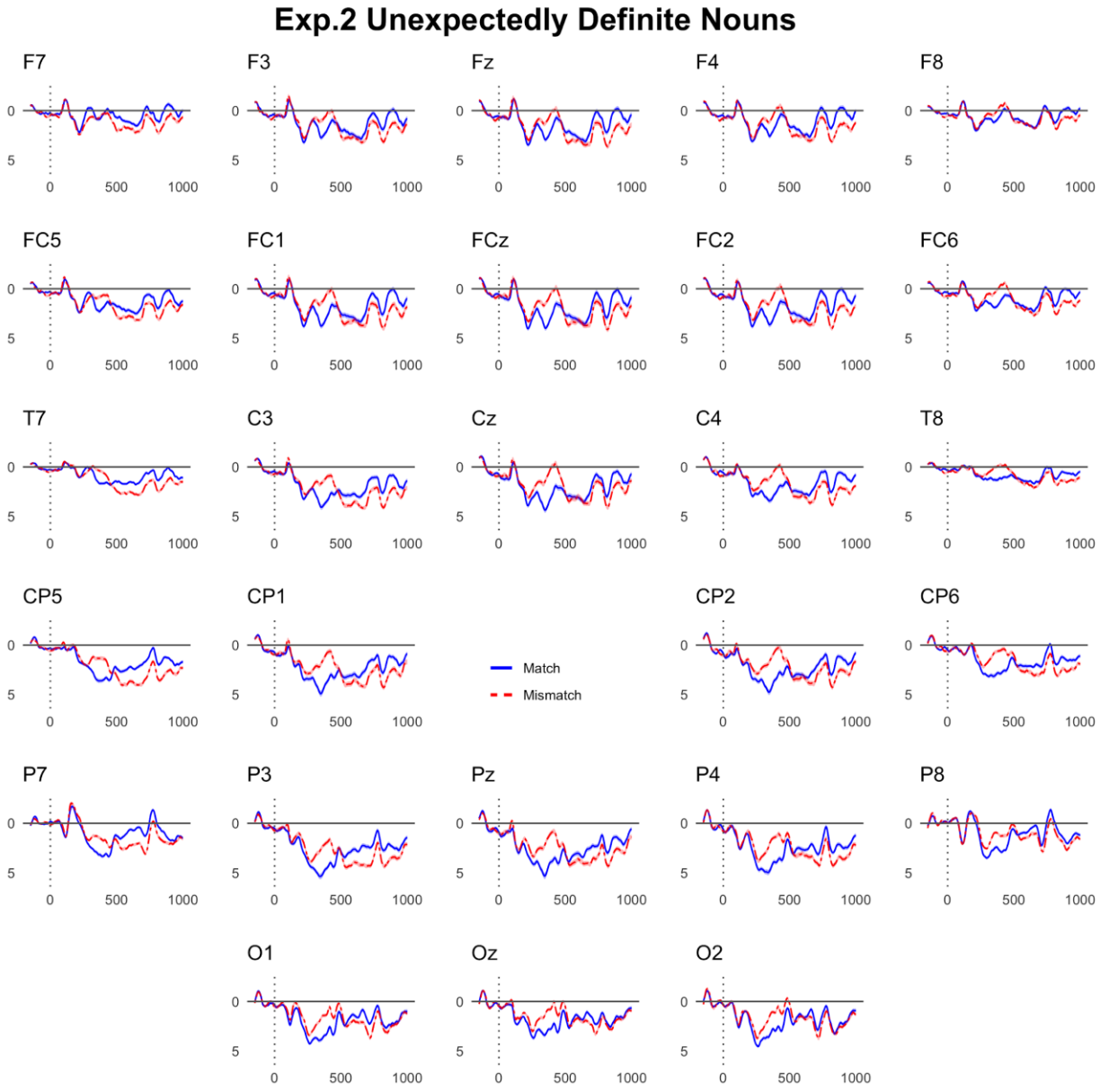
Supplementary Figure 6. Grand-average ERPs for unexpectedly definite articles at all channels in Experiment 2.



Supplementary Figure 7. Grand-average ERPs for expectedly definite nouns at all channels in Experiment 2.



Supplementary Figure 8. Grand-average ERPs for unexpectedly definite nouns at all channels in Experiment 2.



Appendix. Exploratory tests for 'de' versus 'het'.

Unlike previous research (Kochari & Flecken, 2018; Otten & Van Berkum, 2009), our pre-registered analyses explicitly controlled for the general effect of the article ('de/het'). However, like previous research, our analyses did not take into account potential interactions between article form and the effects of interest. Such interactions could be relevant, because the articles may be used in different ways during predictive processing (e.g., Brouwer, Sprenger & Unsworth, 2017; Loerts, Wieling & Schmid, 2013, for examples using the visual-world eye-tracking methodology). Therefore, we performed exploratory analyses to compare the gender-mismatch effects for 'de' and 'het'. First, we repeated the article analyses with a three-way interaction term between 'article-form' (de/het), 'gender' (expected/unexpected) and 'definiteness' (expected/unexpected), with all deviation-coded factors. We here report these analyses for the combined data sets (N=128).

Analyses for the 5 ROIs (shown in Figure A1) indeed suggested that the interaction between definiteness and gender depended on article-form in some of the ROIs (anterior, 300-500 ms: $\beta = 0.95$, SE = 0.47, t=1.99 , p = 0.046 ; 500-700 ms: $\beta = 0.93$, SE = 0.51 , t= 1.83 , p = 0.07; posterior, 300-500 ms: $\beta = 0.69$, SE = 0.44 , t= 1.55 , p = 0.12; 500-700 ms: $\beta = 1.06$, SE = 0.48 , t= 2.23 , p = 0.026; occipital, 300-400 ms: $\beta = 0.32$, SE = 0.37 , t= 0.87 , p = 0.38). We performed follow-up analyses for expectedly definite articles only, for which the interaction pattern of interest was not confounded by expected definiteness. At anterior ROIs, the gender-mismatch effect went into opposite directions for 'de' and 'het', with enhanced negativity for 'het' and a positivity for 'de', both in the 300-500 ms time window (interaction $\beta = -0.88$, S.E. = 0.41, Z = -2.15, p =0.031; de: $\beta = -0.47$, S.E. = 0.27, p =0.083; het, $\beta = 0.44$, S.E.= 0.27, p

=0.10) and the 500-700 ms time window (interaction β = -0.89, S.E. = 0.43, Z = -2.05, p =0.040; de: β = -0.28, S.E. = 0.28, p =0.32; het, β = 0.61, S.E.= 0.28, p = 0.03). At posterior ROIs, the gender-mismatch effect was stronger for 'het' than for 'de' in the 300-500 ms time window (interaction β = -0.58, S.E. = 0.36, Z = -1.62, p =0.11; de: β = 0.33, S.E. = 0.24, p =0.17; het, β = 0.90, S.E.= 0.24, p <.0001) and the 500-700 ms time window (interaction β = -0.91, S.E. = 0.39, Z = -2.31, p =0.021; de: β = 0.16, S.E. = 0.26, p =0.55; het, β = 1.06, S.E.= 0.26, p <.0001). At the occipital ROI, the strong gender-mismatch effect differed little between 'de' and 'het' (interaction β = -0.15, S.E. = 0.27, Z = -0.55, p =0.58; de: β = 0.80, S.E. = 0.19, p <.0001; het, β = 0.94, S.E.= 0.19, p <.0001).

In sum, while the occipital ROI did not appear sensitive to which article elicited the mismatch effect, the posterior (N400) ROI showed greater sensitivity for 'het' compared to 'de', whereas the anterior ROI differentiated between 'het' and 'de' by showing effects in different directions. The different mismatch effects for 'de' and 'het' depended primarily on differences between matching conditions, and we therefore do not claim that processes associated with prediction mismatch or prediction revision differed between these articles. In addition, we emphasize that our study was not designed with these analyses in mind, and the associated results should be interpreted with caution. The contexts with common-gender or neuter-gender nouns as best completions may have differed in unknown but relevant ways, for example in the ERP responses associated with the words preceding the articles, which could then distort the article-elicited ERPs. Also relevant, because the ratio of common/neuter gender predictable nouns in our study matched the higher frequency of 'de' compared to 'het' in natural language corpus counts (Van Berkum, 1997), participants in our gender-mismatch design saw more

gender-mismatching ‘het’ than ‘de’ articles. At this point, it is unclear whether the results indeed reflect a genuine processing difference for common and neuter gender, as has been reported for other paradigms (e.g., Deutsch & Wijnen, 1985).

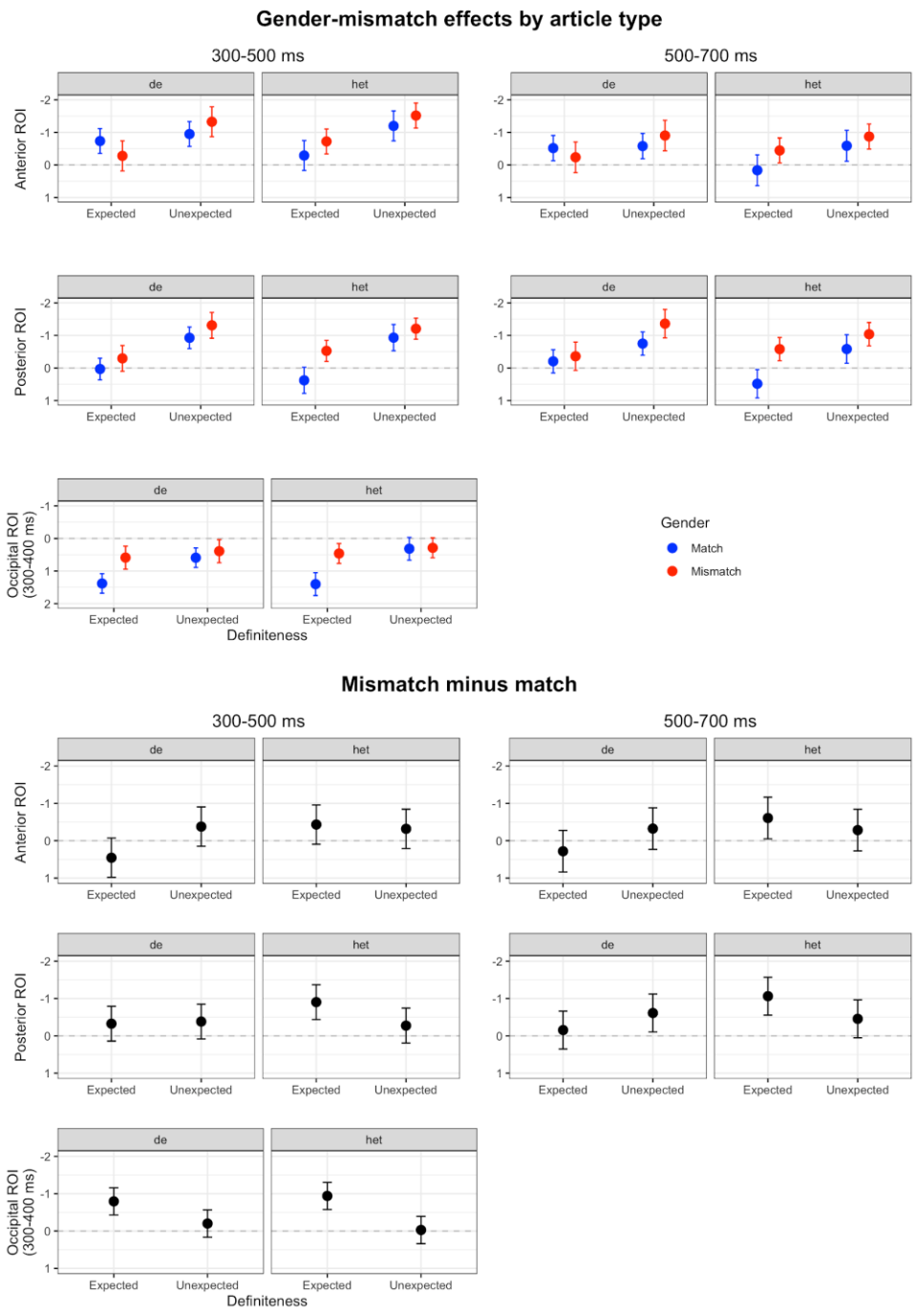


Figure A1. Gender-mismatch effects for 'de' and 'het' at all ROIs, based on combined data from Experiment 1 and 2. Upper graphs show mean voltage (μV) and confidence interval per condition, bottom graphs show the mean difference (mismatch minus match) and confidence interval.