From oscillations to language: behavioural and electroencephalographic studies on cross-language interactions

Von Grebmer Zu Wolfsthurn, S.

CHAPTER 8

Neural correlates of gender agreement processing in Spanish: P600 or N400?

This article was submitted for review as: Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (2022). Neural correlates of gender agreement processing in Spanish: P600 or N400?. Brain and Language.

Abstract: The P600 component was previously established as a robust index for syntactic processing, particularly with respect to gender agreement violations. However, studies showing an N400 component for gender agreement violations in isolated noun-phrases in Spanish challenge this particular interpretation. Here, we measured event-related potentials during a syntactic violation paradigm to examine the neural correlates of gender agreement violations in determiner-noun pairs in Spanish, e.g., ([*el nube] vs. [la nube] - *the cloud*). Based on previous literature, we predicted larger P600 amplitudes for gender violations [*el nube] vs. correct pairs [la nube]. However, we also probed a potential N400 component for violations. We used generalised additive mixed models to flexibly model voltage amplitudes over time. Results showed a P600 effect for gender agreement violations compared to non-violations, but

no evidence for an N400 effect. These results have critical implications for characterising the underlying neural correlates of gender agreement processing in Spanish.

Keywords: *native language comprehension, gender agreement violations, event-related potentials, P600 component, N400 component, generalised additive mixed models*

8.1 Introduction

Language comprehension is a remarkably complex process because it involves not one, but several simultaneous encoding and integration processes (Friederici et al., 2004; Nieuwland, 2019; Skeide & Friederici, 2017; Sung, Yoo, Lee & Eom, 2017; Walenski, Europa, Caplan & Thompson, 2019). To that end, studying imperfect or "faulty" language input is fundamental in characterising the different cognitive processes underlying language comprehension. Grammatical gender, the focus of this study, has been a particularly suitable candidate for exploring these processing mechanisms and the corresponding neural correlates, especially in the context of gender agreement violation paradigms (Barber & Carreiras, 2005; Beatty-Martínez, Bruni, Bajo & Dussias, 2021; Hasting & Kotz, 2008; Neville, Nicol, Barss, Forster & Garrett, 1991; Osterhout & Mobley, 1995). Grammatical gender (hereafter *gender*) operates as a classification system for nouns (Corbett, 1991). It is considered both an abstract lexical and syntactic feature (Cantone & Müller, 2008; Corbett, 1991; Schriefers & Jescheniak, 1999; Sá-Leite et al., 2019). Spanish, the target language in our study, is characterised by a two-value gender system with a masculine and feminine gender value. Here, the definite determiner *la* marks the feminine gender value, e.g., [la_f nube_f] *the cloud*, whereas the definite determiner *el* marks the masculine gender value, e.g., [el_m libro_m] *the book*.

In this study, we examined the neural correlates of processing gender agreement violations within determiner-noun phrases (hereafter NPs) such as [la nube] vs. [*el nube] in Spanish. For this, we combined a gender agreement violation paradigm with electroencephalography (EEG) and event-related potentials (ERPs). The use of ERPs has been paramount for characterising the neural correlates of gender agreement, in particular within sentences (Friederici et al., 1999; Kaan, 2007; Kotz, Holcomb & Osterhout, 2008; Osterhout & Nicol, 1999; Steinhauer & Connolly, 2008; Swaab et al., 2011). However, as will be discussed below, the specific neural underpinnings of processing gender agreement violations in isolated NPs such as [la nube] vs. [*el nube] remain debated in the literature. This is particularly the case for Spanish and for isolated determiner-noun NPs that are examined outside of a sentence context. This is a critical issue because it taps directly into the broader question of whether gender agreement processing is qualitatively different for isolated NPs compared to NPs within sentences. The following sections provide an overview of the neural correlates of gender agreement processing in NPs, with a particular focus on Spanish as the target language in this study.

8.1.1 P600 and LAN effects for noun-phrase violations

Prior research has identified two primary ERP components relevant to gender agreement processing in NPs: the P600 component and the left anterior negativity (LAN) (Hasting & Kotz, 2008; Swaab et al., 2011). The P600 component is a positive-going oscillation associated with a broad topographic distribution in a time window between 500 ms to 900 ms and with a peak around 600 ms post-event onset (Friederici et al., 1999; Osterhout & Holcomb, 1992; Steinhauer et al., 2009). Research also suggested that the P600 component can be further divided into two functionally different stages: an early stage between 500 ms to 700 ms with a broad topographic distribution linked to syntactic integration; and a later stage between 700 ms to 900 ms with a centro-parietal topographic distribution linked to syntactic re-analysis and repair (Alemán Bañón et al., 2012; Barber & Carreiras, 2005; Hagoort & Brown, 2000; Molinaro et al., 2008). The so-called P600 effect is reflected in higher voltage amplitudes for gender agreement violations compared to syntactically correct structures. For example, using a probe verification task, Gunter et al. (2000) found higher P600 voltage amplitudes in ungrammatical German sentences containing a gender agreement violation at the determiner-noun level such as in the example in [1], compared to grammatical sentences as in the example in [2]. Similar P600 effects were also reported in Italian by Molinaro et al. (2008) for gender agreement violations at the determiner-noun level [3] compared to non-violations [4]; see also Hagoort and Brown (1999) for comparable findings with Dutch sentences. More relevant to this study, Barber and Carreiras (2005) reported a P600 effect in Spanish for determiner-noun gender agreement violations [5] compared to grammatical structures [6]; see also Wicha et al. (2004) for similar findings in Spanish.

[1] *Sie bereist* **den_m $Land_n$* *auf einem kräftigen Kamel.*

[She travels *the_m land_n on a strong camel]

[2] *Sie bereist* *das_n $Land_n$* *auf einem kräftigen Kamel.*

[She travels the_n land_n on a strong camel]

[3] *Le olive farcite con* **la_f $peperone_m$* *sono ottime.*

[The olives stuffed with *the_f bell pepper_m are excellent]

[4] *Le olive farcite con* *il_m $peperone_m$* *sono ottime.*

[The olives stuffed with the_m bell pepper_m are excellent]

[5] **La_f $piano_m$* *estaba viejo y desafinado.*

[*The_f piano_m was old and off-key]

[6] *El_m $piano_m$* *estaba viejo y desafinado.*

[The$_m$ piano$_m$ was old and off-key].

The second component previously reported in the context of gender agreement violations is the left anterior negativity (LAN). It is a negative-going wave linked to a left anterior topographic distribution between 300 ms and 500 ms, although some studies also reported a broader distribution (Coulson, King & Kutas, 1998; Kaan, 2007; Martín-Loeches et al., 2005; Molinaro, Barber, Caffarra & Carreiras, 2015; Neville et al., 1991; Osterhout & Mobley, 1995; Padrón, Fraga & Acuña-Fariña, 2020). The LAN effect, i.e., more negative amplitudes for syntactic violations compared to non-violations, is commonly interpreted as reflecting early automatic syntactic processing. Subsequently, it was frequently reported as a pre-cursor to the P600 effect to form a biphasic LAN/P600 pattern in studies on gender agreement processing (Barber & Carreiras, 2005; Kaan, 2007; Molinaro et al., 2008; Steinhauer & Drury, 2012). However, results have varied in that respect, with some studies failing to provide evidence for a LAN effect for gender agreement violations (Hagoort, 2003; Wicha et al., 2004). Therefore, the specific circumstances under which a LAN effect is elicited in combination with a P600 effect in gender agreement violation contexts are still subject to debates, see Alemán Bañón et al. (2012) and Molinaro et al. (2011) for discussions.

8.1.2 ERP effects for noun-phrase violation processing in Spanish

The general consensus from the studies discussed above is that the P600 component is a reflection of processes connected to advanced syntactic processing such as syntactic integration and structural re-analysis or repair, for example for processing gender agreement violations in determiner-noun NP constructions (Hagoort et al., 1993; Osterhout & Holcomb, 1992; Steinhauer & Drury, 2012). In these studies, the P600 effect emerged when processing a gender agreement violation [*el$_m$ nube$_f$] compared to [la$_f$ nube$_f$] *the cloud* (Barber & Carreiras, 2005). Moreover, as pointed out before, the

LAN effect can precede the P600 effect, indexing earlier syntactic processes during gender agreement violation (Barber & Carreiras, 2005; Molinaro et al., 2008). However, this interpretation of the P600 component and its relevance to gender agreement processing is challenged by studies focusing on Spanish as the target language and when the processing of isolated NPs instead of sentence-embedded NPs is examined. For example, a study by Barber and Carreiras (2003) investigated gender and number agreement violations in Spanish adjective-noun pairs. The following four conditions were tested: a syntactic violation of gender [7], a syntactic violation of number [8], a double syntactic violation of gender and number [9], and a control condition [10]. Participants judged the grammaticality of each pair in this task after the noun and adjective were shown sequentially on the screen.

[7] **faro*$_{m-sg}$ *alta*$_{f-sg}$ [*lighthouse$_{m-sg}$ high$_{f-sg}$]
[8] **faro*$_{m-sg}$ *altos*$_{m-pl}$ [*lighthouse$_{m-sg}$ high$_{m-pl}$]
[9] **faro*$_{m-sg}$ *altas*$_{f-pl}$ [*lighthouse$_{m-sg}$ high$_{f-pl}$]
[10] *faro*$_{m-sg}$ *alto*$_{m-sg}$ [lighthouse$_{m-sg}$ high$_{m-sg}$]

Results showed that the gender, number and the double-violation condition elicited more negative amplitudes compared to the control condition between 300 ms and 500 ms post-stimulus onset in centro-parietal regions. This is a pattern consistent with the so-called N400 effect. As one of the most well-studied ERP components, the N400 component has been fundamental in contributing to our current understanding of language processing (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011; Swaab et al., 2011). Broadly speaking, the N400 component has a negative oscillatory tendency and is generally located in centro-parietal regions. It usually peaks around 400 ms post-event onset, with a component latency between 300 ms and 500 ms (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011; Lau et al., 2008; Molinaro et al., 2015; Van Petten, Kutas, Kluender, Mitchiner & McIsaac, 1991). The N400 effect, i.e., more negative voltage amplitudes for structures with semantic violations, is typically reported for sentences such as *"I take my coffee with milk and <u>*dog</u>"*, compared to semantically plausible sentences such as *"I take my coffee with milk and sugar"* as shown in the seminal study by Kutas and Hillyard (1980). Therefore, the findings from Barber and Carreiras (2003) are unexpected in that they link the N400 effect to processing gender agreement violations in isolated adjective-noun pairs in Spanish, but not the P600 effect. Importantly, similar results were reported in a subsequent two-part study by Barber and Carreiras (2005). In Experiment 1 of their study, the authors explored gender and number violations in Spanish speakers in isolated determiner-noun pairs such as *the piano*, and in noun-adjective pairs such as *tall lighthouse*; the latter as examined in Barber and Carreiras (2003). The design included a gender violation condition for determiner-noun pairs [11], a number violation condition for determiner-noun pairs [12] and a control condition [13]; and similarly, a gender violation condition for adjective-noun pairs as [14], a number violation condition for adjective-noun pairs [15] and a control condition [16].

[11] **la*$_{f-sg}$ *piano*$_{m-sg}$ [*the$_{f-sg}$ piano$_{m-sg}$]
[12] **los*$_{m-pl}$ *piano*$_{m-sg}$ [*the$_{m-pl}$ piano$_{m-sg}$]
[13] *el piano*$_{m-sg}$ [the$_{m-sg}$ piano$_{m-sg}$]
[14] **faro*$_{m-sg}$ *alta*$_{f-sg}$ [*lighthouse$_{m-sg}$ tall$_{f-sg}$]
[15] **faro*$_{m-sg}$ *altos*$_{m-pl}$ [*lighthouse$_{m-sg}$ tall$_{m-pl}$]
[16] *faro*$_{m-sg}$ *alto*$_{m-sg}$ [lighthouse$_{m-sg}$ tall$_{m-sg}$]

Relevantly, in Experiment 2 from Barber and Carreiras (2005), the authors embedded the determiner-noun pairs and adjective-noun pairs in sentences. Example stimuli sentences were *"*La*$_{f-sg}$ *piano*$_{m-sg}$ *estaba viejo y desafinado"* [*The$_{f-sg}$ piano$_{m-sg}$ was old and off-key] for the gender violation condition for determiner-noun pairs, and *"*El*$_{m-sg}$ *faro*$_{m-sg}$ *es alta*$_{f-sg}$ *y luminoso"* [*The$_{m-sg}$ lighthouse$_{m-sg}$ is tall$_{f-sg}$ and bright] for the gender violation condition for adjective-noun pairs. In half of the trials, the violations were placed at the beginning of the sentence, and in the middle of the sentence for the remaining half of the trials. Results from Experiment 1 revealed a broadly distributed N400 effect for the violation conditions in centro-parietal regions between 300 ms and

500 ms. This was in line with previous findings by Barber and Carreiras (2003). In contrast, results from Experiment 2 showed a LAN effect for both pairs, followed by a P600 effect. Taken together, the results from Experiment 1 do not support the involvement of the P600 component in processing gender agreement violation in isolated NPs. On the other hand, results from Experiment 2 favour the classical interpretation of a biphasic LAN/P600 effect connected to gender agreement processing in NPs (Hagoort & Brown, 1999; Steinhauer & Drury, 2012).

In sum, most of the studies presented above yield relatively homogeneous findings: first, studies robustly yield the P600 effect for gender agreement violations in a sentence context (Barber & Carreiras, 2005; Gunter et al., 2000; Hagoort et al., 1993), reflecting the involvement of the P600 in syntactic processing. However, this general interpretation of the P600 effect and its connection to gender agreement processing is called into question by studies where the violation is in an isolated NP instead of a sentence context: studies on gender agreement processing in isolated NPs such as [*la$_f$ piano$_m$] *the piano* in Spanish have been unique in that they yielded an N400 effect for agreement violations (Barber & Carreiras, 2003, 2005). Therefore, and to expand on existing research on the neural correlates of gender agreement violations in NPs in Spanish, we employed a syntactic violation paradigm with NPs containing a violation at the determiner-noun level. In this, our study has important implications for the neural underpinnings of processing gender agreement violations in Spanish: we first explored whether there were different ERP components linked to gender processing in isolated NPs; and second, we investigated whether the P600 component was the primary ERP component linked to gender agreement violation processing in both isolated NPs and NPs within a sentence context, or whether the N400 effect was also at play. In this, we also probed for the presence of a LAN effect.

8.1.3 The current study

The focus of the current study was to explore the EEG signal underlying the processing of gender agreement violations within NPs in Spanish. More specifically, we probed the elicitation of the P600 effect for gender agreement violations compared to non-violations, as well as LAN and N400 effects. Here, we employed a syntactic violation paradigm and presented participants with determiner-noun NPs that were either correct, i.e., *non-violations* (la_f $nube_f$ [the cloud]) or incorrect, i.e., *violations* (*el_m $nube_f$), while we measured their EEG. We specifically opted for NPs to exclusively focus on the ERP components linked to the processing of syntactic violations in the absence of sentence-related contextual effects. In addition, we asked participants to fill in the LEAP-Q, which is a language proficiency and experience questionnaire used to describe participants' linguistic profile (Marian et al., 2007). This was done to accurately capture their exposure to other languages besides Spanish because participants were tested in a non-native environment. Participants also completed the LexTALE-Esp, a Spanish vocabulary size task (Izura et al., 2014) which we used as a covariate in the analyses.

Research questions and hypotheses

The main research question of this study was the following: what are the ERP components linked to the processing of syntactically correct vs. incorrect Spanish NPs in Spanish native speakers? Behaviourally, we predicted higher accuracy and shorter response times (RTs) for *non-violation* trials compared to *violation* trials. More importantly, based on the previous literature, we predicted that *violations* would elicit more positive P600 amplitudes compared to *non-violations*, thereby generating a P600 effect. Further, and to be consistent with results from prior studies, we investigated whether the P600 component was elicited as an isolated effect, within a biphasic LAN/P600, or if instead we found evidence for an N400 effect. A LAN effect would be reflected in more negative amplitudes for violations compared to non-violations in left anterior regions, whereas an N400 effect would be reflected in more negative amplitudes for violations compared to non-violations in centro-parietal regions.

8.2 Methods

8.2.1 Participants

We recruited 40 native Spanish speakers (28 females) for this study, in line with previous work (Barber & Carreiras, 2005; Von Grebmer Zu Wolfsthurn et al., 2021a). Participants were between 18 and 35 years old with $M = 28.00$ years of age ($SD = 3.92$). Eligibility criteria included the absence of psychological or neurological disorders, no language or reading impairments, no second language learnt before the age of three, normal or corrected-to-normal vision and hearing and right-handedness. Using the LEAP-Q, we determined that all participants acquired at least one additional language to Spanish at the time of testing. See Appendix 8.A for an overview of the languages acquired by the participants. For the analyses, we included a total of 34 participants, 22 of which were female (see section 8.3.3 for data exclusion). Mean age of the included participants was $M = 27.85$ years ($SD = 3.93$). On average, participants started acquiring Spanish at $M = 0.265$ years of age ($SD = 0.567$). They reported being fluent in Spanish at $M = 3.40$ years of age ($SD = 1.85$), and started reading in Spanish around $M = 4.87$ years of age ($SD = 1.59$) before reaching reading fluency at $M = 6.90$ years of age ($SD = 1.62$). Participants' daily exposure to Spanish was $M = 41\%$ ($SD = 17.34$) compared to $M = 43.64\%$ ($SD = 17.48$) for their first foreign language (L2) and $M = 13.38\%$ ($SD = 17.33$) for their second foreign language (L3). Self-reported proficiency in Spanish was $M = 9.82$ ($SD = 0.459$) for oral production, $M = 9.91$ ($SD = 0.288$) for aural comprehension and $M = 9.85$ ($SD = 0.359$) for written comprehension on a scale from one to ten (ten indicating the highest proficiency).

8.2.2 Materials and design

Stimuli for the LexTALE-Esp were identical as described in Izura et al. (2014), with the difference that we converted the manual version of this task into an E-prime2 script (Schneider et al., 2002). The stimuli for the syntactic violation paradigm consisted of 224 highly frequent Spanish nouns and their corresponding definite determiners. The stimuli nouns were selected from the MultiPic database (Duñabeitia et al., 2018) and the Spanish Frequency Dictionary (Davies & Davies, 2017). Nouns followed a balanced masculine:feminine gender value ratio and were controlled for frequency and syllable length.

8.2.3 Procedure

Before the experimental session, participants were instructed to complete the LEAP-Q (Marian et al., 2007), for which the results were reported in section 8.2.1. The experimental session took place at the Leiden University Linguistics Laboratories. At the beginning of the session, we provided participants with an information sheet in Spanish and the opportunity to ask questions. Next, in compliance with the ethics code for linguistics research at the Faculty of Humanities at Leiden University, participants were asked to fill out a consent form. During the session, participants completed the LexTALE-Esp to determine their vocabulary size in Spanish (Izura et al., 2014) and the syntactic violation paradigm. For both tasks, we placed participants in front of a computer screen inside a shielded EEG booth. Participants were instructed to sit as still as possible and to avoid any unnecessary contamination of the EEG signal. The response device was a Chronos® box (Psychology Software Tools, Inc). We provided oral and written instructions prior to each task. After completing the session, participants signed the final consent form before receiving a debrief form and a monetary compensation.

LexTALE-Esp

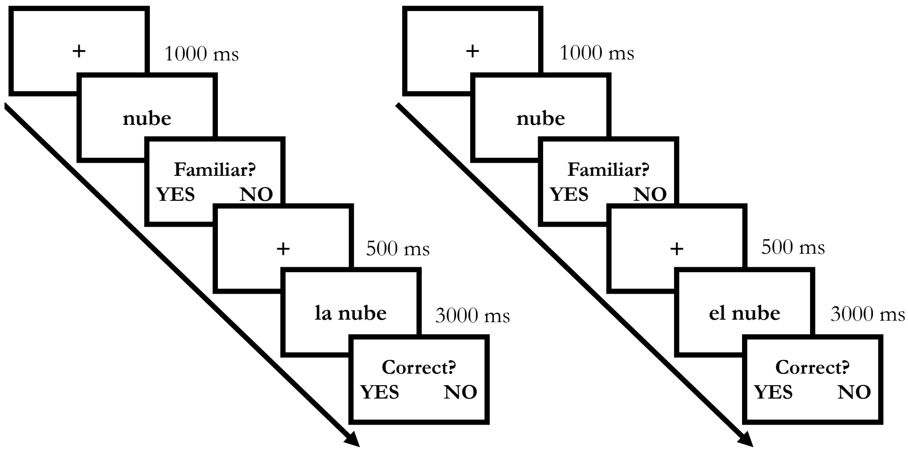
Participants were asked to make a lexical decision whether the letter string displayed on the screen corresponded to a Spanish word or a pseudoword, identical to what was described in Von Grebmer Zu Wolfsthurn et al. (2021a). Following the display of a fixation cross for 500 ms, the letter string remained on the screen until the participant's response. Post-task, we computed LexTALE-Esp vocabulary size scores (*LexTALE-Esp score*) to account for potential differences in terms of vocabulary during the analyses (Izura et al., 2014). Mean LexTALE-Esp scores were M = 77.73 (SD = 18.87), with a maximal score of 100.

Syntactic violation paradigm

We closely modelled this task procedure after earlier work by Von Grebmer Zu Wolfsthurn et al. (2021a) and recorded participants' EEG signal during this task. The core feature of the task was the visual presentation of an NP on the screen. Each NP consisted of a determiner and a noun in Spanish. Half of the NPs contained the correct determiner (non-violation trials), whereas the other half contained the incorrect determiner (violation trials). In a typical trial, participants first saw a black fixation cross for 1,000 ms on the white computer screen. Next, participants saw a single noun (e.g., *nube* [cloud]) in the centre of the screen. Participants were first instructed to indicate via button press if they were familiar with the noun to check for participants knowledge of the word. The noun was displayed until participant's response. Next, we showed participants another fixation cross for 500 ms. Then, participants were exposed to an NP (e.g., *la nube* [the cloud]). Participants had to indicate via button press whether the NP was correct. The NP remained on the screen until participants' response, or until the response limit of 3,000 ms was reached (Figure 8.2.1). In the latter case, we automatically coded the trial as incorrect. The experimenter did not provide feedback during the task. Each NP was only presented once during the task. There were 112 non-violation trials and 112 violation trials, adding to the total of 224 trials.

At regular intervals of 40 trials, we implemented self-paced breaks to restore participants' engagement with the task. Furthermore, we reminded participants via the presentation of an additional instruction screen after one third and two thirds of the trials to respond as accurately and fast as possible. Prior to initiating the main experiment, participants completed a practise round consisting of eight practise trials to get familiar with the procedure and to ask any clarification questions.

Figure 8.2.1: *Trial sequence for the syntactic violation paradigm, with an example for a non-violation trial on the left and an example for a violation trial on the right. Note that the prompts in this figure were translated to English for convenience.*



EEG recordings

We measured the EEG signal via 32 Ag/AgCI active channels arranged according to the international 10/20 montage by BioSemi, see Appendix 8.B. In addition, we used six external channels: two channels were attached to the outer canthus of the left and the right eye, respectively, to measure the horizontal electrooculogram (HEOG); two channels were attached above and below the left eye of participants to measure the vertical electrooculogram (VEOG); and finally, two channels were attached to the mastoid bone behind the left and right ear. All channels were referenced online at the

Common Mode Sense (CMS), while the Driven Right Leg (DRL) was used to capture ground circuit noise. We used the ActiView software (ActiView806-Lores) by BioSemi to configure the channels' impedances below 15 kΩ and to then generate the EEG recordings. The sampling frequency was 512 Hz, resulting in voltage amplitudes being measured approximately every 1.96 ms. The EEG signal was measured continuously during the syntactic violation paradigm.

8.3 Results

8.3.1 Behavioural data exclusion

We excluded the same participants in these analyses as in the EEG analysis, see section 8.3.4.

8.3.2 Behavioural data analysis

For our behavioural data, we followed a generalised linear mixed effects model (GLMM) approach to model our outcome variables accuracy and RTs using the *lme4* package (Bates et al., 2020) in R, Version 4.1.2, and in RStudio, Version 2021.09.0 (R Core Team, 2020). We specified a binomial distribution for accuracy, and a gamma distribution with the identity link function for RTs (Lo & Andrews, 2015). For RTs, we only modelled correct trials. For both accuracy and RTs, we considered *violation type* (violation vs. non-violation) in the fixed effects structure, as well as *LexTALE-Esp* scores, *noun gender* and *terminal phoneme* as covariates. For the random effects structure, we included random intercepts for *participant* and *item*, as well as by-*participant* random slopes for the effect of *violation type*.

For the model fitting procedure, we first constructed a theoretically plausible maximal model for each outcome variable with a maximal random effects structure as supported by our data (Barr, 2013; Matuschek et al., 2017). In the case of singular fit or non-convergence of our maximal model, we simplified our random effects structure. We used the default treatment contrast as our baseline for all models. The models for accuracy were fitted via the Laplace approximation and the models for RTs were fitted using the maximum likelihood (ML) method for reasons of model comparison. However, the final model for RTs was refitted using the restricted maximum likelihood (REML) method (Mardia et al., 1999). After model convergence, we checked the correlation structure between fixed effects using the *vcov()* function and *kappa.mer()* from the *JGmermod* package (Grafmiller, 2020) to identify potentially problematic collinearity in our fixed effects structure. We also performed model diagnostics to check our residual patterns using the *simulateResiduals()* function from the *DHARMa* package (Hartig, 2020).

After a positive evaluation of the model diagnostics, we checked for the statistical relevance of our covariates to avoid over-fitting our model. For this, we systematically compared models with and without a particular covariate term using the *anova()* function. Insignificant contribution of a covariate to the model fit was reflected in a non-significant χ^2-test and virtually identical Information Criterion values (AIC and BIC) for the two models (Akaike, 1974; Neath & Cavanaugh, 2012). Subsequently, the covariate was excluded from the model fitting procedure. However, if there was a significant difference in model fit as reflected in significantly smaller AIC and BIC values across models, the covariate was included in the model. For model parameters, we interpreted absolute test-statistic values larger than 1.96 as statistically significant (Alday et al., 2017). We obtained p-values using the *lmerTest* package (Kuznetsova et al., 2020). Model parameters for accuracy are reported as odds ratios, and all model parameters can be found in the Appendix.

8.3.3 Behavioural data results

Mean accuracy and RTs by violation type are displayed in Table 8.3.1.

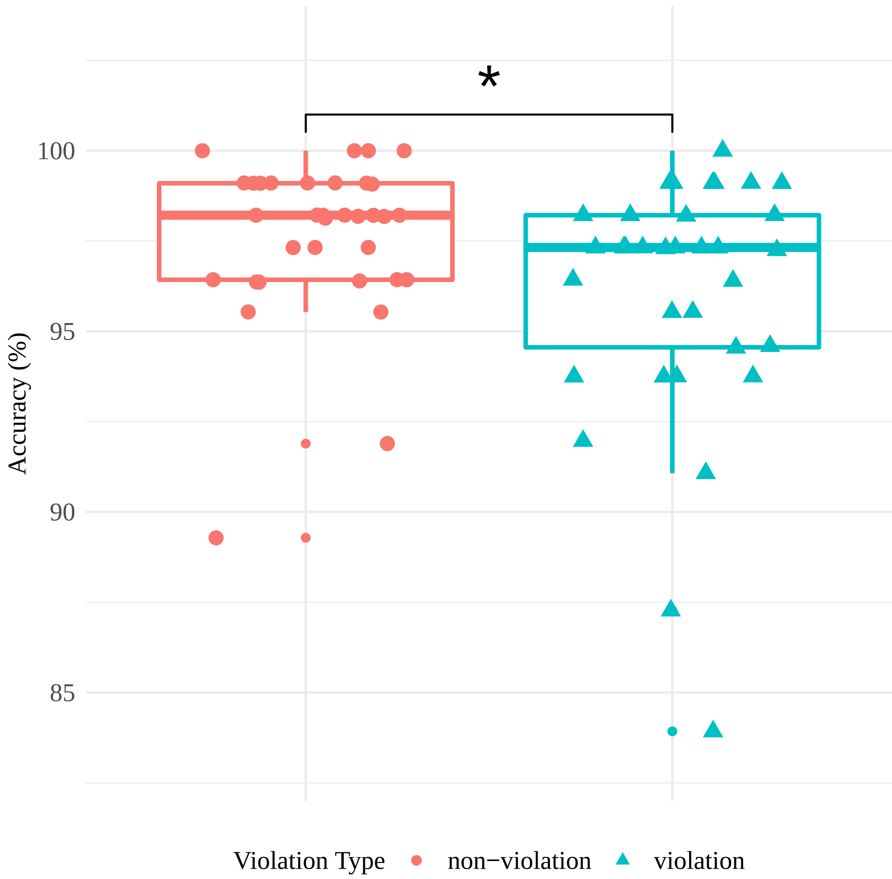
Table 8.3.1: *Mean accuracy and RTs by violation type (n = 34).*

Violation type	Accuracy (%)	SD	RTs (ms)	SD
non-violation	97.62	15.24	822.47	362.80
violation	96.05	19.48	858.16	340.79
Difference	1.57		35.69	

Accuracy

For accuracy, the model including by-*participant* random slopes for *violation type* yielded singular fit. We therefore dropped the random slopes from the model. Next, we excluded *terminal phoneme* from the fixed effects structure due to non-convergence of the model. The model of best fit included *violation type* as fixed effect, as well as *LexTALE-Esp* and *noun gender* as covariates. Further, we included random intercepts for *participant* and *item*. Therefore, the best-fitting models was the following: accuracy ∼ violation type (violation vs. non-violation) + LexTALE-Esp + noun gender (feminine vs. masculine) + (1|participant) + (1|item). Critically, participants were significantly more accurate for *non-violation* trials compared to *violation* trials with $\beta = 0.589$, *95% CI*[0.434, 0.800], $z = -3.39$, $p = 0.001$. See Figure 8.3.1 for a visualisation of these results, and Appendix 8.C for detailed model specifications.

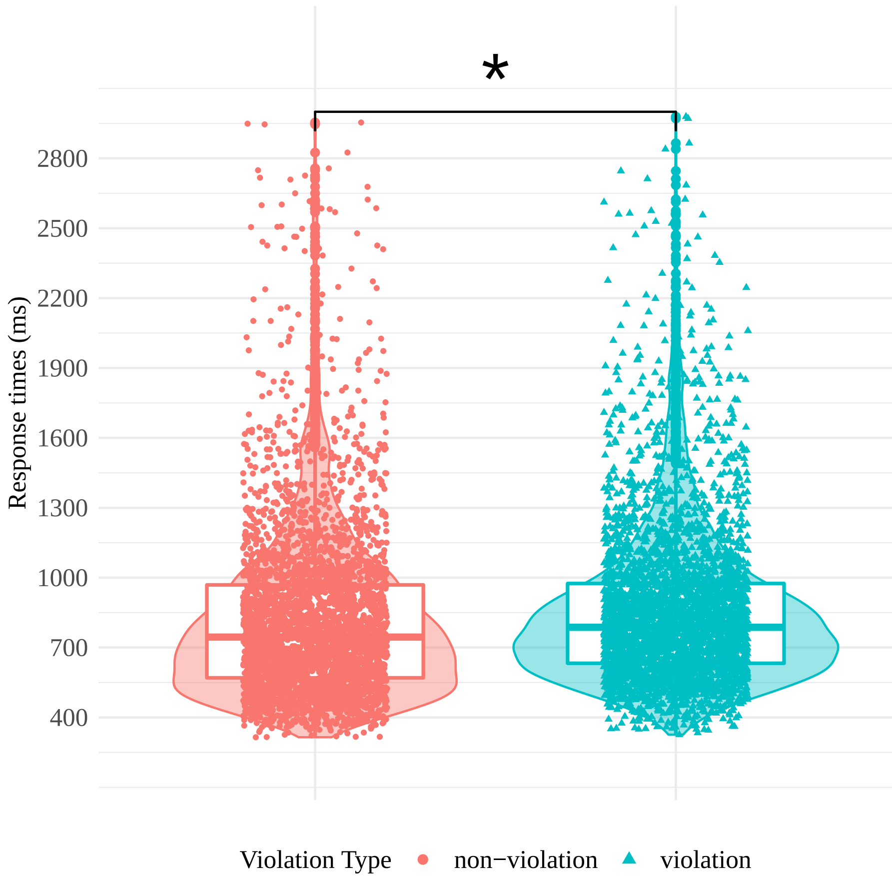
Figure 8.3.1: *Mean accuracy for each violation type (n = 34).*



Response times

For RTs, we dropped the covariate *terminal phoneme* from the maximal model due to non-convergence. Further, *noun gender* did not significantly improve the model fit. The best-fitting model contained *violation type* as fixed effect, as well as *LexTALE-Esp* as a covariate. In addition, we included a by-*participant* random slope for *violation type*, and random intercepts for *item*. The best-fitting model was: RTs ∼ violation type (violation vs. non-violation) + LexTALE-Esp + (violation type|participant) + (1|item). Here, participants were statistically faster for *non-violation* trials compared to *violation* trials, with $\beta = 33.80$, *95% CI* [26.25, 41.35], $t = 8.77$, $p < 0.001$. See Figure 8.3.2 for a visualisation of the results, and Appendix 8.D for detailed model specifications. Taken together, these behavioural results match the accuracy data and imply a significant effect of *violation type* on both accuracy and RTs, in line with our predictions.

Figure 8.3.2: *Mean response times for each violation type (n = 34).*



8.3.4 EEG data exclusion

One EEG dataset was excluded prior to data pre-processing due to a failure of the participant at following the task instructions; and another for a technical recording failure. Further, we established criteria for inclusion in the subsequent statistical analyses. First, we only included trials where participants had indicated that they were familiar with the stimulus noun. Second, we only modelled the EEG signal for correct trials, i.e., where participants had made a correct grammatical judgment on a non-violation or a violation trial. Third, trials which contained an artefact (muscle contractions, jaw movements, etc.) were excluded from any further analysis. Finally, heavily contaminated datasets (i.e., more than 30% trials lost to artefacts) were not included in further statistical analysis (see Appendix 8.E for by-violation type trial exclusion rates for familiar and correct trials). After application of these criteria, four datasets were excluded, adding to a total of 34 included participants. We excluded the same participants in the behavioural analyses as in the EEG analyses.

8.3.5 EEG data pre-processing

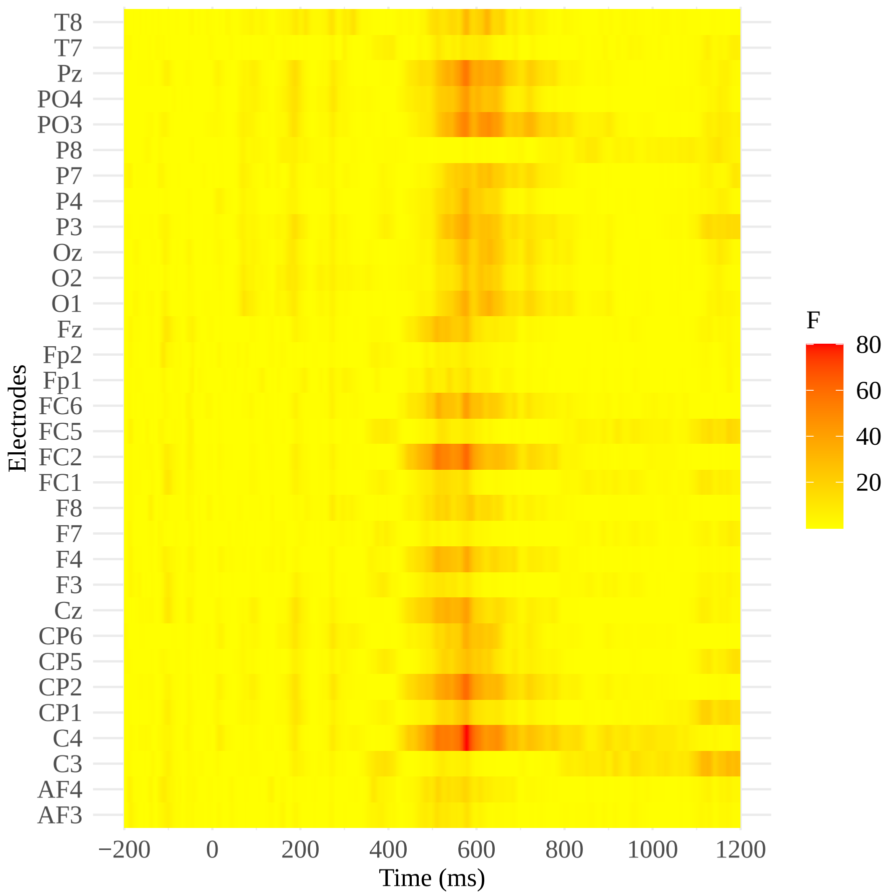
Prior to the statistical analyses of our EEG data, we first performed EEG data pre-processing to increase the signal-to-noise ratio using BrainVision Analyzer (Brain Products, GmbH, Munich). First, we re-referenced the signal from the implicit reference channel to the average of the two mastoid channels. Next, we separately performed linear derivation for the two VEOG and HEOG channels to derive a single VEOG and HEOG channel. We applied offline high-pass filters of 0.1 Hz and at low-pass filters of 30 Hz before we performed residual drift correction for the newly generated VEOG and HEOG channels. In this, we defined a maximum amplitude of ± 200 μV for the HEOG channel, and ± 800 μV for the VEOG channel. Then, we corrected for blink activity using ocular independent component analysis (ICA). Next, we performed artefact rejection to mark bad intervals using the following criteria: for the gradient, we allowed a maximal voltage step of 50 μV/ms, a maximal difference in 100 ms - intervals of 200 μV; maximal amplitudes of ± 200 μV, and the lowest allowable amplitude in 100 ms - intervals of 0.5 μV. As a final step, we generated epochs around stimuli markers from familiar and correct trials from -200 ms pre-event to 1,200 ms post-event. We applied baseline correction to each segment using the activity 200 ms pre-event as baseline. We then exported the voltage amplitudes for each uncontaminated segment for each channel and each participant. We exported a total of 32 channels: *Fp1, Fp2, AF3, AF4, Fz, F3, F4, F7, F8, FC1, FC2, FC5, FC6, Cz, C3, C4, CP1, CP2, CP5, CP6, T7, T8, Pz, P3, P4, P7, P8, PO3, PO4, Oz, O1* and *O2*. Post-export, we used the *BVAtoR* package (Bonneville, 2020) to combine the information from the three default export files (.dat file, .vmrk file and .vhdr file) into a customised data frame containing voltage amplitudes for each time point, channel, violation type and participant as well as the covariates. Finally, each channel was assigned to one of nine topographic regions: left anterior, mid anterior, right anterior; left central, mid central, right central; and finally, left posterior, mid posterior and right posterior regions.

8.3.6 EEG data analysis

A typical approach to analyse ERP data is to generate a priori hypotheses about the region of interest (ROI), i.e., where to expect task-relevant effects; and the time window of interest, i.e., when to expect task-related effects (Friederici et al., 1999; Wicha et al., 2004). In this study, we took a more data-driven approach which allowed us to flexibly model our EEG data without making assumptions about these two parameters. First, we performed a cluster-based permutation analysis to identify a potential ROI. Second, we used generalised additive mixed models (GAMMs) to model *voltage amplitudes* over time and to determine a time window of interest for our effects. Upon termination of these analyses, we further explored whether or not this approach would yield comparable ROIs and time windows of interest compared to the previous literature (Friederici et al., 1999; Kutas & Federmeier, 2011; Martín-Loeches et al., 2005; Osterhout & Mobley, 1995; Wicha et al., 2004).

To determine our ROI, we conducted a permutation analysis using the *permu.test()* function from the *permutes* package (Voeten, 2019) in R. In this, we calculated F-values to reflect differences in voltage amplitudes as a function of *violation type* for each channel (Maris & Oostenveld, 2007; Voeten, 2019). See Figure 8.3.3 for a visualisation of the permutation analysis outcome. In line with the previous literature, visual inspection of the outcome suggested centro-parietal channels as potential ROI around 600 ms, as we would expect for an N400 or a P600 effect. Based on this outcome, we selected the channels *CP1, CP2, CP5, CP6, Pz, P3, P4, PO3* and *PO4* in bilateral centro-parietal regions to model potential N400 and P600 effects. Notably, the outcome did not suggest any significant voltage amplitude modulation prior to this time window, as we would have expected for a LAN effect. Nevertheless, we also probed the presence of the LAN in a separate statistical analysis.

Figure 8.3.3: *Permutation analysis outcome for n = 34. Larger F-values are shown in darker colours and denote an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes.*



Next, to determine our time window of interest, we used a GAMM approach. GAMMs have only recently been applied in ERP research on language processing mechanisms (De Cat et al., 2015; Meulman et al., 2015; Tremblay & Newman, 2015). They represent an extension of (generalised) linear mixed models (LMMs) and have the primary advantage of allowing researchers to flexibly model the complex oscillatory trend of *voltage amplitudes* over *time*. This is done via the inclusion of non-linear terms, so-called (penalised) splines or *smooths*, alongside the linear terms. Non-linear terms are described in terms of a set number of so-called basis functions. These are automatically determined depending on the complexity of the effect of the non-linear term. Moreover, GAMMs are particularly powerful because of the maximum likelihood estimation to calculate unbiased model parameters for the full dataset, even in the case of missing data. Importantly, this is based on the assumption that data are missing at random (MAR) or completely at random (MCAR). GAMMs also allow for the inclusion of by-participant and by-item effects in the form of random slopes and random intercepts. Note that this is also featured in LMMs, see Frömer et al. (2018). Given our data, we determined that GAMMs were a viable approach to model ERP effects of *violation type*. Moreover, we were particularly curious about the predictive power of our models to determine potential time windows of interest. Therefore, we first performed a GAMM analysis using the ROI previously determined in our permutation analysis to explore potential P600 effects. Then, we conducted a similar analysis for the LAN based on a pre-defined ROI based on the literature, given that the permutation analysis did not yield any concrete ROI for LAN effects.

Both EEG analyses for the P600 and the LAN were modelled after previous work by Meulman et al. (2015) and De Cat et al. (2015). We followed a GAMM approach in R using the *mgcv* package (Wood, 2021). Here, we used the *bam()* function for large datasets. Our model fitting procedure was as follows: We constructed a theoretically plausible model which included *voltage amplitudes* as outcome variable, *time* as non-linear term, *violation type* as ordered linear term, *channel* as covariate, the interaction effect between *time* and *violation type*, and between *time* and *channel*, as well as random intercepts for *participant* and *item*, random slopes for the effect of *violation type*, *channel* and *time* for each *participant*, and a random slope for *time* for each *item*. To avoid over-fitting the model, we initially opted to fit a more simplistic model for which we omitted any additional covariates from the fitting procedure. The model was fitted using fast restricted maximum likelihood (fREML) estimation. After fitting the model, we used the function *gam.check()* to inspect the model fit and the model residuals, and to identify any potential issues related to the number of basis functions, concurvity (similar to collinearity for linear models) or influential data points. The model was first fitted using a Gaussian distribution, but inspection of the residual distribution revealed heavy tails in the residual histogram and quantile-quantile plots. We therefore re-fitted the model using a scaled t-distribution with the *identity* link function on the response scale (Meulman et al., 2015). We also checked for spatial and temporal autocorrelation in the model's residuals and applied a correction via the ρ parameter for AR1 error, if applicable (Baayen, Vasishth, Kliegl & Bates, 2017; De Cat et al., 2015; Meulman et al., 2015). As a final step, we plotted the model's predicted differences using the *plot_diff()* function from the *itsadug* package (Van Rij et al., 2020). This function includes simulation-based calculations on the statistical difference in voltage amplitudes between the *non-violation* and *violation* trials.

8.3.7 EEG data results

P600 results

We visualised mean voltage amplitudes for *non-violation* and *violation* trials in centro-parietal channels indicated in the permutation test (CP1, CP2, CP5, CP6, Pz, P3, P4, PO3, PO4) in Figure 8.3.4. Visual inspection of the first 200 ms post-stimulus onset showed the characteristic early visual processing response in the form of the P1/N2 complex (Bakos et al., 2020; Gamboa Arana et al., 2020). This is followed by a large positive-going oscillation with diverging voltage amplitudes for *non-violation* and *violation* trials starting around 450 ms post-stimulus onset and peaking around 650 ms. This signal is consistent with the characteristics of a P600 component and potentially a P600 effect, but not an N400 component in this ROI. Mean voltage amplitudes subsequently converged around 800 ms post-stimulus onset following a downward oscillatory trend. The full complexity of the data is reflected in by-participant means for violation type in Figure 8.3.5. As evident from both figures, voltage amplitudes changed dynamically over time, which cannot easily be modelled with a conventional LMM

(Meulman et al., 2015). In contrast, GAMMs can accurately capture this dynamic change, while avoiding the generation of a priori assumptions about the time window of interest.

Figure 8.3.4: *Mean voltage amplitudes by violation type in centro-parietal regions (CP1, CP2, CP5, CP6, Pz, P3, P4, PO3, PO4) for n = 34.*

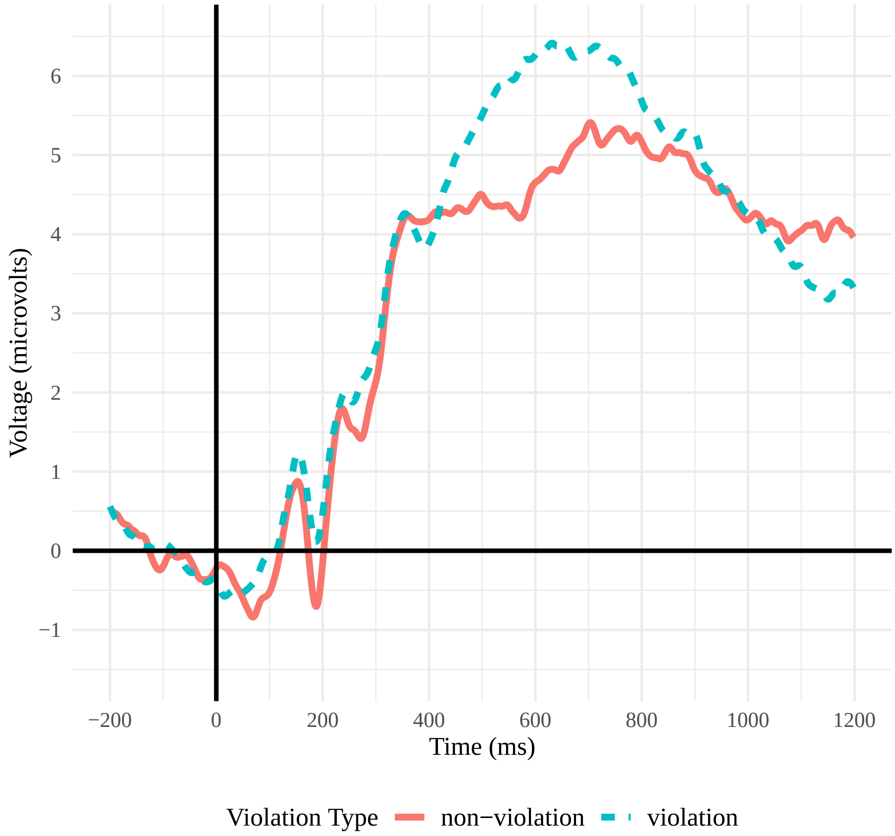
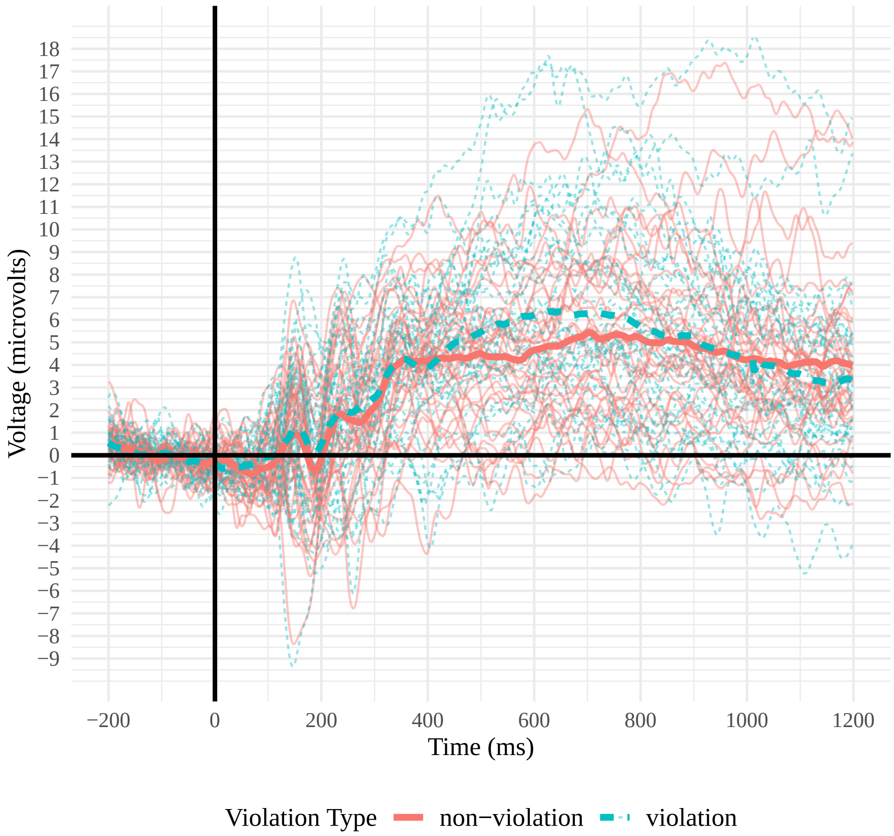
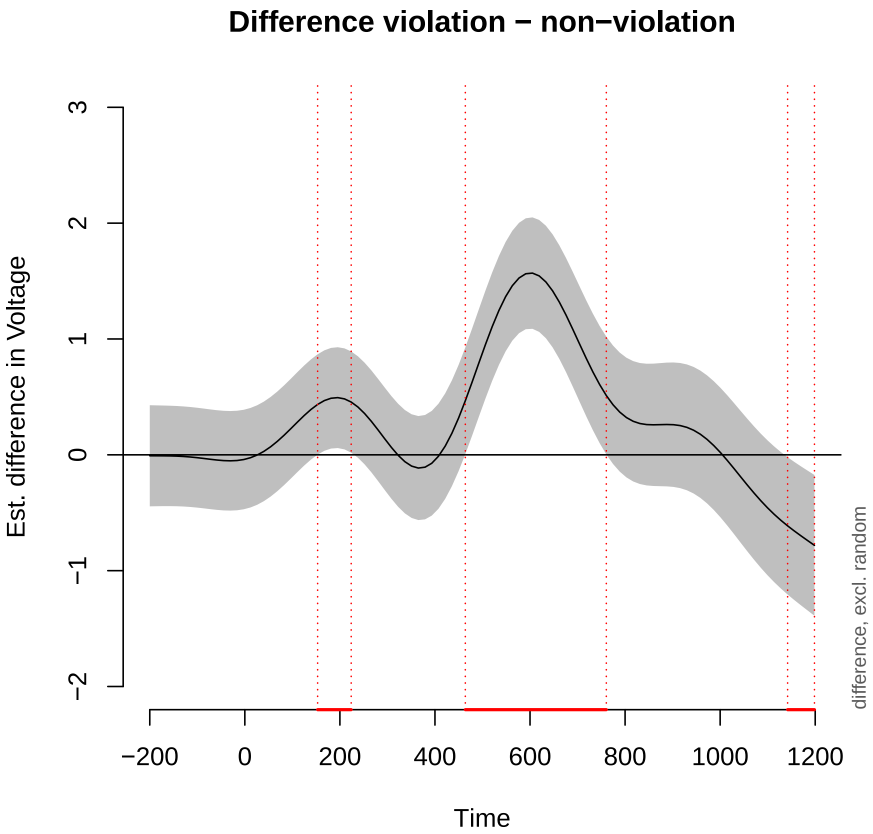


Figure 8.3.5: *Mean voltage amplitudes by violation type in centro-parietal regions (CP1, CP2, CP5, CP6, Pz, P3, P4, PO3, PO4) for n = 34. The thicker lines represent average voltage amplitudes by violation type, whereas the finer lines represent by-violation type averages for each participant.*



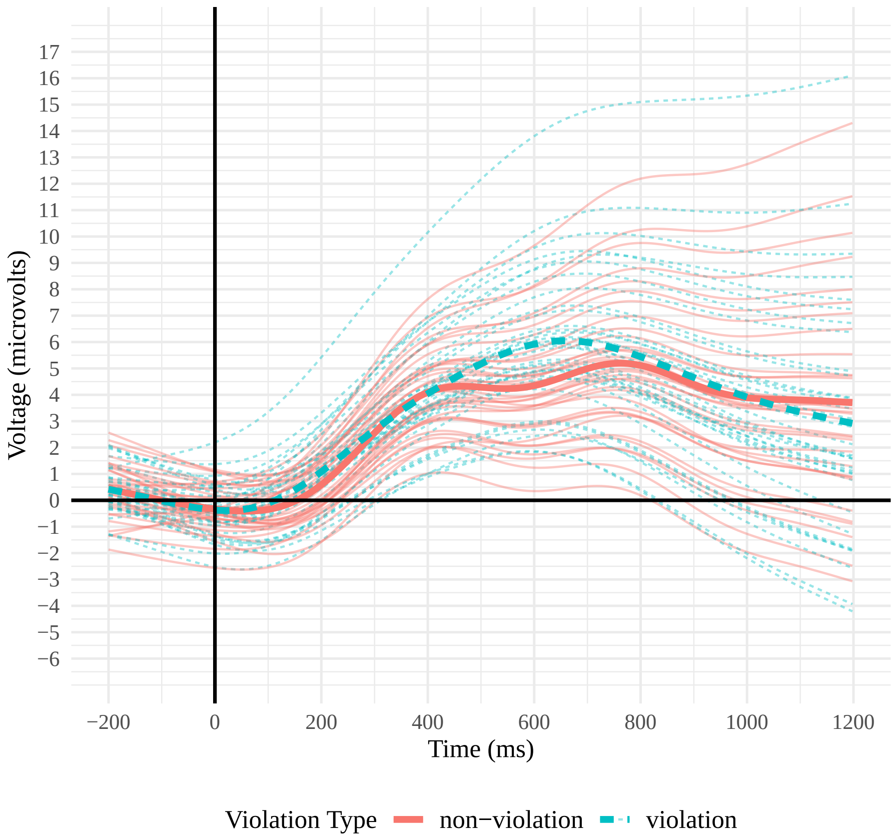
Model parameters for linear and non-linear (smooth) terms are reported in Appendix 8.F. Critically, the model yielded a significant interaction effect of *violation type* and *time* with $F = 3489.47$, $p < 0.001$ for non-violation trials compared to violation trials. This indicated a statistical difference in voltage amplitude between non-violation and violation trials over time. We visualised this predicted difference in voltage amplitudes between non-violation trials and violation trials in Figure 8.3.6. Voltage amplitudes were significantly higher for violation trials compared to non-violation trials in the time window between 464 ms and 761 ms, thereby reflecting a P600 effect. This time window of the effect is highlighted by the dashed vertical lines and the bold line along the x-axis in Figure 8.3.6. Importantly, this time window of our P600 effect by the model is consistent with previous reports on the temporal locus of the P600 effect (Friederici et al., 1999; Osterhout & Mobley, 1995).

Figure 8.3.6: *Predicted differences in voltage amplitudes (μV) for violation and non-violation trials for channels CP1, CP2, CP5, CP6, Pz, P3, P4, PO3 and PO4 (n = 34.). Note that random effects are excluded here.*



Channel emerged as a significant covariate. We also included random slopes and random intercepts for by-*participant* and by-*item* effects. The fitted model was the following: voltage amplitudes ∼ violation type + channel + s(time) + s(time, by = violation type) + s(time, by = channel) + s(participant, channel, bs = “re”) + s(participant, violation type, bs = “re”) + s(participant, time, bs = “re”) + s(participant, bs = “re”) + s(item, time, bs = “re”) + s(item, bs = “re”). Further, the model captured a variance of 7.59%, which is relatively low but not unusual given the large individual variability in the EEG signal (Meulman et al., 2015). Finally, we plotted the average voltage amplitudes for each participant for *violation type* as predicted by our model in Figure 8.3.7. The thicker lines represent the predicted means for each *violation type*, and the remaining lines represent the predicted means for each participant. See Appendix 8.F for the full model parameters.

Figure 8.3.7: *Predicted mean voltage amplitudes (μV) for violation and non-violation trials for each participant for channels CP1, CP2, CP5, CP6, Pz, P3, P4, PO3 and PO4 (n = 34). Note that random effects are excluded here.*



LAN results

The permutation analysis outcome in Figure 8.3.3 did not suggest a difference in voltage amplitudes for *violation type* in a time window prior to the P600 effect, as would be expected for a LAN effect. Nevertheless, we examined the effect of *violation type* on voltage amplitudes in left anterior regions. Based on previous literature, we selected the channels *Fp1, AF3, F3* and *F7* as our ROI

(Martín-Loeches et al., 2005; Molinaro et al., 2015). We then performed an identical analysis as described above. See Figure 8.3.8 for a visualisation of mean voltage amplitudes over time and Figure 8.3.9 for mean voltage amplitudes for each participant for violation type. The visualisation of voltage amplitudes revealed a positive-going peak around 200 ms post-stimulus onset, followed by a decrease in voltage amplitudes back to baseline around 300 ms. Voltage amplitudes then follow a positive trend, briefly diverging around 450 ms and converging around 700 ms post-stimulus onset.

Figure 8.3.8: *Mean voltage amplitudes by violation type for left anterior regions (Fp1, AF3, F3, F7) for n = 34.*

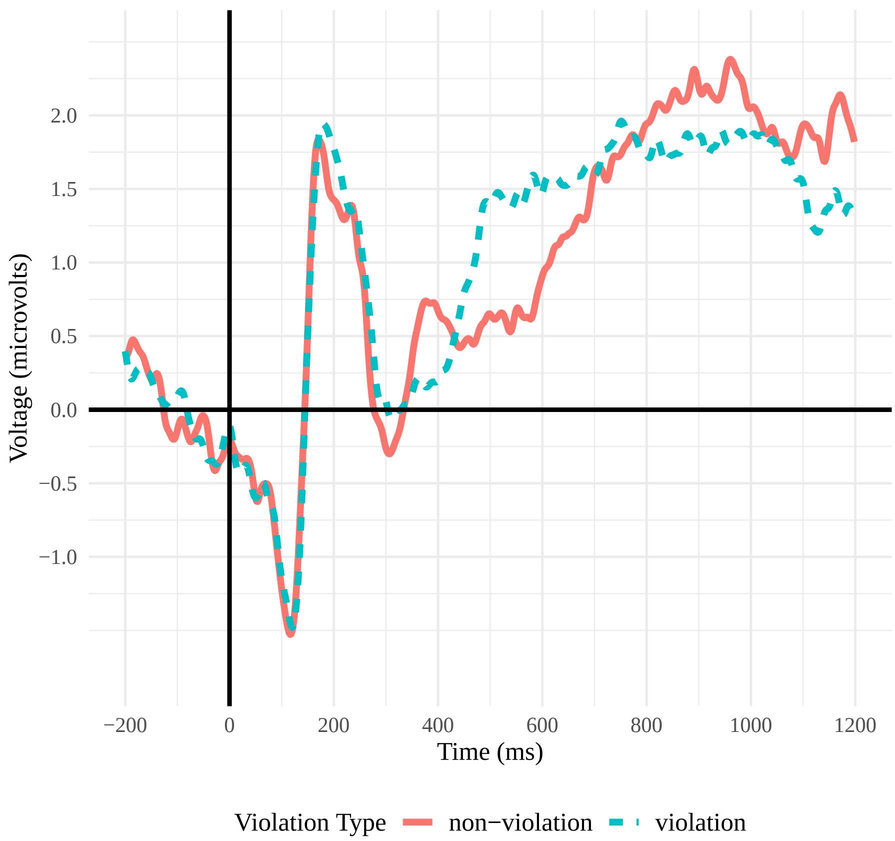
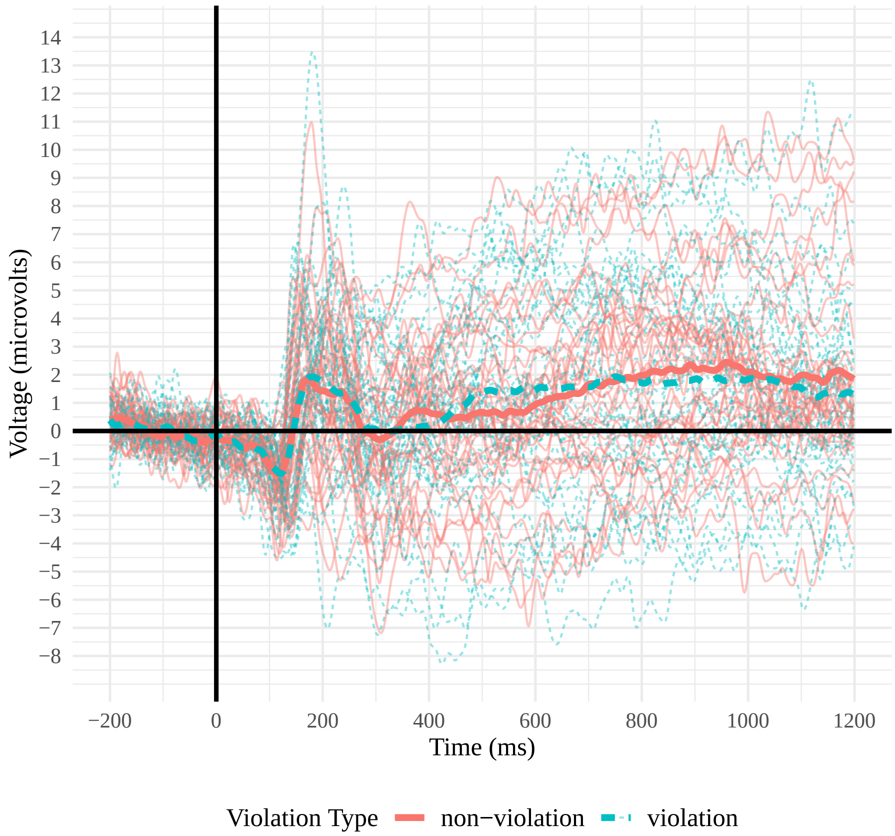
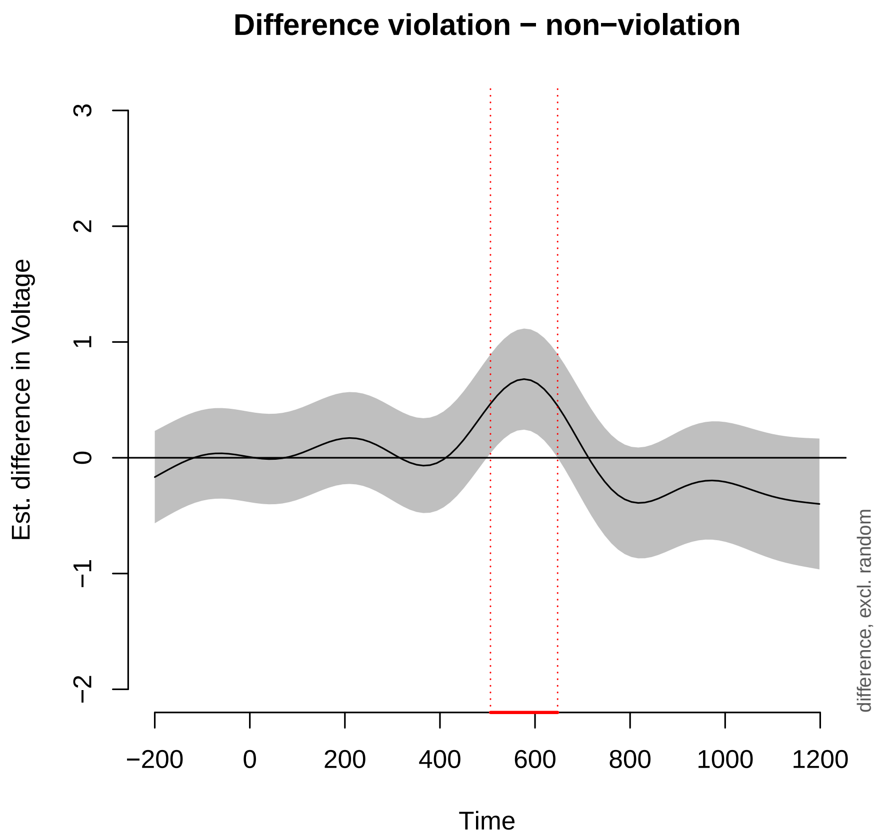


Figure 8.3.9: *Mean voltage amplitudes by violation type for left anterior regions (Fp1, AF3, F3, F7) for n = 34. The solid line represents average voltage amplitudes by violation type, whereas the finer lines represent by-violation type averages for each participant.*



The interaction effect between the smooth term *time* and *violation type* was significant with $F = 413.48$, $p < 0.001$. This implies a statistical difference in voltage amplitudes between violation trials and non-violation trials over time. We visualised this in Figure 8.3.10, which shows more positive voltage amplitudes for violation trials compared to non-violation trials between 506 ms and 648 ms post-stimulus onset. However, previous studies reported more negative amplitudes to violation trials compared to non-violation trials around 300 ms to 500 ms post-stimulus onset (Kaan, 2007; Molinaro et al., 2015). We therefore argue that these results are not compatible with a traditional LAN effect, as we will expand on in detail in the discussion section.

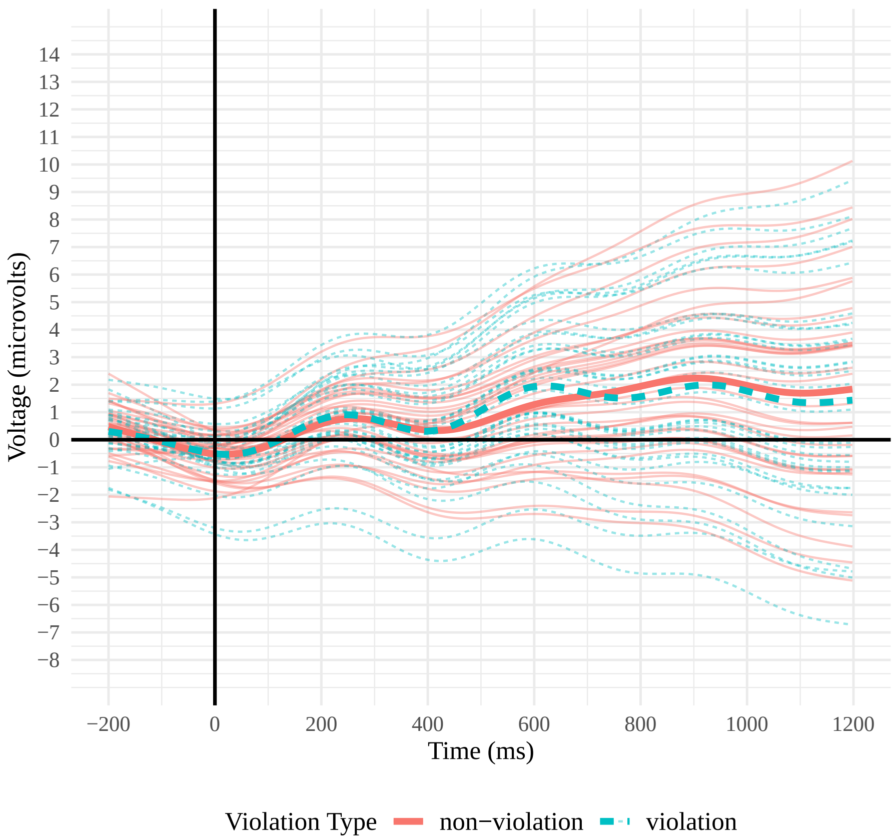
Figure 8.3.10: *Predicted differences in voltage amplitudes (μV) for violation and non-violation trials for channels Fp1, AF3, F3 and F7 (n = 34.). Note that random effects are excluded here.*



Similar to the P600 analysis, *channel* had an effect on voltage amplitudes and was included as a covariate, as were the random intercepts and random slopes for *participant* and *item* effects. The fitted model was as follows: voltage amplitudes ∼ violation type

+ channel + s(time) + s(time, by = violation type) + s(time, by = channel) + s(participant, bs = "re") + s(participant, violation type, bs = "re") + s(participant, channel, bs = "re") + s(participant, time, bs = "re") + s(item, bs = "re") + s(item, time, bs = "re"). This model captured 4.22% of the variance. Predicted by-participant means are shown in Figure 8.3.11, with the thicker lines representing mean voltage amplitudes by *violation type*. See Appendix 8.G for the full model parameters.

Figure 8.3.11: *Predicted mean voltage amplitudes (μV) for violation and non-violation trials for each participant for channels Fp1, AF3, F3 and F7 (n = 34). Note that random effects are excluded here.*



8.4 Discussion

The primary contribution of this study was the exploration of the neural correlates linked to processing gender agreement violations in Spanish. There has been a controversy with respect to the elicited ERP effects: on the one hand, several studies involving gender agreement violations in sentence-embedded NPs embedded found P600 effects (Barber & Carreiras, 2005; Gunter et al., 2000; Kaan, Harris, Gibson & Holcomb, 2000; Molinaro et al., 2008; Wicha et al., 2004). On the other hand, studies on Spanish reported N400 effects for gender agreement violations in NPs outside of a sentence context (Barber & Carreiras, 2005, 2003). These specific findings challenged the classical interpretation of the P600 effect as being linked to gender agreement violations, at least for Spanish. Instead, they put forward the N400 effect as an alternative candidate, which warranted a thorough investigation. Therefore, our study aimed to examine first, the ERP components connected to processing gender agreement violations in NPs such as [*el nube] compared to grammatical NPs such as [la nube]; and second, whether the P600 component indeed played a significant role in gender violation processing in isolated NPs in Spanish. For this, we recorded participants' EEG while they judged the grammaticality of determiner-noun NPs, e.g., [la nube] in Spanish. We not only probed P600 effects and N400 effects, but also a potential LAN effect, which is frequently reported in combination with the P600 effect for these types of syntactic violations (Barber & Carreiras, 2005; Molinaro et al., 2015; Steinhauer & Drury, 2012). Moreover, to explore non-linear effects within the EEG data, we examined non-linear oscillations in voltage amplitudes over the entire segment length using generalised additive mixed models, GAMMs to determine a time window of interest for our effects (Meulman et al., 2015). This is a suitable approach to analysing complex EEG data as it does not require a priori hypothesising of a time window of interest for a given ERP component (Meulman et al., 2015).

For the behavioural measures of *accuracy* and *RTs*, we predicted higher accuracy and shorter RTs for non-violation trials compared to violation trials. This would reflect differential processing of syntactically correct vs. incorrect NPs. In line with our predictions, results showed that participants were statistically both more accurate and faster at making a grammaticality judgement for non-violation trials compared to violation trials (Friederici et al., 1999; Neville et al., 1991). Therefore, these results indeed suggest an overall processing advantage for syntactically correct vs. incorrect structures, even when the speakers are tested in a non-native environment.

For ERPs, we first investigated a potential P600 effect. This would be reflected in more positive voltage amplitudes for violation compared to non-violation trials (Barber & Carreiras, 2005; Hagoort et al., 1993). Moreover, we examined a potential LAN effect and N400 effect, which are typically reflected in more negative voltage amplitudes for violation trials compared to non-violation trials. We conducted two analyses on different ROIs: one analysis for centro-parietal regions to examine the P600 effect and the N400 effect, and one analysis for left anterior regions to investigate a LAN effect. The ROI for the first analysis was derived from the permutation analysis, whereas the ROI for the second analysis was based on previous literature. For the first analysis, the critical findings were the following: first, we found more positive voltage amplitudes for violation trials than for non-violation trials between 464 ms and 761 ms post-stimulus onset for the centro-parietal channels CP1, CP2, CP5, CP6, Pz, P3, P4, PO3 and PO4. This is an index for the classical P600 effect for gender agreement violations at the determiner-noun level in Spanish. This finding adds to the evidence of the crucial involvement of the P600 component in syntactic integration and repair processes not only for sentence-embedded NP violations, but also for isolated NPs in Spanish (Hagoort et al., 1993; Osterhout & Nicol, 1999). Second, we did not find evidence for an N400 effect, neither in the permutation analysis nor in the statistical analysis: our data showed that voltage amplitudes were comparable for both violation and non-violation trials over time until the onset of the P600 effect at 464 ms, thereby not yielding an N400 effect. Therefore, this finding contrasts with previous studies on gender agreement violations in isolated NPs in Spanish (Barber & Carreiras, 2003, 2005).

As discussed in the introduction, previous literature highlighted the crucial involvement of the P600 component in processing gender agreement violations in NPs within sentences across several languages. However, the N400 component was also put forward as an alternative critical ERP component for gender agreement processing, see for example Barber and Carreiras (2005) and Barber and Carreiras (2003). These previous findings and their implications were highly relevant to the current study because they suggested that there may be a measurable difference in processing gender agreement violations in isolated NPs vs. NPs embedded within sentences. By extension, the ERP correlates may differ depending on whether there the violation is embedded within a sentence context. Returning to the overarching question we asked in this study about whether gender agreement violations within isolated NPs elicited different ERP components compared to violations within NPs embedded in a sentence reported in the literature, our results do not support this notion. More specifically, our results demonstrated that P600 effects are linked to gender agreement violations in isolated NPs. Therefore, the findings from the current study suggest similar neural signatures for processing gender agreement violations in isolated NPs as previously reported for sentence-embedded NPs (Barber & Carreiras, 2005; Gunter et al., 2000; Molinaro et al., 2008).

With respect to the second analysis performed to probe a potential LAN effect, we found that voltage amplitudes were significantly different for violation compared to non-violation trials between 506 ms to 648 ms post-stimulus onset for channels Fp1, AF3, F3 and F7. More specifically, we found more positive amplitudes for violation trials compared to non-violation trials. This is in contrast with previous research suggesting a time window of interest for the LAN effect between 300 ms and 500 ms, as well as more negative amplitudes for violation trials (Kaan, 2007; Molinaro et al., 2015). Critically, the time window of this anterior effects overlaps with the P600 effect, which had an onset of 460 ms post-stimulus onset. Based on the topographic characteristics of the LAN as reported in the literature (Barber & Carreiras, 2005; Hagoort, 2003), we argue that our findings are inconsistent with a LAN effect. Instead, we suggest that this effect reflects the early stage (500 ms to 700 ms) of the P600 component. This stage was previously associated with indexing sensitivity to syntactic integration and to a broader topographic distribution (Alemán Bañón et al., 2012; Barber & Carreiras, 2005; Hagoort & Brown, 2000; Molinaro et al., 2008). Therefore, we argue that both the left anterior effect between 506 ms - 648 ms and the centro-parietal effect between 464 ms - 761 ms belong to the early stage of a broadly distributed P600 component. Importantly, by extension, our findings do not show evidence for a biphasic LAN/P600 pattern but instead show an isolated and broadly distributed P600 effect for gender agreement violations (Gunter et al., 2000; Molinaro et al., 2008). Interestingly, we did not find any differences in voltage amplitudes in the later stage of the P600 component between 700 ms - 900 ms, which was previously linked to syntactic analysis and repair and a posterior topographic distribution (Alemán Bañón et al., 2012). Therefore, the duration of the P600 effect was shorter compared to previous studies (Barber & Carreiras, 2005; Friederici et al., 1999) and localised to the syntactic integration stage of the P600 component (Alemán Bañón et al., 2012).

Taken together, our results suggested the following: first, they revealed a P600 effect for gender agreement violations in isolated determiner-noun NPs. More specifically, there was a difference in voltage amplitudes between grammatical and ungrammatical NPs in the early stage of the P600 component, previously associated with syntactic integration. Second, and in contrast to previous research (Barber & Carreiras, 2003, 2005), we did not find evidence that an N400 component was linked to processing NP violations in Spanish. Finally, our data did not support the notion of a biphasic LAN/P600 pattern. Therefore, with respect to our research question, our results indicated that the P600 component was the primary neural correlate relevant to processing gender agreement violations in isolated NPs in Spanish.

A novel aspect of this study was the application of a permutation analysis in combination with generalised additive mixed models (GAMMs) to analyse our ERP data (Meulman et al., 2015; Tremblay & Newman, 2015). We first conducted a permutation analysis to determine our ROI, followed by a GAMM analysis to detect our time window of interest to explore our ERP effects. In this, we avoided generating a priori assumptions about both of these parameters. Results from the permutation analysis provided us with a ROI involving centro-parietal regions to explore P600 effects. This is in line with previous literature on P600 effects (Friederici et al., 1999; Hagoort et al., 1993). Next, the results from the GAMM analysis provided us with a precise time window of interest for the P600 effect for *violation type* on voltage amplitudes. The time window of interest of 464 ms to 761 ms for the P600 effect as predicted by the model was consistent with previous work on the P600 effect, as was the more anterior portion of the P600 effect between 506 ms and 648 ms (Barber & Carreiras, 2005; Friederici et al., 1999; Hagoort & Brown, 2000; Molinaro et al., 2015; Swaab et al., 2011). These results confirmed that GAMMs are a suitable and powerful analysis approach while representing an extension to linear mixed models (Frömer et al., 2018): not only did GAMMs accurately capture the complex non-linear trends of voltage amplitudes over time, but their relatively straightforward implementation in R using the *mgcv* package (Wood, 2021) allows researchers to remain unbiased with respect to the timing of the ERP effects (Meulman et al., 2015). Naturally, caution is warranted with respect to over-fitting the data and the evaluation of the model diagnostics to assess the goodness of fit. Nevertheless, we argue that GAMMs were a paramount component of this study and should be considered for similar ERP research on language comprehension in the future.

8.4.1 Conclusions and future directions

This study contributed new evidence to the discussion on which ERP components were linked to processing gender agreement violations in isolated NPs in Spanish. Our participants performed a syntactic violation paradigm including determiner-noun NPs while their EEG was recorded. Our results highlighted a broadly distributed P600 effect in the early syntactic integration stage for processing gender agreement violations compared to non-violations in Spanish. In contrast, we found neither evidence for a LAN effect, nor an N400 effect. Our findings are relevant for two reasons: first, the confirm the importance of the P600 effect in processing gender agreement violations, even for violations occurring in isolated NPs. Second, they challenge previous work reporting an N400 effect for gender agreement violations in NPs (Barber & Carreiras, 2005, 2003). For a more nuanced investigation of gender agreement violations in isolated NPs vs. NPs in sentences, future studies should extend the current design and directly compare the ERPs elicited in both linguistic configurations. Moreover, it remains unclear whether the consistency of the ERP signature across NPs is specific to the target language Spanish, or whether there are indeed languages with functionally and neurally different ERP signatures as a function of the sentence context (or lack thereof) of the gender agreement violation.

CRediT author contribution statement

Sarah Von Grebmer Zu Wolfsthurn: Conceptualisation, Methodology, Validation, Investigation, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review and Editing, Visualisation. **Leticia Pablos-Robles**: Conceptualisation, Methodology, Writing-Review and Editing, Supervision. **Niels O. Schiller**: Conceptualisation, Writing-Review and Editing, Supervision, Funding Acquisition.

Declaration of competing interests

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

Acknowledgements

We thank our research assistants Philipp Flieger and Ifeoluwa Olusayo Oloruntuyi for their valuable help in the data collection. We are also grateful to Franz Wurm and Marjolein Fokkema for their feedback on the statistical analyses, and to our anonymous reviewers from a previous publication for highlighting the potential relevance of the results of this particular study. The statistical analyses were performed using the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. Finally, we want to thank all of our participants for their participation in this study.

Funding statement

This project has received funding from the European Union's Horizon2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N^{o} 765556 - The Multilingual Mind.

Data availability statement

The data that support the findings of this study are openly available in the Open Science Framework at: `https://osf.io/4m9ky/?view_only=55cac00c9acf429b81aec5fe11d8d266`

Citation diversity statement

Recent papers have made an effort to highlight the underrepresentation of first authors which identify as female and/or as members of a minority group in the academic environment (Dworkin et al., 2020; Zurn et al., 2020). This gap has potentially widened since the start of the global COVID-19 pandemic (Viglione, 2020). Therefore, we report the ratio of represented gender in our reference list on the basis of a binary female-male classification of the first and last authors. Our references consisted of 20% woman/ woman authors, 46% man/ man, 16% woman/ man and finally, 9% man/ woman authors. For comparison, these ratios for the field neuroscience are 6.7% for woman/ woman, 58.4% for man/ man, 25.5% woman/ man, and lastly, 9.4% for man/ woman authors (Dworkin et al., 2020). An obvious limitation of this current statement is the rather coarse nature of this female/male classification. However, as information about preferred gender becomes more readily available, we are optimistic that such statements will become more nuanced and representative.

Appendix

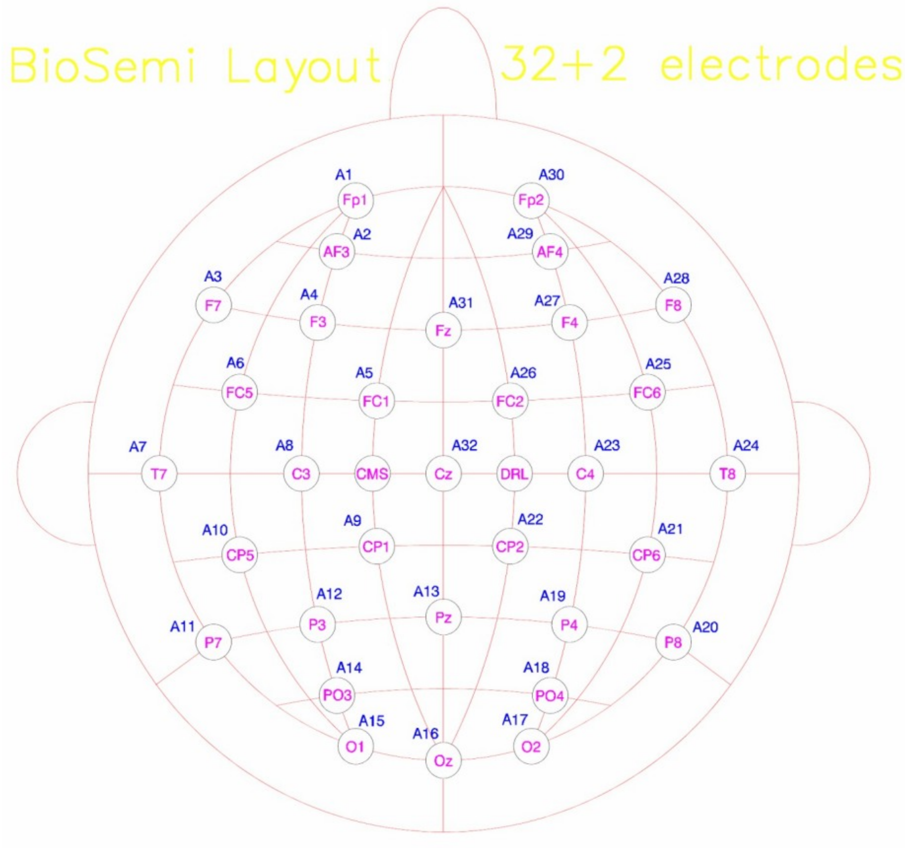
8.A Linguistic profile of participants

Table 8.A.1: *Overview of the native and non-native languages acquired by the participants included in the analysis (n = 34).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Spanish	n = 34					**34**
English		n = 32	n = 2			**34**
German		n = 2	n = 2	n = 2		**6**
Dutch			n = 6	n = 7	n = 3	**16**
French			n = 7	n = 1	n = 1	**9**
Italian			n = 5	n = 3		**8**
Catalan			n = 1			**1**
Japanese			n = 1	n = 1		**2**
Luxembourgish				n = 1		**1**
Total	**34**	**34**	**24**	**15**	**4**	

8.B EEG montage

Figure 8.B.1: *10/20 32-channel montage from BioSemi including CMS and DRL but excluding external channels (www.biosemi.com/headcap.htm).*



8.C Model parameters: accuracy

Table 8.C.1: *Specification of model of best fit for accuracy for n = 34.*

Formula: accuracy ∼ violation type (violation vs. non-violation) + LexTALE-Esp + noun gender (feminine vs. masculine) + (1|participant) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	8.12 [3.52, 18.74]	4.91	< 0.001
Violation type [violation]	0.589 [0.434, 0.800]	-3.39	**0.001**
LexTALE-Esp	1.02 [1.01, 1.03]	4.25	< 0.001
Gender [masculine]	1.51 [1.11, 2.04]	2.65	0.008
Random effects			
σ^2	3.29		
$\tau_{00 Item}$	0.27		
$\tau_{00 Participant}$	0.20		
ICC	0.12		
$N_{Participant}$	34		
N_{Item}	224		
Observations	7,580		
Marginal R^2/ Conditional R^2	0.071 / 0.187		

8.D Model parameters: response times

Table 8.D.1: *Specification of model of best fit for RTs (ms) for n = 34.*

Formula: RTs ∼ violation type (violation vs. non-violation) + LexTALE-Esp + (violation type|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	713.39 [706.22, 720.55]	195.13	< 0.001
Violation type [violation]	33.80 [26.25, 41.35]	8.77	$<$ **0.001**
LexTALE-Esp	1.70 [0.901, 2.49]	4.18	< 0.001
Random effects			
σ^2	0.12		
$\tau_{00 Item}$	1543.33		
$\tau_{00 Participant}$	7446.57		
$\tau_{11 Participant[violation]}$	3090.40		
$\rho_{01 Participant}$	-0.42		
ICC	1.00		
$N_{Participant}$	34		
N_{Item}	224		
Observations	7,340		
Marginal R^2/ Conditional R^2	0.128 / 1.000		

8.E EEG data: by-violation type trial rejection rates

Table 8.E.1: *By-violation type trial exclusion rates for familiar and correct trials due to artefacts (n = 34). Note that the total number of familiar and correct trials was 7,258 as indicated in brackets.*

Violation type	**Trials excluded**	**Trials excluded (%)**
non-violation	173	2.38
violation	179	2.47
Total	352 (7,258)	4.85

8.F Model parameters: P600 component

Table 8.F.1: *Model parameters of the GAMM model for the effect of violation type and time on voltage amplitudes for the P600 ROI with channels CP1, CP2, CP5, CP6, Pz, P3, P4, PO3 and PO4 (n = 34). Estimated degrees of freedom (edf) provide a measure for the complexity of the smooth terms. The edf parameters for our smooth terms suggested that voltage amplitudes follow a highly non-linear tendency.*

Formula: voltage amplitudes ∼ violation type + channel + s(time) + s(time, by = violation type) + s(time, by = channel) + s(participant, channel, bs = "re") + s(participant, violation type, bs = "re") + s(participant, time, bs = "re") + s(participant, bs = "re") + s(item, time, bs = "re") + s(item, bs = "re")

Linear terms	**Estimate**	**SE**	**t-value**	**p-value**
(Intercept)	2.74	0.35	7.90	< 0.001
Violation type [violation]	0.295	0.238	1.24	0.215
Channel [CP5]	-1.04	0.198	-5.29	<**0.001**
Channel [P3]	0.004	0.197	0.019	0.985
Channel [Pz]	0.722	0.197	3.66	< 0.001
Channel [PO3]	0.396	0.197	2.01	0.045
Channel [PO4]	0.375	0.197	1.90	0.057
Channel [P4]	0.511	0.197	2.59	0.009
Channel [CP6]	-0.373	0.197	-1.89	0.059
Channel [CP2]	0.268	0.197	1.36	0.174
Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Time)	8.99	9.00	6736.07	< 0.001
s(Time):Violation type [violation]	8.99	9.000	3489.47	< **0.001**
s(Time):Channel [CP1]	8.96	9.00	297.604	< 0.001
s(Time):Channel [CP5]	8.95	9.00	706.482	< 0.001
s(Time):Channel [P3]	1.00	1.00	1.92	0.166
s(Time):Channel [Pz]	8.84	8.99	71.89	< 0.001
s(Time):Channel [PO3]	8.87	8.99	149.05	< 0.001

s(Time):Channel [PO4]	8.90	8.99	346.16	< 0.001
s(Time):Channel [P4]	7.72	7.97	275.58	< 0.001
s(Time):Channel [CP6]	8.91	8.99	359.00	< 0.001
s(Time):Channel [CP2]	8.95	8.99	130.50	< 0.001
s(Participant)	0.010	33.000	29.17	0.064
s(Participant, Channel)	269.20	297.00	865928.45	0.009
s(Participant, Violation type)	60.48	66.00	86774042.73	< 0.001
s(Participant, Time)	32.99	33.00	177593401.23	< 0.001
s(Item)	220.88	222.00	465329.02	< 0.001
s(Item, Time)	221.62	222.00	512839.12	< 0.001

8.G Model parameters: LAN component

Table 8.G.1: *Model parameters of the GAMM model for the effect of violation type and time on voltage amplitudes for the LAN ROI with channels Fp1, AF3, F7 and F8 (n = 34).*

Formula: voltage amplitudes ∼ violation type + channel + s(time) + s(time, by = violation type) + s(time, by = channel) + s(participant, channel, bs = "re") + s(participant, violation type, bs = "re") + s(participant, time, bs = "re") + s(participant, bs = "re") + s(item, time, bs = "re") + s(item, bs = "re")

Linear terms	**Estimate**	**SE**	**t-value**	**p-value**
(Intercept)	1.23	0.313	3.93	< 0.001
Violation type [violation]	-0.002	0.217	-0.011	0.991
Channel [AF3]	-0.238	0.151	-1.57	0.117
Channel [F7]	-0.447	0.151	-2.95	0.003
Channel [F3]	-0.451	0.151	-2.98	0.003
Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Time)	8.98	8.99	718.79	< 0.001
s(Time):Violation type [violation]	8.97	9.00	413.48	<**0.001**
s(Time):Channel [Fp1]	1.02	1.02	0.002	0.989
s(Time):Channel [AF3]	7.50	7.90	74.985	< 0.001
s(Time):Channel [F7]	8.61	8.93	153.45	< 0.001
s(Time):Channel [F3]	8.22	8.80	68.66	< 0.001
s(Participant)	0.012	33.00	8.59	1.00
s(Participant, Violation type)	57.88	66.00	16149514.65	1.00
s(Participant, Channel)	106.90	132.00	938593.99	1.00
s(Participant, Time)	32.99	33.00	35777233.25	0.999
s(Item)	219.54	222.00	96047.91	0.993
s(Item, Time)	221.18	222.00	96715.91	0.996