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

Von Grebmer Zu Wolfsthurn, S.

CHAPTER 2

Cross-linguistic interference in late language learners: An ERP study

This article is published as: Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (2021). Cross-linguistic interference in late language learners: An ERP study. Brain and Language, 221, 104993.

Abstract: This study investigated cross-linguistic interference in German low-proficient late learners of Spanish. We examined the modulating influence of gender congruency and cognate status using a syntactic violation paradigm. Behavioural results demonstrated that participants were more sensitive to similarities at the syntactic level (gender congruency) than to phonological and orthographic overlap (cognate status). Electrophysiological data showed that they were sensitive to syntactic violations (P600 effect) already in early acquisition stages. However, P600 effect sizes were not modulated by gender congruency or cognate status. Therefore, our late learners of Spanish did not seem to be susceptible to influences from inherent noun properties when processing non-native noun phrases at the neural level. Our results contribute to the discussion about the neural correlates of grammatical gender processing and sensitivity to syntactic violations in early acquisition stages.

Keywords: *multilingualism, late language learners, cross-linguistic interference, grammatical gender, gender congruency effect, cognate facilitation effect, P600 effect, single-trial EEG analysis, ERPs*

2.1 Introduction

How does our brain implement and represent two or more languages? This is an important question because it bears on language control and processing mechanisms in multilingualism. The current study focuses on *cross-linguistic interference* (also cross-linguistic influence), hereafter *CLI*. CLI is the interaction of the native language and any additional languages, which in turn has effects on the underlying processing mechanisms (Lemhöfer et al., 2008). The influence of the native language (L1) on the second (L2) or third language (L3) and vice versa has been studied across different linguistic domains, for example phonology and syntax (Cárdenas-Hagan, Carlson & Pollard-Durodola, 2007; Pika, Nicoladis & Marentette, 2006). A number of studies focusing on CLI between the languages in a multilingual system showed that it occurred independently of the L2 proficiency (Dijkstra & Van Heuven, 2002; Kroll & Tokowicz, 2005), the linguistic similarity between the languages (Blumenfeld & Marian, 2007; Cutler, Weber & Otake, 2006), the orthographic systems of the languages (Hoshino & Kroll, 2008) and their written scripts (Morford, Wilkinson, Villwock, Piñar & Kroll, 2011). Furthermore, CLI was demonstrated to occur in language production and in comprehension, at different ages of acquisition (AoA) both in adults (Hoshino & Kroll, 2008) and in children (Poarch & Van Hell, 2012a). Finally, CLI was studied in bilinguals as well as in late language learners. The latter are individuals who acquired the L2 or L3 post-puberty in adulthood (Frenck-Mestre, Anton, Roth, Vaid & Viallet, 2005). For the purpose of this study, we will be discriminating between *high-proficient late language learners*, i.e., individuals who have reached high overall attainment levels due to high exposure (Midgley, Holcomb & Grainger, 2011), and *low-proficient late language learners*, i.e., in-

dividuals who have not yet reached high attainment levels due to limited exposure (S. Rossi et al., 2006).

The current study focuses on CLI between *grammatical gender* systems in *low-proficient* late language learners. Broadly speaking, grammatical gender (hereafter *gender*) is a noun classification system. It can be used to form syntactic agreement between determiners and nouns, which in turn may form agreement with pronouns and adjectives. Gender is viewed as one of the most complex grammatical categories (Corbett, 1991) because it represents a lexical as well as a syntactic feature (Klassen, 2016). Gender systems for nouns differ depending on the language. For example, Italian, French and Spanish[1] (Romance family) have a two-way gender value system represented by *masculine* and *feminine*. German (Germanic family) has a system with three gender values: *masculine, feminine* and *neuter*. On the other hand, some languages from the Niger-Congo language family have seventeen or more gender values, e.g., Wolof (Babou & Loporcaro, 2016). In L1 acquisition, gender is known to be mastered relatively early in life (Unsworth et al., 2014). However, there is considerable variation between languages due to differing degrees of gender system transparency (Cornips & Hulk, 2008; M. Schwartz et al., 2015). Gender acquisition in a foreign language frequently represents a challenge for late learners despite high proficiency levels (Franceschina, 2005; Unsworth, 2008). Several studies support the notion of the strong influence of L1 on acquiring grammatical gender in the foreign language. More specifically, the interaction of the grammatical gender systems was shown to result in increased (or decreased) performance in gender assignment and acquisition (Franceschina, 2002; Paolieri, Padilla, Koreneva, Morales & Macizo, 2019; Sabourin et al., 2006).

A theoretical account about the representation of grammatical gender that reflects this notion of interfering gender systems is the

[1] In Spanish, nouns are either masculine or feminine; but note that the neuter gender does exist in the form of (demonstrative) pronouns such as “ello”, “esto” and “aquello”.

gender-integrated representation hypothesis, GIRH (Salamoura & Williams, 2007). According to this hypothesis, gender systems are shared between languages and the same (shared) gender node is activated when L1 and L2 match in gender. Under this view, words from L1 and L2 with a contrasting gender activate different nodes, but these nodes are still shared between languages. In contrast, the *gender-autonomous representation hypothesis*, GARH (Costa et al., 2003), predicts that L1 and L2 gender systems are independent. It also predicts that only language-specific gender nodes are activated. Both hypotheses on the possible organisation of the gender system in L1/L2 allow for testable predictions about CLI. The *GIRH* predicts interference from L1 to L2 and vice versa. This should be manifested in faster processing of *congruent* nouns compared to *incongruent* nouns. Nouns are defined as gender *congruent* when the gender values for nouns match across languages, for example for the noun [forest] in German and Spanish: der_M *Wald* and el_M *bosque*. Nouns are gender *incongruent* when the gender values for nouns do not match across languages, for example for the noun [duck] in German and Spanish: die_F *Ente* and el_M *pato*. Faster processing of congruent nouns occurs provided that the gender values across German and Spanish overlap at the conceptual level for the speaker. This processing facilitation effect was established as the *gender congruency effect* (Klassen, 2016) and is discussed in section 2.1.1. In contrast, the *GARH* does not predict interference and therefore no processing advantage of gender congruent items over gender incongruent items in terms of processing latencies. The *GIRH* has received substantial support in the literature (Bordag & Pechmann, 2007; Lemhöfer et al., 2008; Morales et al., 2016; Salamoura & Williams, 2007).

Critically, there is a second linguistic property which could be a potential contributor to the faster processing of gender congruent items, namely *cognate status*. Cognate status is an intrinsic property to noun stimuli. It is frequently manipulated in multilingual studies, see for example Lemhöfer, Dijkstra and Michel (2004). Cognates are words which overlap in their semantic, phonological and orthographic forms (Janyan & Hristova, 2007; C. Li & Gol-

lan, 2018). For example, the words *trono* [throne] in Spanish and *Thron* [throne] in German represent the category of *cognate* nouns, whereas *bosque* [forest] in Spanish and *Wald* [forest] in German are examples of *non-cognate* nouns. Numerous studies demonstrated faster processing of cognates over non-cognates in adult and child populations (Bosma et al., 2019; Hoshino & Kroll, 2008). This was termed the *cognate facilitation effect* (Costa, Santesteban & Caño, 2005; De Groot & Nas, 1991). The cognate facilitation effect was proposed to reflect CLI of the phonological systems of the two languages (Costa et al., 2005). Word production and comprehension tasks such as lexical decision tasks (Lemhöfer & Dijkstra, 2004), translation tasks (Davis et al., 2010) and picture-naming tasks (Hoshino & Kroll, 2008) showed that cognates were more susceptible to CLI compared to non-cognates.

Further supporting evidence for the cognate facilitation effect came from recent studies using behavioural paradigms in combination with electroencephalography, or EEG (Midgley et al., 2011). EEG is a non-invasive technique of recording brain activity and exploring online cognitive processes (Woodman, 2010). Researchers in EEG studies frequently focus on event-related potentials (ERPs), i.e., distinct brain oscillation patterns that arise in response to a particular stimulus or cognitive process. For example, Midgley et al. (2011) found distinct neural patterns for cognates vs. non-cognates in a semantic decision task. This notion manifested itself in larger N400 amplitudes for non-cognates compared to cognates for L1 English – L2 French highly proficient late learners. The N400 ERP component was previously associated with lexical and semantic integration, as well as lexical pre-activation and prediction (Szewczyk & Schriefers, 2018). The results were interpreted as showing greater ease of lexical and semantic integration for cognates compared to non-cognates.

To this date, it is unclear whether gender congruency and cognate status play a joint role in modulating foreign language processing and the associated neuronal patterns. Both are intrinsic properties of nouns that could drive CLI between two languages.

Previous studies did not systematically control for both properties, for example Costa et al. (2003), Lemhöfer and Dijkstra (2004), but see Lemhöfer et al. (2008). In the current study, we build on previous research by Lemhöfer et al. (2008), who manipulated both gender congruency and cognate status to examine their modulating role in CLI. In addition to behavioural measures as collected in Lemhöfer et al. (2008), we collected electrophysiological data to characterise CLI from an electrophysiological perspective in late language learners. To the authors' knowledge, this represents a unique constellation in terms of design and the population of interest.

2.1.1 The gender congruency effect

While the *cognate facilitation effect* demonstrates the interaction of the phonological systems of different languages, the *gender congruency effect* reflects the interaction amongst the grammatical gender systems of the languages within a multilingual system (Bordag & Pechmann, 2007; Klassen, 2016; Morales et al., 2016). It manifests itself in faster processing of gender congruent nouns compared to gender incongruent nouns. Therefore, this effect supports the notion of gender system interference and the *GIRH* of grammatical gender representation in bilinguals and late learners. The majority of studies supporting *GIRH* focused on intermediate ($\geq$ 3 years of language exposure) to highly proficient speakers (Bordag & Pechmann, 2007; Lemhöfer et al., 2008; Morales et al., 2016). In other words, the focus was on balanced or close to balanced simultaneous bilinguals, early bilinguals, and highly proficient late learners within the B2/C1/C2 proficiency range according to the Common European Framework of Reference for Languages (Council of Europe, 2001), hereafter *CEFR*. Only a few studies focused on CLI effects and the gender congruency effect in late language learners with low to moderate proficiency (Hahne, 2001; S. Rossi et al., 2006). Therefore, it remains unclear whether the findings from these studies (Bordag & Pechmann, 2007; Costa et al., 2003; Lemhöfer et al., 2008; Morales et al., 2016) are applicable to late language learners (AoA $>$ 12 years of age) with low

to intermediate proficiency levels within the A1/A2/B1/ B2 range and low exposure to the language (< 3 years). Our study aimed to contribute with new cross-linguistic evidence to the study of CLI of grammatical gender systems in late language learners with low to moderate (< 3 years of exposure) proficiency levels in the B1/B2 range. Further, a central focus of our study was to characterize the neural signature of CLI for which we hypothesised distinct neuronal patterns, as discussed in section 2.1.2.

2.1.2 CLI of grammatical gender and neural signatures

The majority of studies mentioned above aimed to characterise CLI in multilingual speakers from a behavioural perspective. Relatively recently, studies began to focus on the neural components of CLI in combination with ERPs (Ganushchak, Verdonschot & Schiller, 2011; Midgley et al., 2011). In the literature, there is an ongoing debate about the elicitation of the P600 ERP component in late language learners (Steinhauer et al., 2009; Van Hell & Tokowicz, 2010). The P600 effect was linked to syntactic phrase violations. It is reflected in a positive deflection of the EEG signal around 600 ms after stimulus onset in centro-parietal regions (Nichols & Joanisse, 2019; Osterhout, Mclaughlin, Pitkänen, Frenck-Mestre & Molinaro, 2006; S. Rossi et al., 2006; Steinhauer et al., 2009). More specifically, studies using this paradigm described more positive P600 amplitudes for syntactic violations compared to non-violations.

The P600 effect was reliably reported as an index for syntactic violations in highly proficient late learners with several years of exposure (Foucart & Frenck-Mestre, 2011; Gillon-Dowens, Vergara, Barber & Carreiras, 2010). For these learners, the P600 was also found to have a more bilateral distribution compared to the P600 effect in monolinguals (S. Moreno, Bialystok, Wodniecka & Alain, 2010). In contrast, studies on less proficient late learners (AoA > 12 years of age) frequently did not find a P600 effect in syntactic violation paradigms. This was the case for example for Chinese

late learners of English (Weber-Fox & Neville, 1996), Russian late learners of German (Hahne, 2001) and Japanese late learners of German (Hahne & Friederici, 2001). These results contrast with work by S. Rossi et al. (2006) who provided evidence of a smaller and delayed P600 effect around 1,000 ms in a syntactic violation paradigm in low-proficient Italian late learners of German. Another study on low-proficient English late learners of Spanish found evidence for a P600 effect in the violation trials between 500 ms and 900 ms (Tokowicz & MacWhinney, 2005). However, it neither showed a reduction nor delay of the P600 amplitudes. Moreover, Foucart and Frenck-Mestre (2011), who employed a syntactic violation paradigm combined with EEG on French monolinguals and proficient German-French late learners, found larger P600 amplitudes for gender congruent violation trials compared to incongruent violation trials (i.e., matching gender values affected L2 syntactic processing in L2 French speakers) for German-French late learners. This indicated distinct neural patterns for processing gender congruent vs. gender incongruent nouns. Taken together, the findings above from studies with late language learners with low to moderate proficiency levels and low exposure have provided contradictory findings until now: they found either an absent, a smaller or a delayed P600 effect. However, research in proficient late learners also suggests a modulation of the P600 effect as a function of gender congruency (Foucart & Frenck-Mestre, 2011), which warrants further investigation. It is worth adding that studies on syntactic violation processing in native speakers show varying findings in terms of ERP patterns. On one hand, some studies on Spanish native speakers reported N400/P300 effects for syntactic violations in noun-adjective and determiner-noun pairs (Barber & Carreiras, 2005, 2003). On the other hand, studies also reported biphasic N400/P600 patterns for syntactic and semantic violations in sentences (Martín-Loeches, Nigbur, Casado, Hohlfeld & Sommer, 2006; Wicha, Moreno & Kutas, 2004). Critically, these studies include a strong semantic component in the experimental design. Finally, the ERP literature on native speakers also revealed the more typical P600 effect for syntactic violations, frequently in combination with the LAN, in languages such as Spanish, English and Ger-

man (Barber & Carreiras, 2005; Hasting & Kotz, 2008; Molinaro, Barber & Carreiras, 2011; Münte, Matzke & Johannes, 1997; Osterhout & Mobley, 1995). In our study on late language learners, we exclusively focused on the P600 effect as it was more reliably reported for syntactic violations for multilinguals and language learners (Foucart & Frenck-Mestre, 2011; S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005).

2.1.3 The current study

Behavioural and electrophysiological research on CLI in German late language learners of Spanish (AoA > 12 years of age) with low to moderate proficiency levels (B1/B2) is scarce. Therefore, the present study explored the modulating role of *gender congruency* and *cognate status* on CLI effects. Our speakers were native German speakers who were late learners of Spanish (AoA > 16 years of age) with low exposure to Spanish (< 3 years). The aim of the study was threefold: first, we explored CLI from a behavioural and an electrophysiological (EEG) perspective. More specifically, we considered grammatical gender processing in the comprehension domain. Secondly, we examined a potential interaction effect of gender congruency and cognate status on processing accuracy and response latencies by employing a syntactic violation paradigm with Spanish noun phrases (NPs). The NPs consisted of "determiner + noun" sequences (Foucart & Frenck-Mestre, 2011). Within these NPs, we systematically manipulated gender congruency and cognate status of the target nouns with respect to the determiner they appeared with. This was to further explore the interaction between the gender congruency effect and the cognate facilitation effect. Lastly, we expanded on existing research on CLI effects in symmetric gender systems (e.g., Italian and Spanish) for highly proficient bilinguals (Paolieri et al., 2019; Salamoura & Williams, 2007) by examining CLI effects in German late learners of Spanish. Critically, we placed a strong focus on exploring ERP signatures in late language learners with low proficiency levels alongside behavioural measures. This design represented a significant extension of previous studies (Costa et al., 2003; Foucart & Frenck-Mestre, 2011) in

that these studies manipulated gender congruency alone, and not cognate status, and they examined highly proficiency speakers.

Further, we aimed to relate proficiency levels to inhibitory skills in late learners. Inhibitory skills are closely related to CLI: bilinguals and late language learners need to successfully manage interference between the languages, for example, in situations in which one language is more appropriate. It has previously been proposed that bilinguals and late language learners regulate parallel activation by means of inhibition (Bialystok, Craik & Luk, 2008). In bilingual research, the Stroop task has been frequently used to characterise bilingual inhibitory skills (Costa, Albareda & Santesteban, 2008; Goldfarb & Tzelgov, 2007). However, it has been scarcely implemented for late language learners. Therefore, we incorporated an implicit measure of these skills into our design, i.e., an adapted version of the *Stroop task* (MacLeod, 1992). The Stroop task is characterised by the absence or presence of a conflict inherent to a stimulus, traditionally resulting in a *congruent* and *incongruent* condition, respectively. The *Stroop effect* quantifies the difference in response times between the congruent and incongruent condition. Critically, a smaller Stroop effect has been previously associated with better inhibitory skills (Bialystok et al., 2008) and higher proficiency (Lev-Ari & Peperkamp, 2013) in bilingual studies. For example, Blumenfeld and Marian (2013) presented evidence that high-proficient Spanish – English bilinguals yielded smaller Stroop effects compared to low-proficient Spanish – English bilinguals. Relevant for our purposes, Lemhöfer and Broersma (2012) associated LexTALE vocabulary size scores between 60% and 80% as B2 level for Dutch – English bilinguals (Table 2.1). Scores below 60% were associated with level B1 and lower. This implies a positive link between proficiency and vocabulary size. Importantly, it was previously shown that participants who scored higher on this task had more in-depth knowledge of a language, not only in terms of knowing specific lexical items, but also with respect to their knowledge of phonological and orthographic rules (Diependaele, Lemhöfer & Brysbaert, 2013). Since it is still an open question whether these results hold for late language learners with low to moderate profi-

ciency levels, we further investigated the relationship between inhibitory skills and vocabulary size. In the present study, vocabulary size was measured using the LexTALE-Esp (Izura, Cuetos & Brysbaert, 2014). This Spanish version is based on the original LexTALE by Lemhöfer and Broersma (2012).

Table 2.1: *Relation between English proficiency levels and LexTALE scores in native Dutch speakers, from Lemhöfer and Broersma (2012).*

CEFR level	**CEFR description**	**LexTALE score**
C1 and C2	lower and upper advanced/proficient user	80% - 100%
B2	upper intermediate	60% - 80%
B1 and lower	lower intermediate and lower	below 59%

In addition to the Stroop task and the LexTALE-Esp, we measured participants' EEG in a syntactic violation paradigm while they were visually exposed to sets of Spanish NPs. The results from this study have important implications for characterising multilingual gender processing in low-proficient late language learners. Moreover, they provide further insight into the acquisition and the representation of grammatical gender in those speakers. Taken together, we investigated the following questions: whether processing accuracy and response latencies of Spanish NPs were modulated by gender congruency and cognate status, whether a P600 effect was present in late learners, and finally, whether this P600 effect was modulated by gender congruency or cognate status.

Hypotheses

For the Stroop task, we expected an effect of *condition*: if the word *left* (links/ izquierda) appeared on the left side of the screen (representing a *congruent* trial), we predicted faster response times (hereafter RTs) compared to when the word *left* appeared on the right side of the screen (representing an *incongruent* trial). The same applied to the word *right* (rechts/ derecha). Moreover, in line with previous results, we hypothesised that a smaller Stroop effect reflecting better inhibitory skills would be positively correlated

with a higher vocabulary score in the LexTALE-Esp (Blumenfeld & Marian, 2013; Lev-Ari & Peperkamp, 2013).

For the violation paradigm task, we predicted higher accuracy rates and shorter RTs for *non-violation* trials compared to *violation* trials (i.e., trials where the gender assignment was correct vs. incorrect). This would be behavioural evidence for the differential processing of NPs containing a syntactic violation compared to NPs which did not. More importantly, we also predicted higher accuracy and shorter RTs for *congruent* trials (i.e., gender values matched across German and Spanish) compared to *incongruent* trials (i.e., gender values did not match). This in turn would be evidence for CLI of the grammatical gender systems of German and Spanish. Critically, we expected a joint effect between *gender congruency* and *cognate status*: *congruent cognate* trials should elicit faster RTs and higher accuracy rates compared to *incongruent non-cognate* trials. This would also be evidence for CLI of the phonological and orthographic word forms during NP processing in Spanish.

For the EEG data recorded during the violation paradigm task, we first expected a P600 effect for late learners (S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005). This would be reflected in more positive voltage amplitudes for *violation* trials compared to *non-violation* trials, which represents the P600 effect. Second, we expected at least some modulation of P600 effect sizes due to an interaction of gender congruency and cognate status: we hypothesised a larger P600 effect for *congruent cognates* as a result of facilitatory interference from German compared to *incongruent non-cognates*. For the latter, we hypothesised that interference from German would slow down processing of a violation. In line with our behavioural hypotheses, this would be neural evidence for a P600 effect in late learners. Further, it would be neural evidence for CLI of the German and Spanish gender systems while processing Spanish violation NPs.

2.2 Methods

2.2.1 Participants

We recruited thirty-three right-handed native German participants (twenty-seven females) from the University of Konstanz (Germany) with a B1/B2 Spanish proficiency level in accordance with the CEFR (Council of Europe, 2001). All participants received a monetary compensation for their participation. Mean age of participants was 23.06 years (SD = 2.47). At the time of testing, none of the participants reported any learning or reading disorders, hearing impairments, visual impairments, psychological or neurological impairments. Participants' linguistic profile was assessed through an adapted version of the LEAP-Q Language Experience and Proficiency Questionnaire (Marian, Blumenfeld & Kaushanskaya, 2007), carried out with the original author's permission. The LEAP-Q was distributed prior to the experimental session via a home-based administration. This was done in an effort to obtain both an exhaustive description of participants' language use as well as descriptive proficiency measures. Furthermore, we increased ecological validity because no experimenter was present and social pressure was therefore reduced (Johnson & Fendrich, 2005; Rosenman, Tennekoon & Hill, 2011). Complying with the Ethics Code for linguistic research at the Faculty of Humanities at Leiden University, participants signed an informed consent form prior to their participation.

LEAP-Q: Linguistic profile of participants

In the LEAP-Q, the vast majority of the thirty-three participants (n = 31) indicated that English was their first foreign language (L2) after acquiring L1 German, with M_{AOA} = 8.90 (SD_{AOA} = 1.90). The remaining two participants learnt French as their first foreign language (L2) with M_{AOA} = 8.5 (SD_{AOA} = 2.5). Sixteen participants reported learning Spanish as a second foreign language (L3). Fifteen participants disclosed Spanish as their third foreign language (L4). Finally, two participants reported Spanish as their fourth foreign language (L5). See Appendix 2.A for details about

the linguistic background reported by the participants. With regards to Spanish, the main language of interest in this study, the mean age of acquisition was $M_{AOA} = 16.29$ ($SD_{AOA} = 2.39$). The self-reported fluency age was $M_{AOA} = 18.53$ ($SD_{AOA} = 2.29$), and reading onset age was $M_{ROA} = 17.27$ ($SD_{ROA} = 3.03$). A total of thirty-one participants spent on average $M = 0.96$ years ($SD = 0.69$) in a Spanish-speaking country (e.g., Spain, Chile, Argentina, Puerto Rico, Colombia). On a scale from zero to ten (ten corresponding to reporting maximal proficiency), participants quantified their speaking proficiency with $M = 6.76$, ($SD = 1.00$); listening comprehension proficiency with $M = 7.34$ ($SD = 0.92$); and finally, reading proficiency with $M = 7.18$ ($SD = 1.07$). At the time of testing, participants were exposed to Spanish through interaction with Spanish native speakers, radio shows, television, reading or self-instructions on average $M = 3.12$ ($SD = 2.31$) on a scale from zero to ten (with ten being maximally exposed to the language). This compared to an average exposure of $M = 5.20$ ($SD = 2.48$) for the L2, and to $M = 1.34$ ($SD = 2.04$) for their L3. Further, participants indicated the order of known languages in terms of which language they felt they were most proficient in at the time of testing (current perceived proficiency). Despite the fact that most of the participants formally acquired Spanish as their L3, four participants nevertheless reported Spanish as their current perceived L2, and twenty-six participants as their current perceived L3. In other words, most participants reported that their Spanish levels were equivalent to their L3. This was taken as a proxy indicator for their confidence in their language skills for Spanish.

2.2.2 Materials and design

During the experimental session, participants completed three experimental tasks. We first measured their Spanish vocabulary size in the LexTALE-Esp (Izura et al., 2014). They then completed a Stroop task to measure inhibitory skills (MacLeod, 1992). Finally, participants performed a violation paradigm task which examined grammatical gender processing in Spanish (Foucart & Frenck-Mestre, 2011). Participants' EEG was measured exclusively

during the last task. All three experimental tasks were programmed in E-Prime 2.0 (Psychology Software Tools, Inc.).

LexTALE-Esp

For the LexTALE-Esp, we transformed the original Spanish pen-and-paper version (Izura et al., 2014) into a computer-based equivalent for administration in the laboratory.

Stroop task

For the Stroop task, the stimuli were translation equivalents of the words *left* and *right* in German and Spanish (*links, rechts*, and *izquierda, derecha*, respectively). Participants were asked to respond to a target word while ignoring its location on the screen. This task served as a measure for inhibitory skills, which was subsequently correlated with performance on the LexTALE-Esp to establish a potential correlation between inhibitory skills and proficiency.

Violation paradigm task

For the violation paradigm task, stimuli nouns were taken from the MultiPic database (Duñabeitia et al., 2018) and the Spanish Frequency Dictionary (Davies & Davies, 2017). The MultiPic database includes 750 coloured drawings of common objects. They were standardised for name agreement across a range of languages, including Spanish, British English, German, Italian, French, Dutch (Belgium) and Dutch (The Netherlands). We selected the nouns where participants had provided the highest percentage of the correct name of the object, and items where they most often gave the most frequent name of the object across German and Spanish. We also selected additional highly frequent nouns from the Spanish Frequency Dictionary (Davies & Davies, 2017). We then assigned each noun a gender congruency status (congruent or incongruent in German and Spanish) and a cognate status (cognate or non-cognate in German and Spanish) on the basis of the semantic, orthographic and phonological overlap these nouns had across German and Spanish. We omitted identical cognates (e.g., *das Taxi*

– *el taxi* [the taxi]), nouns which take a plural form in German or Spanish (e.g., *die Brille – las gafas* [the glasses]), professions with a biological gender (e.g., *die Tänzerin – la bailarina* [the female dancer]), English loan words (e.g., *der Boomerang – el boomerang* [the boomerang]), ambiguous gender assignment cases which commonly elicit debates among native Spanish speakers, such as *el ancla* [the anchor] (feminine gender but determiner takes the masculine form due to initial stress and */a/* onset); and lastly, nouns which had two translation equivalents with opposing genders (e.g., *der Esel – la mula/el burro/el asno* [the donkey]) to avoid ambiguity. An additional relevant feature of the stimuli for this task was the systematic matching of terminal morphemes to the natural distribution of word endings in Spanish (Appendix 2.B). This was modelled after work by Clegg (2011). As discussed by Sá-Leite, Fraga and Comesaña (2019), terminal phonemes cannot be taken as strict cues to infer the grammatical gender, despite being probabilistic cues. In addition, excluding nouns on the basis of their terminal phonemes would drastically decrease the ecological validity of our stimuli. Finally, we created a balanced masculine-to-feminine stimuli ratio with 55.36% and 44.64%, respectively, in line with research by Eddington (2002) and Bull (1965). In comparison, in German 38.8% of monomorphemic nouns are masculine, 35.4% are feminine and 25.9% are neuter (Schiller & Caramazza, 2003).

2.2.3 Procedure

LexTALE-Esp

We first administered the LexTALE-Esp task to determine participants' vocabulary size in Spanish. The test consisted of a visual lexical decision task in Spanish. Participants were presented with a letter string on the screen and had to decide via a button press whether or not this letter string was part of the Spanish lexicon. Letter strings were presented along the horizontal midline. Thirty of the eighty-seven items were pronounceable Spanish *pseudowords* (e.g., *grodo*), whereas fifty-seven items were Spanish *words*. Three words were excluded from the original stimulus set due to overlap

with our experimental stimuli. This resulted in thirty pseudoword trials and fifty-seven word trials. Each letter string was presented once. A typical trial was initiated by a fixation cross of a duration of 1,000 ms, followed by a single letter string display until the participant's response. Post-test, we calculated a vocabulary size score (percentage of correctly identified words minus percentage of incorrectly identified pseudowords). The maximum score was 100. Participants were told that incorrectly assigning the word status to a pseudoword would lead to a deduction of points from the final score.

Stroop task

The second task of the study was the Stroop task, which featured a conflict between the target word and the location of the target word. It consisted of two blocks, one for target words in German, and one for Spanish. Each block consisted of ninety-six trials, with a total of 192 experimental trials for both target languages. Trial order was randomised in both blocks. Prior to completing the first block, we included four practise trials to familiarise participants with the procedure. Upon initiation of a trial, a fixation cross was displayed for 500 ms on a white screen, followed by the display of the target word for 1,000 ms. In the first block, participants were presented with exclusively German target words, e.g., "*links*" or "*rechts*" for left or right, respectively. The target word appeared either on the left or the right side of the screen along the horizontal midline. Participants were visually instructed in German to indicate whether the word corresponded to the word "left" or the word "right", while ignoring the location of the target word on the screen. Half of the trials were *congruent*, i.e., the target word and the location of the target word matched, and the other half was *incongruent*, i.e., the target word and the location of the target word did not match. The procedure in the second block was identical, however the instructions and the targets were displayed in Spanish. Therefore, the target words were "*izquierda*" and "*derecha*" for left and right, respectively. We opted to present the Spanish block in second place to induce a bilingual mode in our participants

(Grosjean, 2012; Stocker & Berthele, 2020) in preparation for the subsequent violation paradigm task conducted in Spanish.

Violation paradigm task

As our final task, we implemented an EEG version of a syntactic violation paradigm similar to that used by Foucart and Frenck-Mestre (2011). In contrast to Foucart and Frenck-Mestre (2011), we opted for the presentation of an NP as opposed to a full sentence to reduce automatic prediction of an upcoming noun or gender category. This prediction process was previously linked to an N400 effect (Szewczyk & Schriefers, 2018). We included eight practice trials to familiarise participants with the task procedure. Four of the practice trials contained complex infrequent nouns typically unknown to low-proficient learners of Spanish at the B1/B2 level (e.g., *estiércol* [dirt]). This was an additional measure whether participants were reliable in their answers about familiarity with the noun. The task procedure was as follows: we first presented participants with a fixation cross for 1,000 ms. Then they were visually presented with a single target noun (e.g., *bosque* [forest]) along the horizontal midline. Participants indicated whether or not they were familiar with the noun. We then presented participants with the same target noun within an NP configuration (i.e., determiner + noun: el_M $bosque_M$ [the forest]) for 3,000 ms, or until a button-press response was registered. Participants' task was to indicate as accurately and as fast as possible whether the NP was grammatically correct. While participants were exposed to the NP, their EEG was recorded.

The task design was a full factorial design (2 x 2 x 2) with three independent variables adding to a total of eight conditions: half of the presented noun phrases were *violation* trials, where the determiner was grammatically correct (e.g., el_M $bosque_M$), whereas the other half were *violation* trials (e.g., la_F $bosque_M$). Of both the violation and non-violation trials, we manipulated *gender congruency*, i.e., half of the trials had matching gender values across languages (*congruent* trials), whereas the other half had non-matching

gender values (*incongruent* trials). Finally, we manipulated *cognate status* of nouns: half of the trials were cognates (*cognate* trials), and the other half were not (*non-cognate* trials). See Table 2.2 for a sample set of stimuli. Target nouns were controlled for frequency, and number of syllables ($M = 2.74$, $SD = 0.81$). There were twenty-eight trials for each of the eight conditions, adding up to 224 trials. Trial order was fully randomised to present each participant with a unique order of trials. Participants were given short breaks throughout the task. Furthermore, we reminded participants through a text display to give fast and accurate responses. Upon termination of all three tasks, participants were given a debrief letter and were asked to sign the final consent form.

Table 2.2: *Sample set of stimuli for the violation paradigm task, illustrating the three manipulations: violation type, gender congruency and cognate status.*

		non-violation	
		congruent	**incongruent**
cognate	**German**	der_M $Traktor_M$	die_F $Garage_F$
	Spanish	el_M $tractor_M$	el_M $garaje_M$
		the tractor	*the garage*
non-cognate	**German**	der_M $Wald_M$	die_F $Ente_F$
	Spanish	el_M $bosque_M$	el_M $pato_M$
		the forest	*the duck*
		violation	
		congruent	**incongruent**
cognate	**German**	der_M $Thron_M$	die_F $Pistazie_F$
	Spanish	$*la_F$ $trono_M$	$*la_F$ $pistacho_M$
		the throne	*the pistachio*
non-cognate	**German**	die_F $Treppe_F$	die_F $Reise_F$
	Spanish	$*el_M$ $escalera_F$	$*la_F$ $viaje_M$
		the stairs	*the trip*

EEG recordings

The EEG data were collected with passive electrodes using the BrainVision Recorder software 1.10 (Brain Products GmbH). We used a standard 32-electrode 10/20 montage at a sampling rate of 500 Hz (Appendix 2.C). We recorded the vertical electrooculogram (VEOG) from one external facial electrode placed below the participant's left eye. We also recorded the horizontal electrooculogram (HEOG) from two electrodes at the outer canthus of each eye. The EEG recording was originally referenced to the central electrode Cz. It was later re-referenced offline to the mastoid electrodes TP9 and TP10. The ground electrode was placed on the right cheek of participants. We configured electrodes via the actiCAP 2 software (Brain Products GmbH) to ensure optimal conductivity. Impedances were kept below 10 kΩ. for the cap and eye electrodes, and below 5 kΩ. for the ground and reference electrode.

2.3 Results

2.3.1 Behavioural data exclusion

Stroop task data from one participant were excluded because of a failure to follow the task instructions. For the analysis of the RTs of the Stroop task, we considered only correct trials. For the violation paradigm task, we only included correct and familiar trials in the analysis, i.e., where participants indicated familiarity with the single target noun. This was to minimise the risk of employing (confounding) guessing strategies during the main experimental trials.

2.3.2 Behavioural data analysis

LexTALE-Esp scores were computed offline and added as a variable in the analysis for the Stroop task. We calculated Stroop effects by subtracting the RTs in the congruent condition from the RTs in the incongruent condition. Behavioural data from the Stroop task and the violation paradigm task were analysed using R and RStudio (R Core Team, 2020). For both tasks, we modelled accuracy

and RTs separately following a mixed effect model approach using the *lme4* package (Bates et al., 2020). We employed a generalised linear mixed effect model (*GLMM*) using the *glmer()* function with a binomial distribution and a gamma distribution to model the binomially distributed accuracy data and positively skewed RT data, respectively. In contrast, we fitted a linear mixed model (*LMM*) using the *lmer()* function to generate mixed effects models for our normally distributed RT data. Absolute t-values > 1.96 were interpreted as statistically significant with $\alpha = 0.05$ (Alday, Schlesewsky & Bornkessel-Schlesewsky, 2017). Random effects were chosen to be as maximal as possible without over-parameterisation to balance Type-I errors and power (Matuschek et al., 2017).

We followed a maximal model building approach where we maintained the simplest possible model structure in light of our main manipulations (Bates, Kliegl, Vasishth & Baayen, 2018; Winter, 2019). These were *condition* for the Stroop task, and *violation type, gender congruency* and *cognate status* for the behavioural analysis of the violation paradigm task. For the Stroop task, we also added *LexTALE-Esp score* as a covariate. For the violation paradigm task, we included *LexTALE-Esp score, order of acquisition of Spanish* (i.e., whether Spanish was acquired as L3 vs. L4 vs. L5), *terminal phoneme, Stroop effect* and *target noun gender* as potential covariates. The model selection procedure was as follows: we constructed separate models with different predictor variables (with and without interactions and random slopes). We subsequently performed model fit checks by plotting the model residuals against predicted values. We used the *anova()* function to perform model comparisons and likelihood ratio tests on the basis of the Akaike's Information Criterion, AIC (Akaike, 1974), the Bayesian Information Criterion, BIC (Neath & Cavanaugh, 2012) and the log-likelihood in order to establish the best-fitting model for our data. Where applicable, we performed Tukey corrected post-hoc contrasts to estimate effect sizes using the *emmeans()* function (Lenth et al., 2019).

2.3.3 Behavioural data results

LexTALE-Esp

LexTALE-Esp scores were calculated by subtracting the percentage of incorrect word identifications from correct word identifications (percentage-yes-responses to words minus percentage-yes-responses to pseudowords). All of our speakers fell into the B1 level category or below according to their LexTALE-Esp scores (Lemhöfer & Broersma, 2012). The mean LexTALE-Esp score was $M = 18.91$ ($SD = 20.45$). There was a large variation in scores with a range from –23 to 60. Note that scores were not taken as an absolute measure for proficiency but as a measure for vocabulary size.

Stroop task

We first examined Stroop effects for each target language separately. In a second step, we explored whether Stroop effects correlated with LexTALE-Esp vocabulary size scores. A summary of mean accuracy rates and RTs for the target languages German and Spanish is shown in Table 2.3.

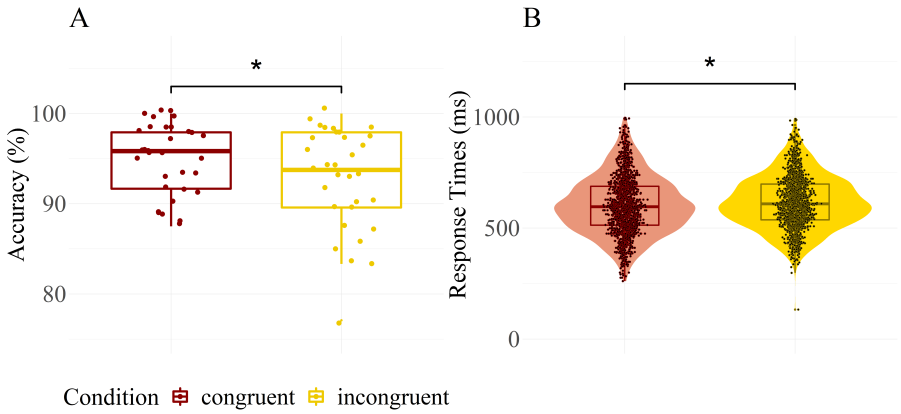
Table 2.3: *Descriptive statistics for both target languages for each condition (n = 32).*

	Accuracy	**RTs (ms)**
German	**Mean (SD)**	**Mean (SD)**
congruent	0.951 (0.216)	603 (130)
incongruent	0.928 (0.258)	616 (119)
Stroop effect	0.023	13
Spanish	**Mean (SD)**	**Mean (SD)**
congruent	0.956 (0.204)	556 (110)
incongruent	0.938 (0.242)	576 (108)
Stroop effect	0.02	20

German Target Block. For accuracy rates, we fitted a GLMM and explored *condition* (*congruent* vs. *incongruent*) as fixed effect, *subject* and *item* as random effects. Our final model for the accuracy data included *condition* as a fixed effect, and *subject* and *item* as random effects (Appendix 2.D). By-*subject* random slopes for *condition* did not significantly improve the model fit $\chi^2(2, n = 32) = 2.02$, $p = 0.364$, and neither did *LexTALE-Esp score* as a covariate, which yielded singular fit. Accuracy for the *congruent* condition was significantly different with $\beta = 0.659$, *95% CI*[0.437, 0.994], $z =$ -1.99, $p = 0.047$ compared to the *incongruent* condition, i.e., participants were significantly more accurate in the congruent condition compared to the incongruent condition (Figure 2.1). The model of best fit was: accuracy $\sim$ condition + (1|subject) + (1|item).

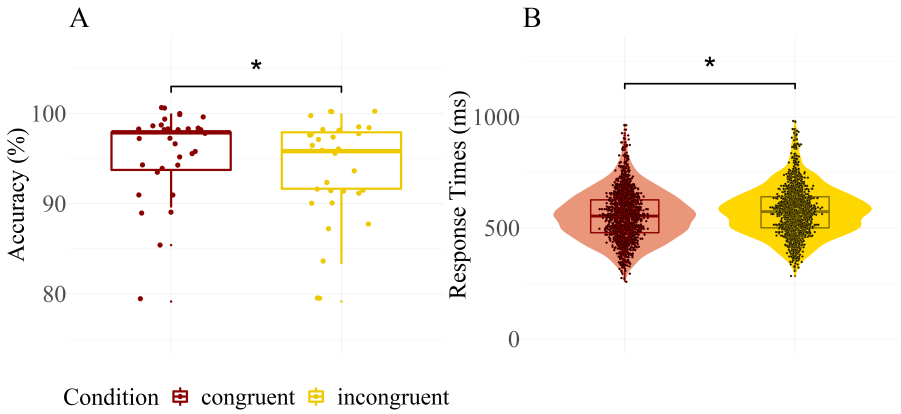
We followed a similar analysis approach for the RTs, for which we fitted an LMM. The model of best fit contained *condition* and *LexTALE-Esp score* as an interaction effect, and *subject* as random effect (Appendix 2.D). However, the interaction effect was not significant with $\beta =$ -0.014, *95% CI*[-0.410, 0.382], $t =$ -0.069, $p =$ 0.945. Participants were significantly faster in the *congruent* condition with $\beta = 13.08$, *95% CI*[2.48, 23.68], $t = 2.42$, $p = 0.016$ compared to the *incongruent* condition (Figure 2.1). The model of best fit was: RTs $\sim$ condition * LexTALE-Esp score + (1|subject). In sum, we found an effect of condition on both accuracy rates and RTs when participants were exposed to the target words in German. In a final step, we calculated the *Stroop effect* (RTs incongruent condition minus RTs congruent condition) for the German targets for each participant in order to explore a correlation between better inhibitory control skills (i.e., a smaller Stroop effect) and higher *LexTALE-Esp scores* indexing vocabulary size in Spanish. The range of the Stroop effect was -57.73 ms to 62.49 ms. We correlated the Stroop effect with the *LexTALE-Esp score* for each participant. We did not find evidence for a correlation between LexTALE-Esp scores and the size of the Stroop effect for German targets ($R = 0.014$, $p = 0.94$).

Figure 2.1: *Mean accuracy rates for each participant (A) and mean RTs (B) for German Stroop targets (n = 32).*



Spanish Target Block. We followed a similar procedure for the Spanish targets. The GLMM of best fit for accuracy rates included *condition* as fixed effect and *subject* as random effect (Appendix 2.E). The model highlighted that participants were once again more accurate in the *congruent* compared to the *incongruent* condition with β = 0.676, *95% CI*[0.491, 0.930], z = -2.41, p = 0.016 (Figure 2.2). The best-fitting model was: accuracy ~ condition + (1|subject). The LMM of best fit for RTs included *condition* as main effect and *subject* as random effect, while *LexTALE-Esp score* did not emerge as a covariate (Appendix 2.E). There was a significant difference for RTs between the *congruent* and the *incongruent* condition with β = 20.55, *95% CI*[13.44, 27.65], t = 5.67, p < 0.001. Participants were statistically faster in the *congruent* compared to the *incongruent* condition (Figure 2.2). The model of best fit was: RTs ~ condition + (1|subject). Finally, we tested whether there was a correlation between the *Stroop effect* and *LexTALE-Esp scores*. With R = 0.024 and p = 0.900, we did not find supporting evidence for a positive correlation between a smaller Stroop effect and higher LexTALE-Esp scores. This mirrored the results from the German targets.

Figure 2.2: *Mean accuracy rates for each participant (A) and mean RTs (B) for Spanish Stroop targets (n = 32).*



Comparison Stroop Effect. For reasons of completeness, we performed additional analyses to compare the Stroop effect across the German and the Spanish block. Participants were overall faster in the Spanish block compared to the German block with β = -43.90, t = -15.908, $p < 0.001$. As reported in previous sections, there was also an effect of *condition*, with participants being significantly faster in the *congruent* compared to the *incongruent* condition with $\beta = 16.47$, $t = 4.23$, $p < 0.001$. However, because the Spanish block was always presented as the second block, we argue that our results are consistent with a simple practice effect. Critically, we did not find evidence for an interaction effect of target language and condition, indicating that the Stroop effect was statistically comparable across the German and the Spanish block. The difference in Stroop effect was not the focus of this study, but should be investigated more closely in future experiments.

Violation paradigm task

In this task, we explored the effect of gender congruency and cognate status on accuracy and RTs. See Table 2.4 for mean accuracy rates and RTs for each condition (N = 33).

Table 2.4: *Descriptive statistics for the violation paradigm task for each condition for familiar nouns (N = 33).*

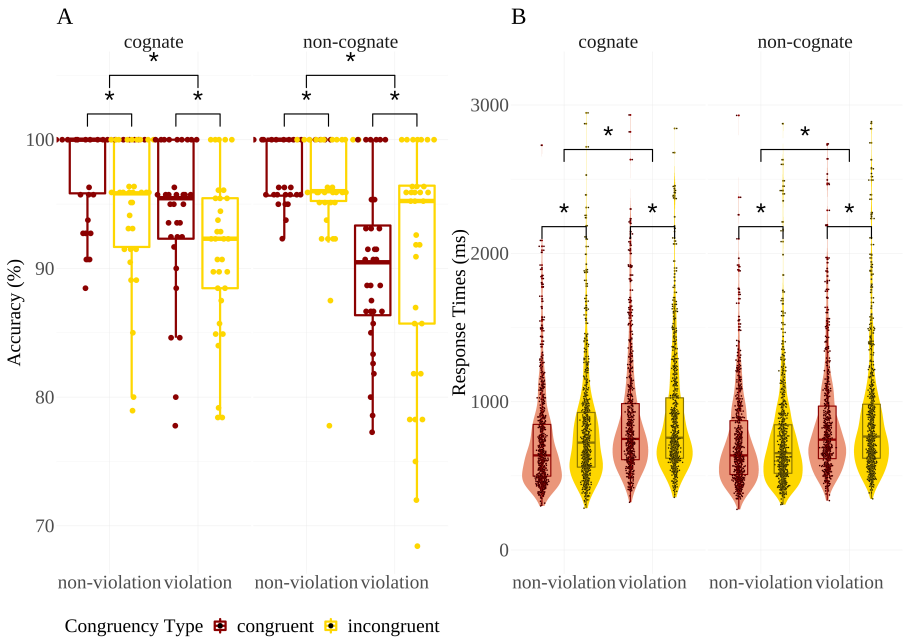
		non-violation	
		Accuracy	**RTs (ms)**
		Mean (SD)	**Mean (SD)**
cognate	**congruent**	0.979 (0.144)	729 (328)
	incongruent	0.948 (0.223)	802 (369)
Difference		0.031	73
non-cognate	**congruent**	0.978 (0.147)	727 (312)
	incongruent	0.959 (0.197)	740 (348)
Difference		0.019	13
		violation	
		Accuracy	**RTs (ms)**
		Mean (SD)	**Mean (SD)**
cognate	**congruent**	0.947 (0.224)	865 (389)
	incongruent	0.915 (0.279)	886 (395)
Difference		0.032	21
non-cognate	**congruent**	0.904 (0.294)	847 (373)
	incongruent	0.914 (0.280)	883 (416)
Difference		-0.010	36

Accuracy. As described above, we employed a GLMM approach for the accuracy of the grammaticality judgment on NPs, which contained either a gender-related violation or not. The model of best fit yielded *violation type* and *gender congruency* as main effects, as well as *subject* and *item* as random effects, with no interactions or random slopes (Appendix 2.F). The covariate *LexTALE-Esp score* was also included as main effect in the model ($\beta = 1.027$, *95% CI*[1.02, 1.04], $z = 4.58$, $p < 0.001$). As predicted, there was a significant difference between *non-violation* and *violation* trials with $\beta = 0.389$, *95% CI*[0.235, 0.644], $z = -3.68$, $p < 0.001$, and *congruent* and *incongruent* trials with $\beta = 0.539$, *95% CI*[0.326, 0.892], $z = -2.40$, $p = 0.016$. Therefore, participants were more accurate in *non-violation* trials and in *congruent* trials (Figure 2.3). Further, the other hypothetical covariates (*order of acquisition of*

Spanish, terminal phoneme, Stroop effect and *target noun gender*) either led to non-convergence or did not improve the model fit. The model of best fit was: accuracy ∼ violation type + gender congruency + LexTALE-Esp score + (1|subject) + (1|item). Note that estimates are provided as odds ratios.

Response times. We followed a similar GLMM approach to model RTs for familiar and correct trials of the violation paradigm task. The model of best fit included *violation type, gender congruency* and *cognate status* as main effects, as well as *subject* and *item* as random effects (Appendix 2.G). *Terminal phoneme* emerged as covariate in the best-fitting model. Instead, *order of acquisition of Spanish, LexTALE-Esp score, Stroop effect* and *target noun gender* were not included in the best fitting model as their inclusion led to non-convergence. The final model showed shorter RTs for *non-violation* trials compared to *violation* trials with $\beta = 116.12$, *95% CI*[102.68, 129.56], $t = 16.94$, $p < 0.001$, for *congruent* compared to *incongruent* trials with $\beta = 34.54$, *95% CI*[20.53, 48.55], $t = 4.83$, $p < 0.001$, and finally, for *non-cognate* compared to *cognate* trials with $\beta = -19.75$, *95% CI*[-32.70, -6.80], $t = -2.99$, $p = 0.003$ (Figure 2.3). Thus, participants were statistically faster in *non-violation* trials, *congruent* trials and *non-cognate* trials. The model of best fit was: RTs ∼ violation type + gender congruency + cognate status + terminal phoneme + (1|subject) + (1|item).

Figure 2.3: *Accuracy (A) and RTs (B) for violation type and gender congruency for the violation paradigm task (N = 33).*



2.3.4 EEG data exclusion

One EEG data set for the violation paradigm task was excluded due to a recording failure. Further, we defined a set of exclusion criteria for the EEG data in order to determine outliers: first, we only included trials in the analysis where participants indicated familiarity with the target noun. As the maximum number of familiar targets was 218 out of 224 trials for this dataset, we adopted a threshold of 218 as the new upper limit for trials upon which further calculations were based. Second, we only included trials were participants accurately detected a (non–) violation. Finally, we explored the signal-to-noise ratio for each condition for each participant via data pre-processing and artefact rejection. Only participants with a remainder of at least 60% of trials were included in the analysis. The total number of rejected trials due to artefacts was 271 (4.94%) out of a total of 5,486 familiar and correct trials. See Appendix 2.H for rejection rates for each condition. On the basis of these exclusion

criteria, we subsequently excluded the EEG data of four additional participants. Therefore, twenty-eight participants were included in the EEG analysis.

2.3.5 EEG data pre-processing

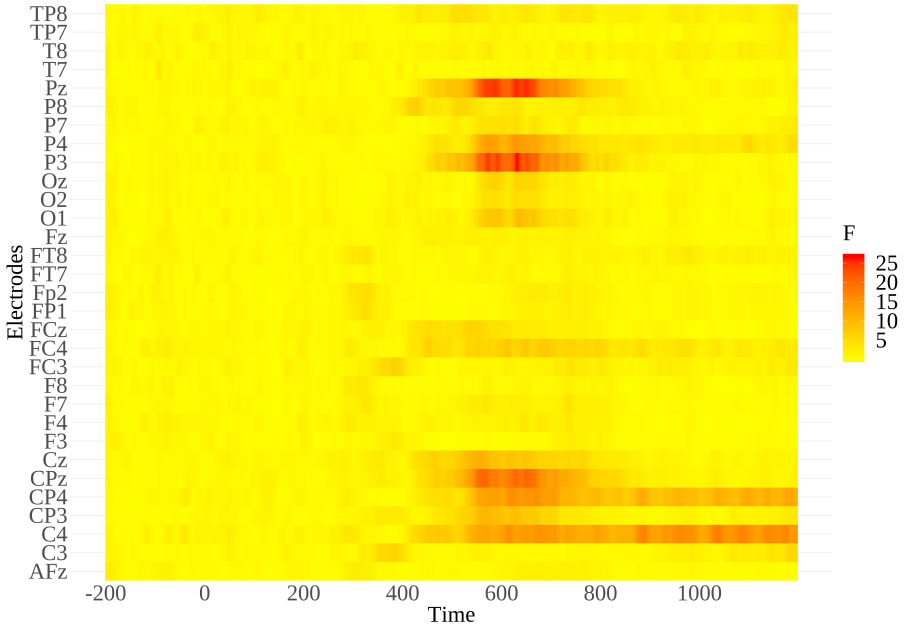
We processed the EEG data using BrainVision Analyzer Version 2.1 (Brain Products GmbH). We followed a classical EEG data pre-processing procedure for language-comprehension related phenomena (Foucart & Frenck-Mestre, 2011). It consisted of visual inspection of the signal, re-referencing, linear derivation for the HEOG electrodes (combining the two electrodes placed at the outer canthus) and filtering at a low-pass filter of 0.1 Hz and a high-pass filter of 30 Hz. We then performed ocular correction and artefact rejection. The HEOG was computed by merging the activity from the two electrodes placed on the outer canthus of the left and right eye. We defined the electrode placed underneath the left eye as VEOG. Offline, we re-referenced the recordings to the average of the left and right mastoid electrodes (TP9 and TP10). Next, signal segmentation and epoching was applied to familiar target words and correct trials only. We generated epochs around the stimulus onsets to explore the voltage amplitudes for the ERP component of interest, namely the P600. We deliberately selected a longer epoch period of 1,400 ms in total because of potentially later P600 effects, which are known to be observed in late learners (S. Rossi et al., 2006). Therefore, we defined the range of the epochs from 200 ms prior to the onset of the target NP to 1,200 ms after the onset of the target NP. Segments marked as bad during artefact rejection were excluded from the analyses. We performed baseline correction for each segment using the average EEG activity in the 200 ms prior to NP onset.

2.3.6 EEG data analysis

After pre-processing and exporting our data, we performed a cluster-based permutation analysis to tentatively explore the regions of interest and the potential time windows associated with

significant modulations of the EEG signal. For this, we used a permutation test from the *permutes* R package (Voeten, 2019) which included the voltage amplitudes for all data electrodes across the entire exported time window of 1,400 ms across conditions. As evident in Figure 2.4, the output of this test demonstrated potentially significant modulations of the EEG signal in the time window between 500 ms and 900 ms post-NP onset for posterior electrodes. This time window and ROI have been previously associated with the P600 (Nichols & Joanisse, 2019; Osterhout et al., 2006; S. Rossi et al., 2006; Steinhauer et al., 2009; Tokowicz & MacWhinney, 2005). Further visual inspection of the output did not suggest significant EEG signal modulation in windows prior to the time window associated with the P600. Finally, we divided electrodes into nine areas of interest in line with standard P600 analysis procedures (Foucart & Frenck-Mestre, 2011), i.e., left anterior, central anterior, right anterior, left medial, central medial, right medial, left posterior, central posterior and right posterior regions. On the basis of the output from the permutation test and previous literature associating centro-parietal regions to the P600 (Osterhout et al., 2006; S. Rossi et al., 2006; Steinhauer et al., 2009), we defined our ROI as the following thirteen electrodes: *CPz, CP3, CP4, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1, O2*. These electrodes we located in left posterior, central posterior and right posterior regions.

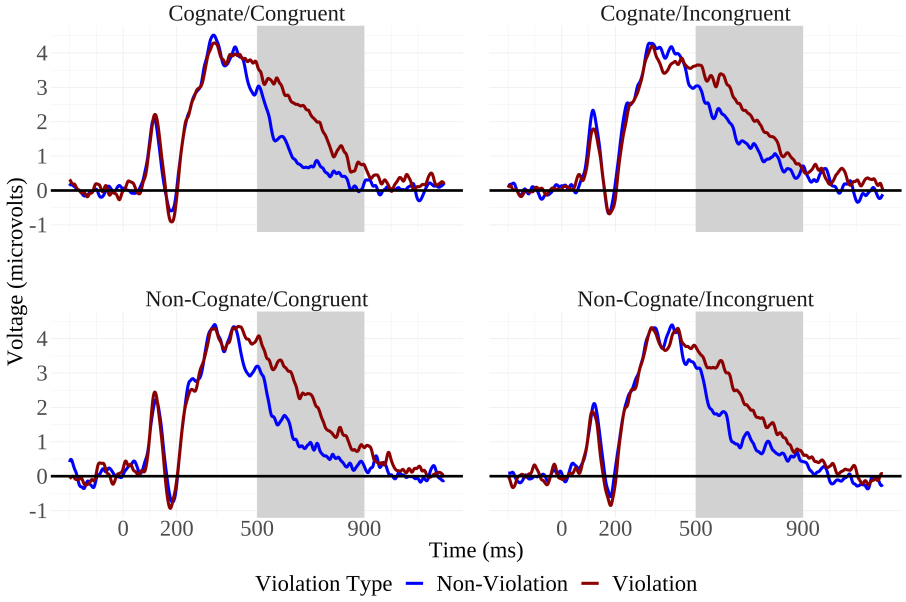
Figure 2.4: *Output of permutation test across conditions for all data electrodes for the exported time window of 1,400 ms including the corresponding F-values (n = 28). Larger F-values are shown in darker colours and denote an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes.*



Next, we followed a linear mixed effects models (LMM) approach on a *single-trial* basis (Frömer et al., 2018) in R and RStudio (R Core Team, 2020) in an effort to expand on the traditional average-type analysis. This latter method has been heavily criticised due to its limitations in terms of equally weighted observations on a by-condition and by-participant basis, and independent factor levels. These assumptions are frequently compromised for reasons of design and during the EEG data pre-processing stages. An alternative method are single-trial LMMs endorsed by an increasing number of researchers since its first application to EEG data in 2011 (Amsel, 2011). These models include both fixed effects and estimates for the random variance between subjects and items, namely random effects (Kornrumpf, Niefind, Sommer & Dimigen, 2016). They can be applied to data sets with variability in effect sizes and to unbalanced designs (Baayen et al., 2008; Fröber et al., 2017). For

the single-trial LMM approach, we included all available voltage values for each epoch of 1,400 ms without averaging across segments from the same condition in order to preserve by-subject and by-item variance. We considered *violation type, gender congruency* and *cognate status* as fixed effects, as well as *hemisphere, LexTALE-Esp score, order of acquisition of Spanish, terminal phoneme, Stroop effect* and *target noun gender* as covariates. We included *subject* and *item* (i.e., the individual NP) as random effects in the single-trial analysis. The model-fitting procedure was similar to the behavioural analyses. We followed a maximal building approach in light of hypotheses while maintaining the simplest possible model structure (Bates et al., 2018; Matuschek et al., 2017; Winter, 2019). Figure 2.5 shows the mean voltage amplitudes for the entire epoch of 1,400 ms for each condition for posterior regions in the P600 time window. Visual data inspection revealed a P1/N2 complex typically linked to early visual processing (X. Cheng, Schafer & Akyürek, 2010; Eulitz, Hauk & Cohen, 2000; Misra, Guo, Bobb & Kroll, 2012; Schendan & Kutas, 2003).

Figure 2.5: *Mean voltage amplitudes for each condition averaged across segments, participants (n = 28) and channels (CPz, CP3, CP4, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1, O2); the P600 time window of interest (500 ms – 900 ms) is highlighted in grey.*



2.3.7 EEG data results

The model of best fit yielded a main effect for *violation type*, as well as by-*subject* random slopes for violation type (Appendix 2.I). *Item* was included as a random effect to capture by-*item* variance. Further, *hemisphere* and *LexTALE-Esp score* were included as covariates. *Gender congruency* and *cognate status* were included in the model but did not show an effect on voltage amplitudes with β = 0.223, *95% CI*[-0.020, 0.466], t = 1.80, p = 0.072 for *congruent* vs. *incongruent* nouns and β = -0.012, *95% CI*[-0.255, 0.231], t = -0.095, p = 0.924 for *cognates* vs. *non-cognates*. *Order of acquisition of Spanish, terminal phoneme, Stroop effect* and *target noun gender* led to non-convergence and were therefore not included. In line with our predictions, voltage amplitudes were significantly higher for *violation* trials compared to *non-violation* trials (β =

0.951, *95% CI*[0.528, 1.37], $t = 4.41$, $p < 0.001$). This reflected a robust P600 effect across all conditions. The model of best fit was: voltage amplitudes ~ violation type + gender congruency + cognate status + hemisphere + LexTALE-Esp score + (violation type|subject) + (1|item).

In a second step, we examined whether P600 effect sizes (voltage amplitudes for non-violation trials subtracted from violation trials) varied as a function of *gender congruency* and *cognate status* more closely. We used an LMM approach to determine effect size variation, averaged across markers from the four conditions. The P600 effect sizes for each condition were $M = 1.11$ ($SD = 2.96$) for *congruent non-cognate* trials, followed by $M = 1.08$ ($SD = 2.90$) for *congruent cognate* trials, $M = 0.937$ ($SD = 2.84$) for *incongruent non-cognate* trials and $M = 0.751$ ($SD = 2.78$) for *incongruent cognate* trials. The model of best fit included an interaction effect for *gender congruency* and *cognate status*. Critically, this interaction did not have a significant effect on P600 effect size with $\beta = 0.134$, $t = 0.315$, $p = 0.755$, and neither did the main effects for *gender congruency* and *cognate status* with β = -0.282, t = -0.794, p = 0.434 and $\beta = 0.039$, $t = 0.157$, $p = 0.876$, respectively. *Hemisphere* and *LexTALE-Esp score* were included as covariates. By-*condition* and by-*hemisphere* random slopes for *subject* were also included in the model. *Order of acquisition* of Spanish and *Stroop effect* did not significantly improve the model fit. The model of best fit was the following: P600 effect size ~ Gender congruency * Cognate status + Hemisphere + LexTALE-Esp score + (Gender congruency * Cognate status|Subject) + (Hemisphere|Subject). In sum, we established a P600 effect and therefore sensitivity to syntactic irregularities for all conditions. However, our results did not demonstrate a modulation of the P600 effect size by *gender congruency* or *cognate status*. These results support our behavioural findings from the violation paradigm task regarding the P600 effect.

2.4 Discussion

The aims of this study were the following: first, to examine whether there was cross-linguistic interference (CLI) of the gender systems in German late learners of Spanish. Secondly, to explore an interaction between the *gender congruency effect* and the *cognate facilitation effect* on grammatical gender processing. Finally, to characterise low-proficient late language learners with low exposure to Spanish (< 3 years) in terms of inhibitory skills and CLI. This was to contribute to the conceptualization of CLI in the multilingual brain.

For the Stroop task, we first predicted a Stroop effect of condition as well as target language. In line with our hypotheses and previous research (Costa, Albareda & Santesteban, 2008; Goldfarb & Tzelgov, 2007), participants were consistently more accurate and faster in the congruent condition compared to the incongruent condition. More centrally, we also studied the association between inhibitory skills and vocabulary size. For this, we correlated the size of the Stroop effect for each participant with the individual vocabulary scores obtained from the LexTALE-Esp task. Contrary to our predictions, we found no correlation between the two variables for neither the German nor the Spanish target words. In other words, we found no evidence that better inhibitory skills on the Stroop task (i.e., a smaller Stoop effect) were associated with a larger vocabulary in Spanish. This result is somewhat surprising given that previous research found such a relationship between proficiency and inhibitory skills (Marian, Blumenfeld, Mizrahi, Kania & Cordes, 2013). Previous research also proposed a relationship between intermediate to high proficiency and higher LexTALE scores (Lemhöfer & Broersma, 2012). On the other hand, research on inhibitory skills and vocabulary size on low-proficient late learners with limited exposure is scarce.

There are three possible interpretations of these findings. The first is concerned with the notion that LexTALE-Esp scores are as-

sociated with overall proficiency in proficient speakers (Lemhöfer & Broersma, 2012), but not in low-proficient late language learners. In other words, the LexTALE-Esp might not be suitable for inferences about grammatical knowledge, phonological awareness or syntactic knowledge for late learners with low exposure to the language. Second, the original LexTALE-Esp was tested on speakers with different L1s, the majority of which were English natives, the rest being speakers of French, German, Italian, Romanian, Portuguese and Polish. Therefore, the test was not exclusively validated for German as L1, but rather for a combined group of different L1s. Performance on the LexTALE-Esp might therefore be susceptible to L1 influences: speakers with an L1 typologically similar to Spanish (e.g., Italian and French) might have an inherent advantage compared to speakers of an L1 with a larger typological distance (e.g., German and Polish), e.g., for cognates (Lemhöfer & Dijkstra, 2004). This advantage in performing the LexTALE-Esp might be independent of their true vocabulary size in Spanish. Finally, inhibitory skills might be a suitable predictor of proficiency in highly proficient bilinguals, but not in late learners with low proficiency levels such as those in the current study. While several studies have established a positive association between inhibitory skills and proficiency in proficient learners (Lev-Ari & Peperkamp, 2013), this association might only emerge once participants have reached higher stages of overall proficiency beyond the B1/B2 levels of the current participants. In light of the current findings, it is problematic to argue for one of the three explanations. Taken together, these results are novel in that they warrant for a more fine-grained investigation of the relationship between vocabulary size scores and overall proficiency and the validity of the LexTALE-Esp for a range of different linguistic populations and proficiency levels. Further experiments with a more homogeneous group with an L1 that is more typologically similar to Spanish (e.g., Italian) are needed, while also examining the effects for different levels of proficiency.

For the behavioural data from the violation paradigm task, we replicated the well-established finding of higher accuracy rates and lower RTs in non-violation trials compared to violation trials in

low-proficient late language learners. Therefore, we added to existing research on high-proficient late language learners (Foucart & Frenck-Mestre, 2011; Lemhöfer et al., 2008). Furthermore, this is behavioural evidence supporting different processing mechanisms for NPs with syntactic violations compared to NPs without violations. More importantly, we found evidence for the *gender congruency effect* and therefore for CLI of grammatical gender systems: late learners with low proficiency were more accurate and faster at processing gender congruent nouns compared to incongruent nouns. This adds to existing similar findings in proficient bilinguals (Klassen, 2016) and also supports the previously discussed *gender-integrated representation hypothesis*, GIRH (Bordag & Pechmann, 2007; Costa et al., 2003; Lemhöfer et al., 2008; Morales et al., 2016).

As previously discussed, a large number of studies examining the *gender congruency effect* and the *cognate facilitation effect* have not systematically controlled for *gender congruency* as well as for *cognate status*: Moreover, studies rarely focused on late learners with low proficiency (Costa et al., 2003; Lemhöfer et al., 2008). Thus, it was unclear whether the processing advantage for cognates compared to non-cognates reported in these studies was driven by phonological and orthographic overlap (i.e., cognate status) or similarities at the grammatical level (i.e., an overlap in terms of gender) in late learners with low proficiency. Contrary to our predictions about the presence of an interaction between the *gender congruency effect* (i.e., faster processing of congruent nouns compared to incongruent nouns) and the *cognate facilitation effect* (i.e., faster processing of cognate compared to non-cognates), we found no effect of cognate status on accuracy rates. For RTs, we found an effect of cognate status in the opposite direction: participants appeared to be slower when making syntactic decisions in cognate trials compared to non-cognate trials. This is a crucial and novel finding. It speaks directly to the respective saliency of two inherent properties of lexical items stored within a bilingual system in late learners: at low proficiency levels and relatively limited exposure to Spanish, German late learners of Spanish were more sensitive to lexico-syntactic similarities at the gender level than to phonological

and orthographic overlap provided by cognates.

For the EEG data from the violation paradigm task, we employed a relatively recent and novel single-trial LMM approach (Frömer et al., 2018) in an attempt to move away from average-style approaches to a more suitable data analysis approach. We found clear evidence for a P600 effect across all conditions. We therefore confirmed the sensitivity of late learners to syntactic violations. However, we did not find evidence for an influence of gender congruency or cognate status on voltage amplitudes. This is indicative of two important aspects: first, that there was no detectable influence of these two noun properties at the electrophysiological level. Second, that we did not find evidence for distinct neuronal patterns associated with CLI of grammatical gender systems or the phonological systems. These are important findings: first, contrary to reports on absent P600 effects in late learners (Hahne & Friederici, 2001; Weber-Fox & Neville, 1996), we provide evidence that late language learners are indeed sensitive to syntactic violations even in early acquisition stages (S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005). Second, participants appeared to be relatively insensitive to both gender congruency and cognate status at early acquisition stages, with limited CLI traceability at the neural level. Importantly, our findings do not suggest an N400 effect, which has been linked to semantic integration processes in both native and non-native processing (Friederici et al., 1999; Molinaro et al., 2011; Münte et al., 1997). We neither find evidence for a biphasic N400/P300 (Barber & Carreiras, 2003, 2005) nor for an N400/P600 pattern (Martín-Loeches et al., 2006; Wicha et al., 2004) reported in native speakers of Spanish. In ERP terms, an N400 would have been reflected in more negative amplitudes for violation trials (incorrect gender value) compared to non-violation trials (correct gender value). Therefore, we concluded the following: first, the presentation of a bare noun prior to the experimental trial did not introduce a semantic component; second, we successfully minimised the semantic context for the syntactic violation identification task we employed; and finally, we reduced guessing strategies that could

be employed by participants[2].

Finally, in contrast to Foucart and Frenck-Mestre (2011), our results from the single-trial analysis on the P600 effect size across conditions did not yield variation as a function of condition. Despite a descriptive tendency for a larger P600 effect for congruent trials compared to incongruent trials, this difference was not significant. This implies that the P600 effect was statistically similar across conditions, which does not provide evidence for a modulation by gender congruency or cognate status. This is a crucial result because it supports the notion of the insensitivity of late language learners at the neural level to inherent properties of nouns such as gender congruency and cognate status, while processing NPs in Spanish under the influence of German.

Our results are highly relevant for three reasons. First, research on language processing mechanisms and the neuronal signatures of gender processing in late second language learners has been scarce, in particular the extent of CLI from the native language. Second, our results showed that late language learners face CLI at the grammatical level: in the case of overlapping syntactic properties across the languages (i.e., gender congruency), this can facilitate processing in the non-native language, but it hinders processing in the case of non-overlapping syntactic structures (e.g., gender-incongruency). These results therefore allow us to characterise the

[2]Given that previous studies overwhelmingly suggested a P600 effect for processing syntactic violations in native speakers (Barber & Carreiras, 2005; Hasting & Kotz, 2008; Osterhout & Mobley, 1995), we did not find it necessary to include a native Spanish control group. Moreover, studies suggested that N400 effects are limited to conditions with a strong semantic violation or semantic integration component (Osterhout & Mobley, 1995; Wicha et al., 2004). Therefore, if we were to repeat our study with native speakers of Spanish, consistent with our current predictions about non-native speakers and the syntactic nature of our task, we predict more positive P600 amplitudes for syntactic violations compared to non-violations. In contrast, given that the N400 effect is mostly elicited in connection to semantic factors, we do not predict an N400 effect in our specific case, neither for hypothetical native speakers nor for those non-native speakers we tested in the current study.

challenges encountered by late learners with low proficiency levels. They provide a basis for an increased focus on these challenges during foreign language teaching. Further, the results are central to characterising the brain mechanisms involved in processing grammatical gender in a foreign language. They are fundamental to broadening our understanding of processing a foreign language at limited proficiency levels. Notably, low-proficient speakers do show sensitivity to syntactic irregularities in their foreign language. The results from this study strongly encourage a wider focus on intermediate and low-proficient language learners in order to characterise the respective underlying processing mechanisms. The results also promote more differentiated testing designs and controlling for the inherent property of gender congruency when investigating the cognate facilitation effect, especially in population samples of late language learners.

2.4.1 Conclusions and future directions

CLI in late learners with low to moderate proficiency levels has not received enough attention in the field of multilingual language processing. Overall, our results support the notion of a *P600 effect* for determiner-noun gender agreement in *late learners* of Spanish with low proficiency levels. This was reflected in the ERP signatures of trials containing syntactic violations compared to trials which did not. Moreover, we present evidence for *cross-linguistic interference* of grammatical gender systems at the behavioural level in the form of the *gender congruency effect*. On the other hand, the underlying neuro-cognitive processes and P600 effect sizes appear unaffected by gender system similarities or overlapping phonological or orthographic forms. Contrary to our predictions, we did not find evidence for a joint effect of *gender congruency* and *cognate status* at the neuronal or behavioural level. Thus, it appears that late language learners are behaviourally more sensitive to similarities in terms of gender, compared to similarities at the phonological and orthographic level. Nevertheless, the results support the *gender-integrated representation hypothesis* (GIRH), even in late learners with relatively low proficiency. The results from this study contrib-

ute to the debate about the sensitivity of late language learners to syntactic violations and to inherent properties of nouns during non-native language processing. Therefore, this study opens up new avenues for the conceptualization of syntactic processing in language learners with limited language exposure as well as cross-linguistic interference in early acquisition stages.

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 want to thank Carsten Eulitz and Oleksiy Bobrov for providing the lab facilities at the University of Konstanz, and Charlotte Englert, Khrystyna Oliynyk, Zeynep Dogan, Fangming Zhang and Anna-Maria Waibel for their help with data collection and participant recruitment. We would also like to thank Theo Marinis and Maria del Carmen Parafita Couto for their feedback. Further, we thank Julian Karch for his input and support in the data analysis, and we are grateful to all of our participants. The statistical analyses were performed using the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. Finally, we thank our anonymous reviewers for their comments and feedback.

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/xvt6c/?view_only=24db45812a36490cacb33b7a298a71fc`

Citation diversity statement

Recent studies have highlighted an inherent bias in academic publishing in that women scientists and scientific from minorities are systematically under-cited in comparison to the papers published in the field (Dworkin et al., 2020; Zurn, Bassett & Rust, 2020). The Citation Diversity Statement serves the purpose of raising awareness of this bias from the perspective of gender. Several recently published studies have included such a statement (Rust & Mehrpour, 2020; Torres, Blevins, Bassett & Eliassi-Rad, 2020). With this statement we explored the gender balance in our reference list. On the basis of the preferred gender of the first and last author, we assigned each reference entry one of the following gender combinations: man/ man, woman/ woman, woman/ man, man/ woman. Our references consisted of 30.5% woman/ woman, 31.4%. man/ man, 22.9% woman/ man and finally, 12.4% man/ woman. Three references were not classified because they did not have first and last author. According to work by Dworkin et al. (2020), this compares to 6.7% for woman/ woman, 58.4% for man/ man, 25.5% woman/ man, and lastly, 9.4% for man/ woman authored references for the field of neuroscience. Note that there are limitations to this classification because it is based on a binary gender distinction, however, we are confident that future work will help us improve this classification system.

Appendix

2.A Linguistic profile: German-Spanish group

Table 2.A.1: *Overview of the languages acquired by the participants of the current study (N = 33) according to the LEAP-Q.*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
German	n = 33					**33**
Spanish			n = 16	n = 15	n = 2	**33**
English		n = 31	n = 2			**33**
Latin			n = 3	n = 1	n = 1	**5**
French		n = 2	n = 11	n = 5		**18**
Russian			n = 1		n = 1	**2**
Swedish				n = 1		**1**
Italian					n = 1	**1**
Arabic					n = 1	**1**
Catalan					n = 1	**1**
Mandarin					n = 1	**1**
Portuguese					n = 2	**2**
Total	**33**	**33**	**33**	**22**	**10**	

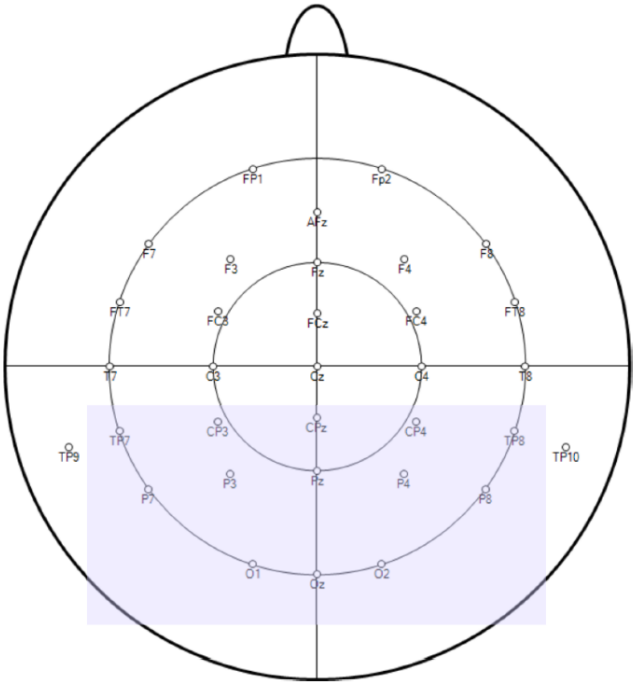
2.B Stimuli: Terminal phoneme distribution

Table 2.B.1: *Terminal morpheme frequencies for the violation paradigm task in this study in bold on the left side of the panel and terminal morpheme frequencies according to Clegg (2011) on the right side of the panel.*

Terminal morpheme	**Count**	**Percentage**	Terminal morpheme	Count	Percentage	Associated gender
o	73	32.59	o	685	30.95	M
a	76	33.93	a	716	32.35	F
e	21	9.38	e	167	7.55	M
r	5	2.23	r	75	3.39	M
l	5	2.23	l	47	2.12	M
z	4	1.79	z	21	0.95	F
s	1	0.45	s	24	1.08	M
ión	17	7.59	ión	301	13.60	F
n	12	5.36	n	45	2.03	F
d	7	3.13	d	121	5.47	F
umbre	1	0.45	umbre	4	0.18	F
Others	2	0.90	Others	7	0.31	M
Total:	**224**	**100**	Total:	2,213	100	

2.C EEG electrode montage

Figure 2.C.1: *Electrode positions following a 10/20 montage. Electrodes included in the analysis are in the shaded area.*



2.D Stroop model parameters: German targets

Table 2.D.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for German targets (n = 32).*

	Formula: accuracy ~ condition (congruent vs. incongruent) + (1|subject) + (1|item)			**Formula**: RTs ~ condition (congruent vs. incongruent) * LexTALE-Esp score + (1|subject) + (1|item)		
Term	**Odds ratio [95% CI]**	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	22.43 [15.48, 32.49]	16.45	< 0.001	600.95 [573.32, 628.58]	42.64	< 0.001
Condition [incongruent]	0.659 [0.437, 0.994]	-1.99	**0.047**	13.08 [2.48, 23.68]	2.42	**0.016**
LexTALE-Esp score				0.100 [-0.937, 1.14]	0.189	0.850
Condition * LexTALE-Esp score				-0.014 [-0.410, 0.382]	-0.069	0.945
Random effects						
σ^2	3.29			12,183.47		
$\tau_{00Subject}$	0.31			3,439.83		
τ_{00Item}	0.02					
ICC	0.09			0.22		
$N_{Subject}$	32			32		
N_{Item}	4			4		
Observations	3,072			2,887		
Marg. R^2/ Cond. R^2	0.012/0.101			0.003/0.222		

2.E Stroop model parameters: Spanish targets

Table 2.E.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for Spanish targets (n = 32).*

	Formula: accuracy ~ condition (congruent vs. incongruent) + (1|subject)			**Formula**: RTs ~ condition (congruent vs. incongruent) + (1|subject)		
Term	**Odds ratio [95% CI]**	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	28.36 [19.45, 41.37]	17.37	< 0.001	555.28 [537.57, 572.99]	61.47	< 0.001
Condition [incongruent]	0.676 [0.491, 0.930]	-2.41	**0.016**	20.55 [13.44, 27.65]	5.67	< **0.001**
Random effects						
σ^2	3.29			9535.55		
$\tau_{00Subject}$	0.57			2403.60		
ICC	0.15			0.20		
$N_{Subject}$	32			32		
Observations	3,072			2,909		
Marg. R^2/ Cond. R^2	0.010/0.156			0.009/0.208		

2.F Model parameters: accuracy

Table 2.F.1: *Model parameters for best-fitting model for accuracy (N = 33).*

Formula: accuracy ∼ violation type (non-violation vs. violation) + gender congruency (congruent vs. incongruent) + LexTALE-Esp score + (1|subject) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	66.35 [37.70, 116.79]	14.54	< 0.001
Violation type [violation]	0.389 [0.235, 0.644]	-3.68	$<$ **0.001**
Gender congruency [incongruent]	0.539 [0.326, 0.892]	-2.40	**0.016**
LexTALE-Esp score	1.03 [1.02, 1.04]	4.58	< 0.001
Random effects			
σ^2	3.29		
τ_{00Item}	2.03		
$\tau_{00Subject}$	0.29		
ICC	0.41		
$N_{Subject}$	33		
N_{Item}	224		
Observations	5,977		
Marginal R^2/ Conditional R^2	0.099/0.471		

2.G Model parameters: response times

Table 2.G.1: *Model parameters for best-fitting model for response times (N = 33).*

Formula: RTs ~ violation type (non-violation vs. violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + terminal phoneme + (1|subject) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	788.42 [772.77, 804.07]	98.75	< 0.001
Violation type [violation]	116.12 [102.68, 129.56]	16.94	< **0.001**
Gender congruency [incongruent]	34.54 [20.53, 48.55]	4.83	< **0.001**
Cognate status [non-cognate]	-19.75 [-32.70, -6.80]	-2.99	**0.003**
Terminal phoneme [d]	-22.08 [-33.10, -11.06]	-3.93	< 0.001
Terminal phoneme [e]	30.48 [18.52, 42.43]	4.99	< 0.001
Terminal phoneme [i]	-69.60 [-83.10, -56.11]	-10.113	< 0.001
Terminal phoneme [ión]	36.28 [25.85, 46.71]	6.82	< 0.001
Terminal phoneme [j]	-14.40 [-32.34, 3.54]	-1.57	0.116
Terminal phoneme [l]	22.14 [10.04, 34.25]	3.59	< 0.001
Terminal phoneme [n]	3.92 [-6.13, 13.98]	0.765	0.445
Terminal phoneme [o]	1.18 [-10.72, 13.08]	0.194	0.846
Terminal phoneme [r]	84.99 [73.43, 96.55]	14.408	< 0.001
Terminal phoneme [s]	326.12 [314.06, 338.17]	53.03	< 0.001
Terminal phoneme [umbre]	45.43 [30.05, 60.78]	5.80	< 0.001
Terminal phoneme [z]	-36.74 [-46.01, -27.46]	-7.77	< 0.001

Random effects	
σ^2	0.14
τ_{00Item}	3104.48
$\tau_{00Subject}$	9014.72
ICC	1.00
$N_{Subject}$	33
N_{Item}	224
Observations	5,636
Marginal R^2/ Conditional R^2	0.262/1.00

2.H EEG data: by-condition trial rejection rates

Table 2.H.1: *Rejection rates for each condition for the EEG data of the violation paradigm task (n = 28).*

Condition	**Rejection rate (%)**	**Rejected trials**
cognate/congruent/non-violation	3.88	28
cognate/congruent/violation	6.18	40
cognate/incongruent/non-violation	6.12	42
cognate/incongruent/violation	6.12	42
non-cognate/congruent/non-violation	3.90	27
non-cognate/congruent/violation	5.19	33
non-cognate/incongruent/non-violation	3.93	28
non-cognate/incongruent/violation	4.39	31
Average	**4.94**	**33.88**

2.I Model parameters: P600 component

Table 2.I.1: *Model parameters for best-fitting model for voltage amplitudes (n = 28).*

Formula: voltage amplitudes ~ violation type (non-violation vs. violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + hemisphere + LexTALE-Esp score + (violation type|subject) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	0.780 [0.089, 1.47]	2.21	0.027
Violation type [violation]	0.951 [0.528, 1.37]	4.41	< **0.001**
Gender congruency [incongruent]	0.223 [-0.020, 0.466]	1.80	0.072
Cognate status [non-cognate]	-0.012 [-0.255, 0.231]	-0.095	0.924
LexTALE-Esp score	-0.025 [-0.048, -0.001]	-2.08	0.037
Hemisphere [midline]	1.78 [1.77, 1.79]	294.48	< 0.001
Hemisphere [right]	0.931 [0.921, 0.942]	177.99	< 0.001
Random effects			
σ^2	65.67		
τ_{00Item}	0.86		
$\tau_{00Subject}$	1.64		
$\tau_{11Subject[violation]}$	0.87		
$\rho_{01Subject}$	0.17		
ICC	0.05		
$N_{Subject}$	28		
N_{Item}	224		
Observations	12,469,236		
Marginal R^2/ Conditional R^2	0.014/0.058		