From oscillations to language

Behavioural and electroencephalographic studies on cross-language interactions

Published by

LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: The Power of Waves, by Yiru Zhan Wang

ISBN: 978-94-6093-419-3
DOI: https://dx.medra.org/10.48273/LOT0633
NUR: 616

From oscillations to language

Behavioural and electroencephalographic studies on cross-language interactions

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus Prof. dr. ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 17 januari 2023
klokke 15.00 uur

door

Sarah Von Grebmer Zu Wolfsthurn

geboren 23 oktober 1994
te Bruneck/Brunico, Italië

Promotor:	Prof. dr. N. O. Schiller
Co-promotor:	Dr. L. Pablos-Robles
Promotiecommissie:	Prof. dr. A. Foucart (Universidad Nebrija)
	Prof. dr. C. C. Levelt
	Prof. dr. T. Marinis (Universität Konstanz)
	Dr. J. Witteman

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.

This thesis is dedicated to my parents
and to my favourite sister. Wewiiisch.

Contents

Acknowledgements

This thesis would have never been finished without the guidance from my supervisors, Prof. dr. Niels O. Schiller and Dr. Leticia Pablos-Robles. Niels, you have provided insightful advice throughout the years, and were always ready to share feedback. Leticia, you got me "unstuck" more times than I can count - thank you for your patience and for being an endless source of inspiration. Thank you, Prof. dr. Theo Marinis, Prof. dr. Carsten Eulitz and Prof. dr. Núria Sébastián-Gallés and your respective labs for providing me with the facilities for the data collection. Thank you also, Prof. dr. Marco Calabria, for supervising me during my secondment on the application of clinical neuropsychology in multilingual research - it has been a fantastic experience.

To Kaja Gregorc and Muna Schönhuber, thank you for keeping this project running so smoothly over the years. Thanks to Elisabeth Süß (gym-buddies!), Angeliki Golegos and Ege Ekin Özer for your support in getting settled - you have made my research stays so much brighter. A huge thanks to all of my fellow MultiMinders, Grazia, Sofia, Jia'en, Theresa, Solange, Maren, Sergio, David, Isabel, Michal, Daniela, Mathilde, Jasmijn, Lari, Konstantina and Juhayna - for all the adventures on Greek islands, Italian lakes, in English cities and German wineries, and for the memes - I could have not asked for a more awesome group of people to be on this journey with. To Michal, who put up with my countless calls and always knew what to say - thank you. To all of my student assistants - Charlotte, Khrystyna, Zeynep, Fang-Ming, Despina, Anna, Philipp and Ife - heroes! To the almost two-hundred multilingual participants - I know that for some of you the lab experience was

nerve-racking, but you trusted me to do my job - thank you. This thesis would not exist without your magnificent brain waves.

A big thanks to my fellow lab colleagues, in particular Evelyn Bosma, Lisette Jager, Tanja Westra, Shaoyu Wang, Ruixue Wu, Jin Wang, Jiaqi Wang and Aranka van Tol - thank you for the thorough lab introduction, for the writing groups, the (more or less) scientific discussions and for letting me go on and on about some lab procedure. To my other dear colleagues: Rasmus Puggaard-Rode, Andrew Wigman, Laura Smorenburg, Ami Okabe and Astrid van Alem - thanks for the crash courses in linguistics and for all the dinners and drinks. A big shout-out to Anneloes van Iwaarden, who always helped me to find my feet. To Katja Lubina - thank you for having our backs (in the present and the future). To Julian Karch and Marjolein Fokkema for their statistical consults - I have learnt so much from you. To Jos Pacilly and to Andreea Geambaşu, for being the best lab technician and lab manager - thank you! A special thanks goes to Lis Kerr and Xander Vertegaal (xls) for the endless chats, dinners, attempts at running marathons, LaTeX advice, cat pictures and probably hundreds of cups of tea - thank you for being in my life.

To Solène Hémery, (Lulu) Yiru Zhan Wang (front cover credits!) and Judith Kaiser - thank you for all the get-away weekends and the brunches while reminiscing about our previous Bristol lives. A big thanks to Anne Fleur van Luenen and Niek Mereu - you are an inspiration for life choices and travel adventures. To Lilli Settari and Ramona Ilmer for your friendship, love and support throughout all these years - having you by my side means everything. “Und lous!” sogatn mir iaz.

Most of all, to my parents, Daniela and Hubert - you taught me patience and persistence, and you always had a piece of advice when the world seemed to be too confusing - this one is for you, Mammina und Mössjö. To my sister, Lisa - you taught me how to be brave and strong, and how to do the right thing instead of the easy one. Ohana. Finally, my deepest gratitude goes towards my superhero partner, Édouard (Bom) Bonneville. Thank you for numerous statistical consults, for coping with my dreadful singing, cheese obsession and all the existential crises, for the debates about changing the world (or becoming professional quiche-makers), the after-lunch-coffees and tea refills, the pep-talks in moments where I thought I could never finish this thing, for your endless patience and your love - I could not have done this without you.

Articles in this thesis

This thesis consists of a series of experiments and their corresponding research articles. Below you can find a list of the articles which were previously accepted for scientific publication or are currently under review. Note that the articles are in ascending order based on the chapter number.

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. [Chapter 2]

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (2021). Noun-phrase production as a window to language selection: An ERP study. *Neuropsychologia, 162*, 108055. [Chapter 3]

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (under review). Processing syntactic violations in the non-native language: different ERP effects typologically similar languages. *Neuropsychologia.* [Chapter 5]

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (2022). Does your native language matter? Neural correlates of typological similarity in non-native production. *Lingue e Linguaggio, 21*(1), 143-169. [Chapter 6]

Von Grebmer Zu Wolfsthurn, S., Gupta, A., Pablos-Robles, L., & Schiller, N. O. (2022). When left is right: the role of typological similarity in multilinguals' inhibitory control performance. *Bilingualism: Language and Cognition, 25*(5), 1-14. [Chapter 7]

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (under review). Neural correlates of gender agreement processing in Spanish: P600 or N400?. *Brain and Language* [Chapter 8]

CHAPTER 1

General introduction

1.1 Background

How does the brain manage multiple languages? What are the underlying behavioural and neural mechanisms of the brain trying to make sense of the world from input that is not in the native language (L1), or to produce output in a non-native language? Anecdotally, when *late language learners* are asked about their strategy to name a particular object in a non-native language, they will often state that they rely on the similarities between their L1 and the non-native language in question. In this thesis, we defined late language learners as those individuals who have acquired a language in addition to their L1 after the age of fourteen (S. Rossi, Gugler, Friederici & Hahne, 2006). We placed a special focus on late language learners due to an noticeable lack of research combining behavioural and neurolinguistic methods when examining this particular population. Moreover, we broadly defined individuals who are able to actively engage and communicate in more than one language as *multilinguals* (Kroll, Dussias, Bice & Perrotti, 2015). For those individuals to be considered multilingual, proficiency levels can vary in terms of their comprehension and speaking proficiency but also within their L1 and any additional languages.

Current evidence overwhelmingly suggests that multilingualism has profound functional and structural consequences for the mind and brain and that it impacts both language-specific mechanisms and more domain-general cognitive mechanisms (Abutalebi, Cappa & Perani, 2001; Bialystok, Craik & Luk, 2012; Kroll et al., 2015; Stein et al., 2012; Wong, Yin & O'Brien, 2016). For example, previous research suggested a differential use of domain-general cognitive control networks for multilinguals compared to monolinguals (Abutalebi et al., 2013; Bialystok et al., 2012; D. W. Green & Abutalebi, 2013). Cognitive control refers to the ability to adjust to a preferred outcome, for example suppressing irrelevant information in a particular context (Abutalebi & Green, 2016; Declerck et al., 2021; D. W. Green, 1998). Multilinguals were consistently found to outperform monolinguals on tasks which required the inhibition of irrelevant information, the switching of attention and management of working memory load (Hilchey & Klein, 2011). Critically, in this thesis we move away from the traditional comparison between multilinguals and monolinguals. Instead, we focus on multilingualism as a gateway to the underlying processing mechanisms of the brain with respect to non-native language comprehension and production. In light of this, we aim to add novel evidence to the current literature (Abutalebi et al., 2001; Berthele, 2021a; Kroll & Bialystok, 2013; Kroll et al., 2015; Puig-Mayenco, González Alonso & Rothman, 2020).

A central theme of this thesis is cross-linguistic influence, or *CLI*, which encompasses the bidirectional influence on processing mechanisms of the L1 and non-native language at the cognitive and neural level (Lago, Mosca & Garcia, 2021; Nozari & Pinet, 2020). Weinreich (1953) was one of the first to describe CLI as a key feature in the context of multilingual language processing (Kroll et al., 2015). CLI speaks directly to the anecdotal accounts mentioned before, whereby learners describe that they form "connections" between their L1 and a non-native language. Given its potentially critical role in non-native language processing, CLI has since featured in several studies which made fundamental contributions to the characterisation of the multilingual language architecture (Clahsen & Felser, 2006a; Hartsuiker, Pickering & Veltkamp, 2004; Jarvis, 2011; Odlin, 1989; B. D. Schwartz & Sprouse, 1996). CLI has its roots in the parallel activation of the languages within the multilingual brain, also in cases where only one language is required in a specific context (Blumenfeld & Marian, 2007; Costa & Pickering, 2019; Guo & Peng, 2006; Kroll et al., 2015; Marian & Spivey, 2003a; Mishra & Singh, 2016). Yet, CLI and its impact on non-native comprehension and production has been scarcely researched in late language learners with lower proficiency levels in comparison to highly proficient bilinguals (Bordag & Pechmann, 2007; Bürki & Laganaro, 2014; Costa, Kovacic, Franck & Caramazza, 2003; Lemhöfer, Spalek & Schriefers, 2008; Morales et al., 2016). Therefore, in this thesis we thoroughly examined two CLI effects in late language learners in more detail, namely the *gender congruency effect*, and the *cognate facilitation effect*.

The gender congruency effect refers to the notion of more efficient processing of gender congruent nouns (i.e., nouns with similar gender categories across languages) vs. incongruent nouns (i.e., nouns with dissimilar gender categories across languages) as a result of the interaction of the grammatical gender systems (Morales et al., 2016; Paolieri, Demestre, Guasch, Bajo & Ferré, 2020). Note that we exclusively focused on grammatical gender as opposed to semantic gender, which is based on the biological features of the noun (Corbett, 1991). Our main language of interest in this thesis was Spanish, which is characterised by a two-way gender system. The definite determiner "la" marks the feminine gender value, and "el" marks masculine gender value, e.g., la_F $mesa_F$ [the table] and el_M $perro_M$ [the dog], respectively. Contrastingly, the cognate facilitation effect refers to the processing advantage of cognates (i.e., words with considerable orthographic and phonological overlap) compared to non-cognates, for example *el volcán - the volcano* vs. *el perro - the dog* in Spanish and English (Amengual, 2012; Bosma, Blom, Hoekstra & Versloot, 2019; Costa, Caramazza & Sebastián-Gallés, 2000; Dijkstra, Miwa, Brummelhuis, Sappelli & Baayen, 2010; Lemhöfer

& Dijkstra, 2004; Strijkers, Costa & Thierry, 2010). As mentioned above, these two CLI effects are robustly reported for highly proficient multilinguals (Bordag & Pechmann, 2007; Lemhöfer et al., 2008; Morales et al., 2016), but have been less extensively studied in late language learners. Therefore, by focusing on late language learners, we obtained critical insights from CLI into multilingual language processing in a relatively understudied population. Among those insights were the quantification of CLI in non-native comprehension and production, the sensitivity of the multilingual brain to the corresponding linguistic features of gender congruency and cognate status, the neural correlates of CLI in late language learners, the characterisation of individual non-native production stages as well as the locus of target language selection. These issues are systematically featured in this thesis across a number of studies, as discussed in more detail in the following sections.

The second central theme of this thesis is whether and how *language similarity* between the L1 and the non-native language affects non-native comprehension and production mechanisms. Defining and measuring language similarity has been subject to debate in the literature (Ringbom & Jarvis, 2009; Van der Slik, 2010). Previous work attempted to quantify language similarity by means of genetic linguistic distance (Cavalli-Sforza, Menozzi & Piazza, 1994), or cognate linguistic distance (McMahon & McMahon, 2005; Van der Slik, 2010). While we acknowledge that this debate around how to exactly measure the similarity between two languages is far from resolved, in this thesis we defined language similarity as the structural morphosyntactic, orthographic and phonological word form overlap across languages (Foote, 2009; Rothman & Cabrelli Amaro, 2010). Note also that throughout this thesis, we will use the term language similarity interchangeably with the terms linguistic similarity and typological similarity. Following this definition, we tested three different multilingual populations with varying degrees of language similarity: native German late learners of Spanish and native Dutch late learners of Spanish, representing the linguistically more dissimilar language pairs; and native Italian late learners of Spanish, representing the linguistically more similar language pair.

Beyond the aforementioned anecdotal accounts, research suggested that learners routinely embed the non-native language in their L1 context during language processing (Ringbom & Jarvis, 2009). Yet, language similarity was only featured in a few studies on non-native language comprehension and production (Dijkstra et al., 2010; Foucart & Frenck-Mestre, 2011; Sabourin, Stowe & De Haan, 2006; Rothman & Cabrelli Amaro, 2010; Zawiszewski, Gutiérrez, Fernández & Laka, 2011; Zawiszewski & Laka, 2020). Evidence from these studies tentatively suggested a processing advantage for speakers of linguistically similar languages, e.g., Italian-Spanish, compared to speakers of linguistically less similar languages, e.g., German-Spanish (Sabourin et al., 2006; Zawiszewski & Laka, 2020). The debate around the role of language similarity and its impact on multilingual processing is critical because it taps directly into the co-existence of two or more languages in a multilingual brain and the subsequent functional and cognitive architecture of each of these languages (Abutalebi, 2008; Clahsen & Felser, 2006b; Tolentino & Tokowicz, 2011). In light of the apparent lack of research, characterising the exact impact of language similarity on language processing mechanisms is a crucial issue. Therefore, in this thesis we quantified the role language similarity plays in multilingual language processing to gain insights into how language similarity guides non-native processing. More specifically, we tested the influence of high or low language similarity on non-native comprehension and production in Italian-Spanish speakers and German-Spanish speakers across several studies. The critical question here was whether or not speakers of highly similar languages, e.g., Italian-Spanish, had an inherent processing advantage compared to speakers of less similar languages, e.g., German-Spanish.

Additionally, we explored whether language similarity affected higher cognitive functioning such as cognitive control in Italian-Spanish speakers and in Dutch-Spanish speakers. Here, a fundamental issue is whether speakers of linguistically similar languages, e.g., Italian-Spanish, develop enhanced cognitive control skills over time compared to speakers of linguistically less similar languages, e.g., Dutch-Spanish (Stocco, Yamasaki, Natalenko & Prat, 2014;

Yamasaki, Stocco & Prat, 2018). Taken together, we investigated the role of language similarity both in the context of non-native comprehension and production, but also in the context of more general cognitive functions such as cognitive control.

1.2 Non-native processing through the lens of electroencephalography

In this thesis, we combined behavioural measures of performance accuracy, response times, naming accuracy and naming latencies with electrophysiological measures, i.e., electroencephalography (EEG). EEG is a non-invasive method to measure event-related brain potentials (ERPs). These ERPs are time-locked to particular events, e.g., the appearance of a stimulus during a task (Woodman, 2010). More specifically, ERPs reflect the systematic oscillatory changes of brain potentials over time. These systematic changes are termed ERP components and can be linked to particular linguistic and cognitive phenomena, for example syntactic or semantic processing, or cognitive control (Friederici, Steinhauer & Frisch, 1999; Steinhauer, White & Drury, 2009; Swaab, Ledoux, Camblin & Boudewyn, 2011). ERPs are typically measured via scalp electrodes in the form of voltage amplitudes. Due to their excellent temporal resolution, ERPs can provide detailed and nuanced insights into the sub-mechanisms of language processing as well as domain-general cognitive mechanism in real time (Bürki & Laganaro, 2014; Christoffels, Firk & Schiller, 2007; Foucart & Frenck-Mestre, 2011; Hahne & Friederici, 2001; E. M. Moreno, Rodríguez-Fornells & Laine, 2008; S. Rossi et al., 2006; Steinhauer et al., 2009; Tokowicz & MacWhinney, 2005; Valente, Bürki & Laganaro, 2014; Woodman, 2010). Therefore, EEG and ERPs are fundamental tools in examining non-native language comprehension and production mechanisms. In this thesis, we focused mainly on the exploration of two ERP components: the P600 component in (non-)native comprehension in Chapters 2, 4, 5 and 8; and the P300 component in non-native production in Chapters 3, 4 and 6. However, we also briefly touched on the left anterior negativity (LAN) component, and the N400 component in Chapters 2, 3 and 8; both of which were commonly linked to language comprehension mechanisms (Barber & Carreiras, 2005; Kutas & Federmeier, 2011; Martín-Loeches, Muñoz, Casado, Melcón & Fernández-Frías, 2005; Swaab et al., 2011).

The P600 component is typically elicited in the context of syntactic violation paradigms (Hagoort, Brown & Groothusen, 1993). It has a maximal peak over centro-parietal regions around 600 ms post-stimulus onset and a component latency between 500 ms to 900 ms post-stimulus onset (Steinhauer et al., 2009; Swaab et al., 2011). Importantly, P600 component amplitudes vary with respect to whether a linguistic structure contains a syntactic violation: evidence suggests larger voltage amplitudes for syntactically incorrect compared to syntactically correct constructions, yielding the so-called *P600 effect* (Swaab et al., 2011). The P600 effect was robustly reported across a range of language combinations and different types of syntactic violations, for example determiner-noun agreement violations or adjective-noun agreement violations (Barber & Carreiras, 2005; Friederici et al., 1999; Gunter, Friederici & Schriefers, 2000; Hagoort & Brown, 2000; Molinaro, Vespignani & Job, 2008; Osterhout & Holcomb, 1992; Steinhauer et al., 2009; Swaab et al., 2011). The consensus is that the P600 component is a neural correlate of successful syntactic integration, re-analysis and repair (Alemán Bañón, Fiorentino & Gabriele, 2012; Barber & Carreiras, 2005; Gouvea, Phillips, Kazanina & Poeppel, 2010). Relevant to this thesis, the P600 effect was less robustly reported for late language learners with lower proficiency levels (Hahne, 2001; Hahne & Friederici, 2001; Weber-Fox & Neville, 1996), but see S. Rossi et al. (2006). This tentatively suggested that late language learners showed decreased sensitivity to syntactic violations. Therefore, in this thesis we used the P600 component to first, shine light onto the underlying neural mechanisms of (non-)native comprehension. Second, we examined how factors such as language similarity or CLI could impact P600 component voltage amplitudes. Third, we studied whether late language learners demonstrated sensitivity to syntactic violations. Finally, we examined whether P600 component amplitudes differed as a function of language similarity. These issues were closely investigated in Chapters 2, 4, 5 and 8.

The second ERP component we focused on in this thesis was the P300 component. This component is frequently reported in centro-parietal regions with a peak around 300 ms after stimulus onset and was more generally associated with cognitive control, working memory load and inhibitory processes (Barker & Bialystok, 2019; Barry et al., 2020; González Alonso et al., 2020; Polich, 2007). The so-called *P300 effect* refers to higher voltage amplitudes elicited in the more cognitively demanding condition, for example the incongruent condition in a classical Flanker task (Eriksen & Eriksen, 1974). This particular task exploits the mismatch between a target stimulus and the surrounding stimuli, e.g., a target arrow in the centre of an array where the remaining arrows point in a different direction to the target arrow (Bosma & Pablos, 2020). More recently, the P300 component was also reported in designs combining both cognitive control and language processing components, see for example Bosma and Pablos (2020) for a Flanker task with intermittent code-switching. Yet, to this date not much research has exploited the potential of the P300 component to study key features such as CLI in non-native language production. Even fewer studies have tackled the time course of non-native production with respect to CLI (Bürki & Laganaro, 2014; Valente et al., 2014). Therefore, in this thesis we used the P300 component to explore how CLI modulated the time course of non-native production. We used the Levelt-Roelofs-Meyer (LRM) model (Levelt, Roelofs & Meyer, 1999) as our theoretical framework. This model describes word production in terms of several individual production stages in the L1, each connected to a specific duration in time (Indefrey & Levelt, 2004; Indefrey, 2011). In this thesis, we utilised the gender congruency effect and the cognate facilitation effect to examine two different theoretical accounts of the mitigation of CLI and the selection of the target language. One account suggested that CLI was resolved prior to the so-called lexical retrieval stage (Gollan, Montoya, Fennema-Notestine & Morris, 2005; Hermans, Bongaerts, De Bot & Schreuder, 1998), whereas a second account suggested that CLI was only resolved after lexical retrieval (Christoffels et al., 2007; Colomé, 2001; Rodriguez-Fornells et al., 2005). Therefore, a novel aspect of this thesis was the examination of the P300 component as a potential index of mechanisms critical to successful non-native language production, namely the mitigation of CLI between the L1 and the non-native language and target language selection. Moreover, we explored whether P300 component amplitudes differed across various populations with a higher or lower language similarity. These issues are examined in detail in Chapters 3, 4 and 6 of this thesis. Taken together, combined with adequate statistical tools, EEG and ERPs form powerful pillars in the quest of disentangling the complex processing architecture within the multilingual brain.

1.3 Open Science and statistics in neurolinguistics

In recent years, there have been growing concerns about the robustness, reproducibility and interpretation of published research findings within the scientific community (Sönning & Werner, 2021). These concerns marked the beginnings of the so-called *replication crisis* in fields such as psycholinguistics or neurolinguistics, but also in linguistics, psychology and neuroscience (Ioannidis, 2005; Shrout & Rodgers, 2018; Sönning & Werner, 2021; Tackett, Brandes, King & Markon, 2019). Sönning and Werner (2021) outlined four problematic aspects of published research in an attempt to capture these concerns: the high rates of false-positive results connected to the (mis)use of statistical techniques, the opaqueness in terms of the methodological details and the analysis procedures, the inaccessibility of the original data, and finally, the perceived insignificance of replication studies and their subsequent classification as "non-original" research. Several studies have made efforts to define the underlying causes of these concerns, which include the incorrect interpretation of statistical outcome parameters, sub-optimal research designs and low statistical power (Ioannidis, 2005; Munafò et al., 2017). On the other hand, research has also focused on defining mitigatory strategies of the replication crisis, such as making data and research materials freely accessible and elaborating on the selection of statistical analysis procedures (Sönning & Werner, 2021). This movement is now widely known as Open Science, see Mirowski (2018) and Vicente-Saez and Martinez-Fuentes (2018) for a detailed discussion.

In the current thesis, we continued these efforts and placed a strong focus on the following: first, we increased the transparency regarding the linguistic background of our multilingual participants, the research method and the analysis procedures. More specifically, in addition to our strong theoretical focus, we provided detailed descriptions of our participants, research design, stimuli and materials. Moreover, we meticulously described and motivated our statistical analysis approaches and reported the detailed model specifications and parameters. Second, we made the data and analysis scripts openly available to the scientific community by creating public repositories on the Open Science Framework, OSF (Foster & Deardorff, 2017). This was done in an effort to encourage replication studies and to be transparent about the data and the analysis details. Third, we went beyond traditional statistical analysis approaches, especially for our highly complex EEG data. This meant that for some of our data, we expanded on more conventional analysis approaches and applied relatively novel, but more suitable statistical techniques. This was done to challenge current analysis practises and to propose concrete alternative approaches. To provide tangible examples, we used more advanced statistical methods to analyse both our behavioural data and our EEG data on a single-trial basis: (generalised) linear mixed effects models (LMMs), see Frömer, Maier and Abdel Rahman (2018); and generalised additive mixed models (GAMMs), see De Cat, Klepousniotou and Baayen (2015), Meulman, Wieling, Sprenger, Stowe and Schmid (2015) and Tremblay and Newman (2015).

An important characteristic of LMMs is the ability to effectively estimate random variance, and to manage missing data and unbalanced datasets (Baayen, Davidson & Bates, 2008; Fröber, Stürmer, Frömer & Dreisbach, 2017). As a result, LMMs yield robust estimates while minimising false-positives and maintaining high statistical power (Barr, 2013; Matuschek, Kliegl, Vasishth, Baayen & Bates, 2017). At the centre of this approach is a linear mixed model, which consists of a fixed effects structure and a random effects structure. The fixed effects structure typically contains the intercept or grand mean across all predictors, the linear predictors and the corresponding interaction effects, if applicable, as well as any potential covariates in line with the experimental design, research question and hypotheses. In contrast, the random effects structure contains terms to capture by-participant and by-item variance, and to estimate the effect of the main predictor variables as a function of the individual subject and item. LMMs are suitable for behavioural analyses, but are also a viable alternative to the more conventional ANOVA-approach for EEG data analysis (Frömer et al., 2018). LMMs for behavioural analyses are featured in all chapters of this thesis, and LMMs for EEG data analyses are included in Chapters 2, 3, 4 and 6 of this thesis.

On the other hand, the generalised additive mixed models, or *GAMMs*, represent an extension of LMMs: in addition to linear predictors, GAMMs also include non-linear predictors which do not assume a linear relationship between the predictor and the outcome variable (Meulman et al., 2015; Tremblay & Newman, 2015). These non-linear terms are fundamentally a collection of so-called "basis functions", which for instance capture model voltage amplitudes over time (De Cat et al., 2015; Meulman et al., 2015). This is a particularly critical feature of GAMMs, making them optimal candidates for analysing the complex and multi-dimensional oscillatory trends of EEG data. From the researcher's perspective, this approach avoids the a priori specification of a time window of interest for a given effect. Instead, it provides a precise estimate for the time window of this effect. Critically, GAMMs also include random intercepts and random slopes to capture random variance, by-participant and by-item effects, which is similar to LMMs. Therefore, GAMMs are a powerful approach to analysing EEG data compared to more conventional methods, in particular when examining populations that are less frequently featured in the current literature. In this thesis, we used GAMMs to flexibly model our EEG data in Chapters 5 and 8.

1.4 The current thesis

As previously outlined, the critical issues of this thesis were the following: first, to examine and quantify cross-linguistic influence (CLI) in both non-native comprehension and production. Second, to characterise the corresponding neural correlates of CLI in the multilingual brain. Third, to explore the sensitivity of late language learners to syntactic irregularities. Fourth, to examine the individual non-native production stages and the selection of the target language in light of CLI. Fifth, to investigate the overarching impact of language similarity between the native language and the non-native language on non-native comprehension and production. Finally, to study the impact of language similarity on cognitive control. Subsequently, our core research questions were concerned with the characterisation of the impact of both CLI and language similarity on the underlying processing mechanisms in late language learners, both from a behavioural and from a neural perspective. These issues were explored across a series of experiments, which are outlined separately in the corresponding chapters. We used a variety of tasks and analysis techniques within and across several participant groups, namely German-Spanish speakers, Italian-Spanish speakers and Dutch-Spanish speakers, to tap directly into each one of our critical issues and the corresponding research questions.

In **Chapter 2** of this thesis, we examined the effects of CLI on non-native comprehension in a syntactic violation paradigm in native German late learners of Spanish. More specifically, we studied how two CLI effects, namely the gender congruency effect and the cognate facilitation effect, modulated participants' ability to correctly identify gender agreement violations in determiner-noun pairs, for example [el pato] vs. [*la pato] *the duck*. In this, we also sought to characterise the neural correlates of syntactic violation processing in the form of the P600 component. This was because previous research on late language learners was scarce and tended to suggest a decreased sensitivity of late learners to syntactic violations. Moreover, we also tested whether and how non-native vocabulary size was related to cognitive control, which is a central feature of managing CLI between two languages. This study therefore contributed to current research first, with insights into the cumulative impact of CLI in late language learners and the potential interaction effect between linguistic features of gender congruency and cognate status on non-native comprehension; next, with a characterisation of late language learners' sensitivity to syntactic violations; and finally, with a description of CLI in non-native comprehension from a neural perspective.

In **Chapter 3**, we reported a study on non-native production conducted with the same German-Spanish speakers from Chapter 2. Non-native production is characterised by CLI as well as the necessity to select a target language over a non-target language. Yet, little research has been done on the impact of CLI on the time course of non-native production. Therefore, we used two CLI effects, namely the gender congruency effect and the cognate facilitation effect, to model the temporal unfolding of the non-native production stages and their corresponding neural signatures. In other words, we examined how CLI impacted the timing of the individual production stage during an overt picture-naming task. Next, we also probed the locus of target language selection in late language learners, with a particular focus on the modulation of the P300 component during this process. The critical contributions of this particular study were a nuanced description of CLI effects during the individual non-native production stages, as well as how these effects impacted the time course of non-native production. Moreover, we tapped not only into the mitigation of CLI, but also into the locus of target language selection and the related neural correlates. Therefore, our findings were directly relevant for the characterisation of the time course of non-native production as well as for the debate around the selection of the target language in late language learners.

Chapter 4 used the previous two chapters as its starting point and examined the same issues surrounding CLI in non-native comprehension and production in a linguistically more similar language pair, namely native Italian late learners of Spanish. In other words, we used an identical theoretical framework and experimental design to determine whether the results from Chapters 2 and 3 would apply to speakers of highly similar languages (Italian and Spanish). Using a syntactic violation paradigm to study non-native comprehension, we investigated how gender congruency and cognate status as representatives of CLI influenced non-native comprehension mechanisms. Next, we studied whether the Italian-Spanish speakers also showed a sensitivity to syntactic violations, as reflected in the P600 effect. Finally, we asked whether the P600 component was modulated by CLI. To examine non-native production, we used a picture-naming task to probe impact of CLI on the time course of non-native production and the locus of target language selection. We were particularly interested in a potential modulation of the P300 component by CLI. Taken together, our basic question in this chapter was whether the findings from the previous two chapters would also prevail with respect to a different multilingual population.

In **Chapter 5** we systematically examined an effect of language similarity on CLI in a syntactic violation paradigm in late language learners with respect to non-native comprehension. More specifically, we brought together the German-Spanish and Italian-Spanish speakers studied in Chapters 2 and 4 to ask the fundamental questions of whether speakers of highly similar languages had a processing advantage over speakers of less similar languages, and whether this notion was traceable both in terms of behaviour and neural correlates. Subsequently, we examined first, whether there were distinct neural correlates (i.e., P600 effects) for syntactic violation processing in speakers of linguistically similar vs. linguistically less similar languages; and second, whether CLI effects at the level of gender and cognates varied as a function of language similarity. Crucially, this study allowed us to investigate language similarity both from the perspective of differential neural signatures across groups, but also in terms of CLI effects and what it meant to be a speaker of highly similar languages at the cognitive level. Therefore, this work had direct implications for characterising the relevance of language similarity in non-native comprehension. Moreover, it contributed novel evidence to the small pool of literature tackling this particular issue.

Chapter 6 complemented the previous chapter as it included the direct comparison of the German-Spanish and Italian-Spanish speakers from Chapters 3 and 4 in terms of non-native production. Previous work on language similarity effects in production is scarce, in particular in combination with electrophysiological measures. Therefore, the central aim of this study was to investigate a potential language similarity effect on CLI in speakers of linguistically similar languages (Italian-Spanish) and in speakers of linguistically less similar languages (German-Spanish) at the behavioural and neural level. In this, we also investigated how gender congruency and cognate status impacted overt picture-naming in late language learners. Similar to Chapters 3 and 4, we particularly focused on the modulation of the P300 component as a function of both language similarity and CLI. The findings from this study were relevant for the in-depth examination of the P300 component as an index for CLI mitigation, and spoke directly to the importance of language similarity in non-native production.

Chapter 7 went above and beyond the role of language similarity in non-native language processing. Instead, the study reported in this chapter examined the impact of language similarity on higher cognitive functions. We described a study exploring the effect of language similarity on inhibitory control performance using a spatial Stroop task (Hilbert, Nakagawa, Bindl & Bühner, 2014). This task is commonly employed to quantify cognitive control in the form of inhibitory control (Stroop, 1935). Expanding on the notion of measurable effects of multilingualism on cognitive control, we asked whether speaking highly similar languages, e.g., Italian and Spanish, had direct consequences for inhibitory control performance compared to speaking less similar languages, e.g., Dutch and Spanish. Therefore, the fundamental goal of this study was to explore whether or not speakers of similar languages (Italian-Spanish) yielded superior inhibitory control skills compared to speakers of less similar languages (Dutch-Spanish). This study not only had implications for the relative impact of language similarity on higher cognitive functioning, but also tapped into the broader notion of language-specific vs. domain-general consequences of multilingualism on the multilingual mind and brain.

The foundations for the study reported in **Chapter 8** were developed after feedback provided by an anonymous reviewer in Chapter 2. This particular reviewer highlighted the controversial nature of the ERP correlates of syntactic violation processing in native speakers of Spanish in the context of isolated gender agreement violations. Intrigued by the disparity of results with respect to the elicitation of ERP components during gender agreement violation processing, we therefore closely examined the corresponding neural correlates in a population of native Spanish speakers. Notably, this is the only study in this thesis focusing on L1 language comprehension. Here, we also placed a particular emphasis on using more advanced statistical methods to model the oscillatory tendencies of the EEG signal over time. Subsequently, this study contributed novel evidence to the debate around which could be the primary ERP component linked to gender agreement processing in Spanish. The study further examined to which extent some of the evidence we observed was related to the amount of context given to readers to identify a gender violation. Finally, the study tapped into the broader question of whether there are functionally and neuronally distinct ERP signatures for language-specific phenomena across different (native) languages.

Finally, **Chapter 9** served the purpose of integrating the findings from each of the studies reported in this thesis. Therefore, in this chapter we synthesised our research findings, their theoretical implications and future directions to form a more general picture of the question we asked at the very beginning of this thesis: how *does* the brain manage multiple languages?

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., individuals 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 & Gollan, 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 German (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 proficiency 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.

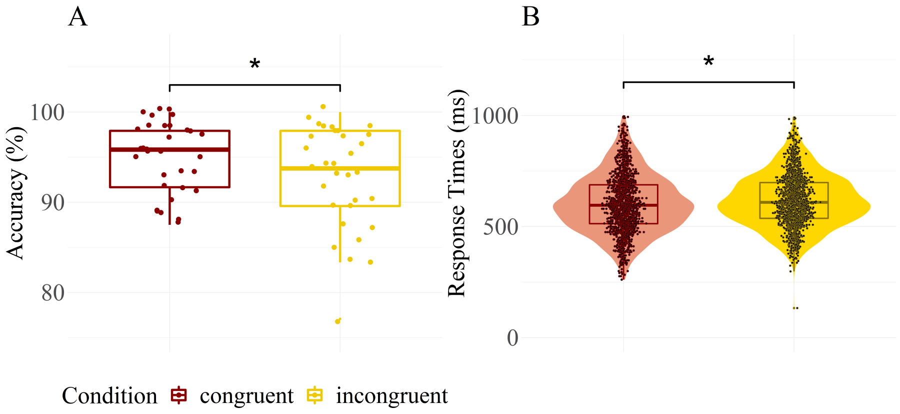
Table 2.3: *Descriptive statistics for both target languages for each condition (n = 32).*

	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 $\chi^2(2, \text{n} = 32) = 2.02$, $p = 0.364$, 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 ∼ condition + (1|subject) + (1|item).

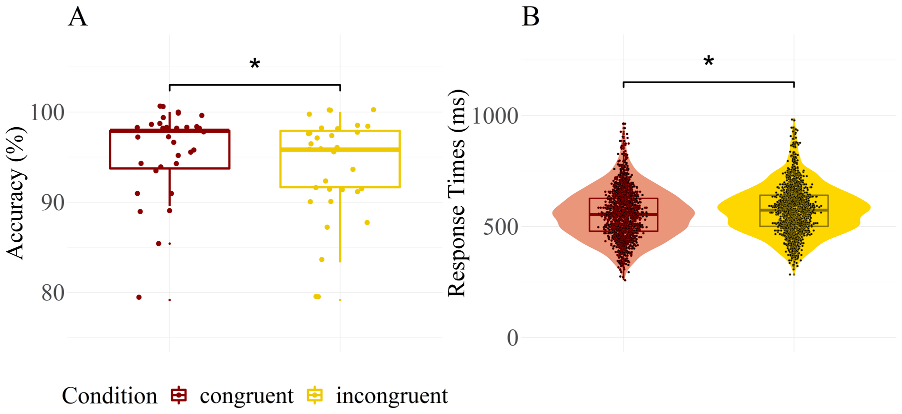
We followed a similar analysis approach for the RTs, for which we fitted an LMM. The model of best fit contained *condition* and *LexTALE-Esp score* as an interaction effect, and *subject* as random effect (Appendix 2.D). However, the interaction effect was not significant with $\beta = -0.014$, *95% 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 ∼ 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).*

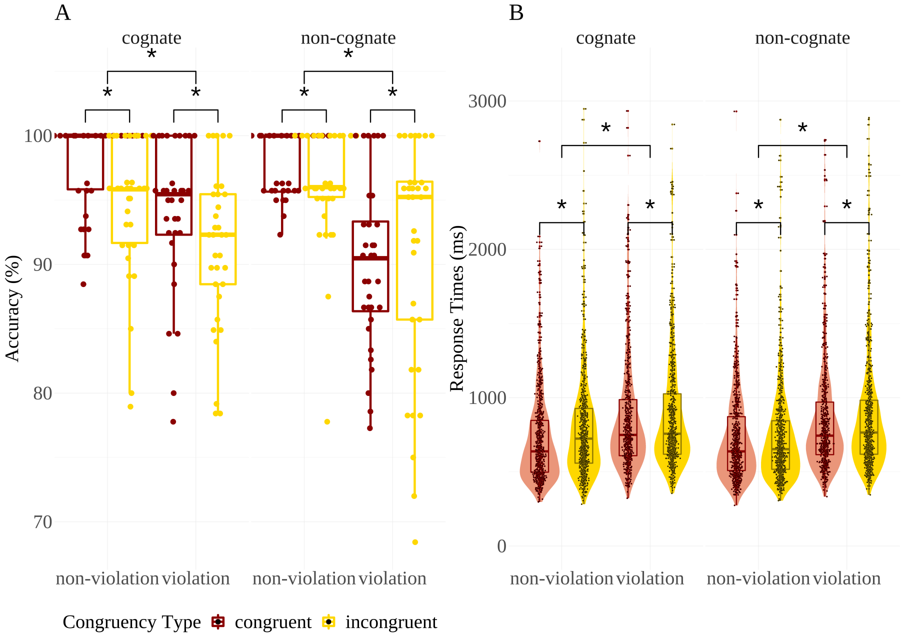
		non-violation	
		Accuracy	**RTs (ms)**
		Mean (SD)	**Mean (SD)**
cognate	**congruent**	0.979 (0.144)	729 (328)
	incongruent	0.948 (0.223)	802 (369)
Difference		0.031	73
non-cognate	**congruent**	0.978 (0.147)	727 (312)
	incongruent	0.959 (0.197)	740 (348)
Difference		0.019	13
		violation	
		Accuracy	**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], $z = \text{-}3.68$, $p < 0.001$, and *congruent* and *incongruent* trials with $\beta = 0.539$, *95% CI*[0.326, 0.892], $z = \text{-}2.40$, $p = 0.016$. 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 β = 116.12, *95% CI*[102.68, 129.56], t = 16.94, p < 0.001, for *congruent* compared to *incongruent* trials with β = 34.54, *95% CI*[20.53, 48.55], t = 4.83, p < 0.001, and finally, for *non-cognate* compared to *cognate* trials with β = -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.

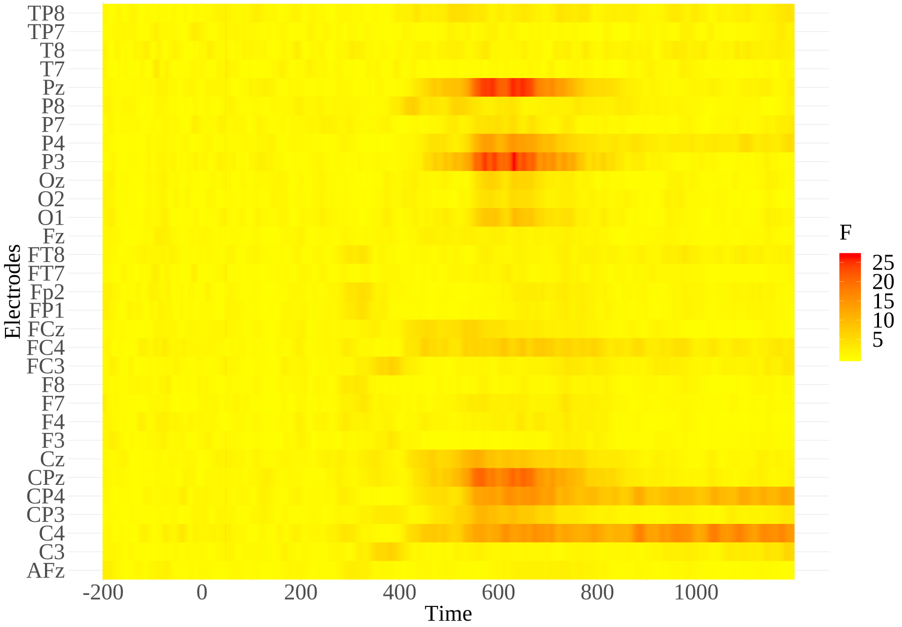
2.3.5 EEG data pre-processing

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

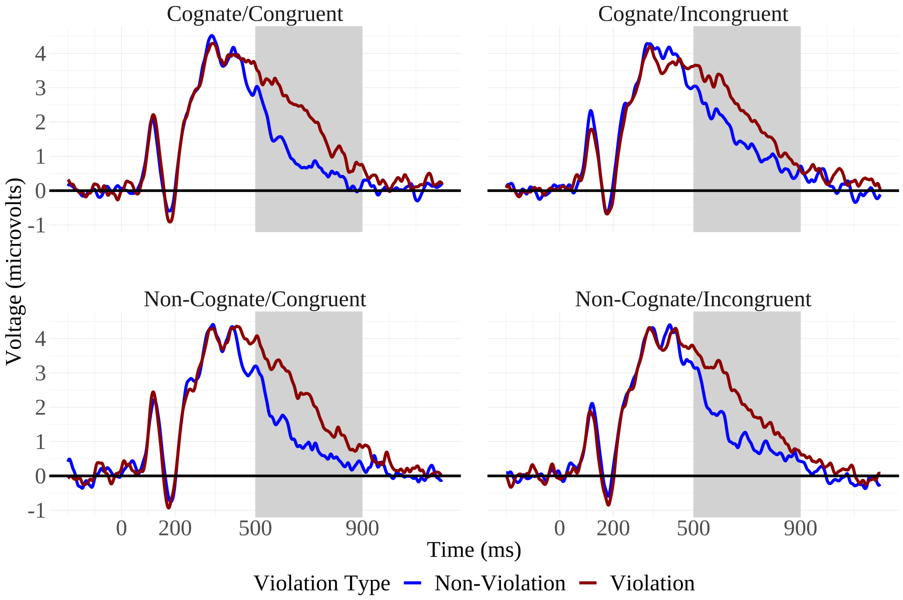
After pre-processing and exporting our data, we performed a cluster-based permutation analysis to tentatively explore the regions of interest and the potential time windows associated with significant modulations of the EEG signal. For this, we used a permutation test from the *permutes* R package (Voeten, 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 β = 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 β = 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 associated 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 contribute 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**	

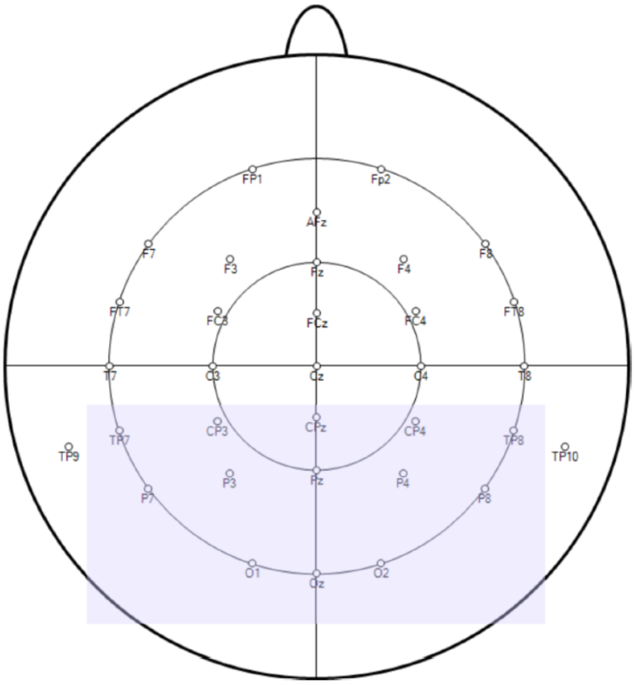
2.B Stimuli: Terminal phoneme distribution

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		
$\tau_{00 Subject}$	0.31			3,439.83		
$\tau_{00 Item}$	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
$\tau_{00 Item}$	3104.48
$\tau_{00 Subject}$	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		
$\tau_{00 Item}$	0.86		
$\tau_{00 Subject}$	1.64		
$\tau_{11 Subject[violation]}$	0.87		
$\rho_{01 Subject}$	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		

CHAPTER 3

Noun-phrase production as a window to language selection: An ERP study

This article is published as: Von Grebmer Zu Wolfsthurn, S., Pablos, L., & Schiller, N. O. (2021). Noun-phrase production as a window to language selection: An ERP study. Neuropsychologia, 162, 108055.

Abstract: Characterising the time course of non-native language production is critical in understanding the mechanisms behind successful communication. Yet, little is known about the modulating role of cross-linguistic influence (CLI) on the temporal unfolding of non-native production and the locus of target language selection. In this study, we explored CLI effects on non-native noun phrase production with behavioural and neural methods. We were particularly interested in the modulation of the P300 as an index for inhibitory control, and the N400 as an index for co-activation and CLI. German late learners of Spanish overtly named pictures while their EEG was monitored. Our results indicate traceable CLI effects at the behavioural and neural level in both early and late production stages. This suggests that speakers faced competition between the target and non-target language until advanced production stages. Our findings add important behavioural and neural

evidence to the underpinnings of non-native production processes, in particular for late learners.

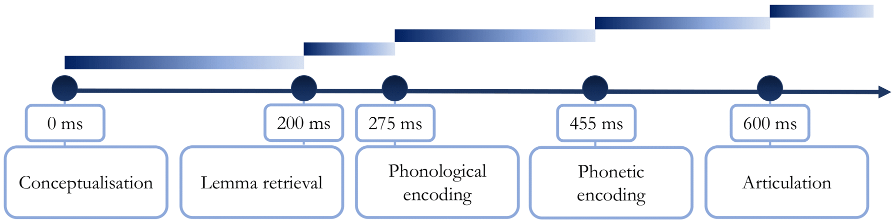
Keywords: *non-native noun phrase production, cross-linguistic influence, target language selection, late language learners, gender congruency effect, cognate facilitation effect, EEG, event-related potentials, P300, N400*

3.1 Introduction

From the speaker's perspective, producing a determiner-noun phrase (NP) e.g., [*the flower*] seems an effortless operation. However, according to current models, word production is a complex, multi-stage process. For example, the LRM model (Levelt et al., 1999) describes three primary stages of single word production. In a picture-naming task, first, the depicted object is conceptualised; second, the concept is lexicalised, i.e., it is given a grammatical, phonological and phonetic form; and finally, the name of the object is articulated. Current research has increasingly focused on characterising the time course of word production. For example, Indefrey and Levelt (2004) and Indefrey (2011) combined behavioural, chronometric and electrophysiological evidence to estimate the time course of each stage in native language word production (Abdel Rahman & Sommer, 2008; Camen, Morand & Laganaro, 2010; X. Cheng et al., 2010; Hanulová, Davidson & Indefrey, 2011; Laganaro et al., 2009; Rodriguez-Fornells, Schmitt, Kutas & Münte, 2002; Schiller, Bles & Jansma, 2003; Van Turennout & Hagoort, 1997; Zhang & Damian, 2009), see Figure 3.1.1[1].

[1] Note that Figure 3.1.1 serves for visualisation purposes and does not claim that the production processing stages are discrete stages or follow a sequential pattern, in line with Levelt et al. (1999) and Indefrey (2011). This notion is subject to an open debate (Camen et al., 2010), which is beyond the scope of our study.

Figure 3.1.1: *Estimated time course of single word production in the native language according to Indefrey and Levelt (2004) and Indefrey (2011).*



Both electroencephalography (EEG) and event-related potentials (ERPs) are particularly valuable tools to explore the neurocognitive processes of native language production. More specifically, the EEG signal yields an implicit measure of the neural signature and the time course of each individual production stage (Aristei, Melinger & Abdel Rahman, 2011; Bürki & Laganaro, 2014; Eulitz et al., 2000; Habets, Jansma & Münte, 2008; Hoshino & Thierry, 2011; Valente et al., 2014). For example, Bürki and Laganaro (2014) found that producing French NPs, such as [*le chat*] "the cat" or NPs including an adjective [*le grand chat*] "the big cat", in comparison to a bare noun [*chat*] "cat" was linked to the topographic stability of the EEG signal between 190 ms and 300 ms and following 530 ms post-stimulus onset. These findings were interpreted as a longer duration of *lexical retrieval* (lemma retrieval in LRM terms) and *phonological encoding* for NPs compared to bare nouns. They were corroborated by longer naming latencies for bare nouns and NPs compared to NPs including an adjective (Bürki & Laganaro, 2014; Lange, Perret & Laganaro, 2015; Schriefers, de Ruiter & Steigerwald, 1999). *Lexical retrieval* has been previously associated with lexical access and grammatical gender processing (Alario & Caramazza, 2002; Badecker, Miozzo & Zanuttini, 1995; Bürki & Laganaro, 2014; Levelt et al., 1999; Strijkers et al., 2010). In contrast, *phonological encoding* was described as the processing of the phonological code of the word and its subsequent syllabification (Levelt et al., 1999). Grammatical gender, hereafter gender, is a noun classification system (Corbett, 1991). More specific to gender processing in Romance languages, research suggested that the activation and selection of determiners in NPs occurred both during *lexical retrieval* and at the early part of the consecutive *phonological encoding* (Alario & Caramazza, 2002; Bürki, Sadat, Dubarry & Alario, 2016; Miozzo & Caramazza, 1999; Sá-Leite, Luna, Fraga & Comesaña, 2020). As shown in Bürki and Laganaro (2014), in these languages producing the phonological forms of determiners and adjectives is partially dependent on the phonological form of the noun, e.g., Spanish [la_F $taza_F$ $roja_F$] vs. English [*the red mug*] (Miozzo & Caramazza, 1999; Sá-Leite et al., 2020; Schriefers, 1992, 1993).

The work by Bürki and Laganaro (2014) and similar studies (Eulitz et al., 2000; Habets et al., 2008; Koester & Schiller, 2008; Lange et al., 2015) characterised the time course of native language word production. However, the time course of non-native production continues to be a complex issue in multilingualism research, especially with respect to the locus of target language selection (Costa, Strijkers, Martin & Thierry, 2009; Hanulová et al., 2011; Hoshino & Thierry, 2011; Strijkers et al., 2010). In light of the increasing prevalence of non-native speakers and multilingual communities (Berthele, 2021b), the need to further characterise the individual production stages in non-native production has become more urgent. In this study, we build upon the theoretical models of native speaker single word production and empirical findings on native NP production (Bürki & Laganaro, 2014; Indefrey, 2011; Indefrey & Levelt, 2004; Levelt et al., 1999). We specifically concentrated on the time course of those production stages preceding the articulation stage, namely *lexical retrieval* and *phonological encoding*. Collecting behavioural and EEG measures, we examined the overt production of determiner + noun NPs in the non-native language Spanish, e.g., [*la flor*] “the flower”, by native speakers of German.

Producing utterances can demonstrably be more challenging in the non-native than in the native language (Pivneva, Palmer & Titone, 2012; Runnqvist, Strijkers, Sadat & Costa, 2011). Studies have found longer and more variable naming latencies in the non-native compared to the native language (Gollan et al., 2005; Hanulová et al., 2011; Ivanova & Costa, 2008; Kroll, Bobb & Wodniecka, 2006). These quantitative differences between native and non-native speech production were reported for various levels of language proficiency (Christoffels et al., 2007; Ivanova & Costa, 2008; Sholl, Sankaranarayanan & Kroll, 1995), and for language pairs with varying phonological and orthographic similarity, e.g., intermediate German learners of Dutch (Christoffels et al., 2007) and highly proficient Greek learners of English (Parker-Jones et al., 2012). The question which arises at this point is: where does this delay in naming latencies originate from? Recent studies explored word frequency and age of acquisition (AoA) as modulating factors of the non-native production processes, see Hanulová et al. (2011) for a discussion. In this study, we focus on another factor that could influence the time course of non-native short utterance production, namely *cross-linguistic influence* (CLI), as described in section 3.1.1. By extension, we aim to explore the following question crucial to non-native production research: during which production stage does the delay in naming latency occur? In section 3.1.2, we outline the electrophysiological correlates of CLI in more detail and we discuss how they offer us an insight into these issues.

3.1.1 Cross-linguistic influence

Non-native speakers face cross-linguistic influence (CLI) during language production and language comprehension (Cárdenas-Hagan et al., 2007; Ganushchak, Verdonschot & Schiller, 2011; Lemhöfer et al., 2008; Morales et al., 2016; Müller & Hulk, 2001; Thierry & Wu, 2007; Von Grebmer Zu Wolfsthurn, Pablos-Robles & Schiller, 2021a). Broadly speaking, CLI is the interaction of the languages within a multilingual system and its influence on the underlying cognitive processing mechanisms. CLI supports the notion that the native and non-native language are co-activated during language production (Guo & Peng, 2006; Hermans et al., 1998; Kroll, Bobb, Misra & Guo, 2008; Lee & Williams, 2001). Co-activation and CLI are rooted into theoretical models. For example, the Revised Hierarchical Model, RHM (Kroll & Stewart, 1994) postulates a conceptual level and separate lexical levels for the native and non-native language with strong lexical connections between the two languages. Critically, the model also suggests that the strength of the connections is modulated by proficiency: as non-native proficiency increases, the connection strength between the non-native lexicon and the conceptual level increases and the involvement of the native language becomes less prominent (Kroll & Stewart, 1994).

For speakers to successfully complete a naming task in either the native or non-native language, it is crucial that co-activation and CLI are resolved prior to articulation. A robust finding in the CLI and language selection literature is the presence of a language control system which mitigates CLI effects and effectively inhibits the non-target language (Abutalebi & Green, 2007; D. W. Green, 1998), but see Verdonschot, Middelburg, Lensink and Schiller (2012). The mitigation of CLI effects and the associated increased cognitive effort is evident at the neural level. For example, increased activation of brain areas involved in language production for non-native compared to native language production was linked to increased error monitoring of competing representations during CLI (Parker-Jones et al., 2012; Rodriguez-Fornells, Kramer, Lorenzo-Seva, Festman & Münte, 2012; E. Rossi, Newman, Kroll & Diaz, 2018). Previous studies on CLI have almost exclusively focused on early acquisition and intermediate to high proficiency levels in the non-native language, thereby leaving a systematic gap in the exploration of the effects of CLI on the time course of NP production in late language learners with lower proficiency levels (Costa et al., 2003; Hoshino & Thierry, 2011; Lemhöfer et al., 2008). Yet, this is a critical issue because studies suggested that proficiency impacted language-related neuro-cognitive mechanisms in multilinguals, shown in that CLI effects were more pronounced at lower proficiency levels (Bosch & Unsworth, 2020; Heidlmayr, Ferragne & Isel, 2021; Steinhauer et al., 2009; Van der Meij, Cuetos, Carreiras & Barber, 2011; White, Titone, Genesee & Steinhauer, 2017; Yip & Matthews, 2007). For example, Sunderman and Kroll (2006) found that compared to lower proficient learners, highly proficient English learners of Spanish were less susceptible to CLI from the native language, and performed better in a picture-naming task. Costa and Santesteban (2004) previously proposed that during production, highly proficient speakers activated only the lexical entry from the target language, thereby effectively avoiding CLI during *lexical retrieval.* Therefore, in this study we directly focused on CLI effects in a group where first, CLI was found to be most prevalent; and second, is frequently understudied in the literature, namely late language learners with intermediate proficiency levels. We defined late language learners as having acquired an additional (non-native) language later in development (AoA $>$ 12 years), see S. Rossi et al. (2006). Moreover, and in contrast to highly proficient late language learners, our group was further characterised by less than three years of exposure to the non-native language and intermediate proficiency levels in the B1/B2 range according to the Common European Framework of Reference for Languages, CEFR (Council of Europe, 2001).

Immediately relevant to CLI effects is the question about when the target language is selected in non-native production. For example, is the target language selected (and CLI resolved), prior to *lexical retrieval*? Or instead, does CLI carry over to later production stages, such as *phonological encoding*? Current debates remain inconclusive with respect to two accounts of the locus of target language selection (Costa, Santesteban & Ivanova, 2006; Hanulová et al., 2011; Hoshino & Thierry, 2011; Sá-Leite et al., 2019). One account suggests that lexical entries from both the target and non-target language are activated, but only the lexical entry corresponding to the target language is selected for subsequent phonological processing (Gollan et al., 2005; Hermans et al., 1998; Lee & Williams, 2001). Under this account, CLI is resolved at lexical retrieval. On the other hand, a second account suggests that the lexical entries from the target and non-target language are both activated and selected for *phonological encoding* (Christoffels et al., 2007; Colomé, 2001; Costa et al., 2000; D. W. Green, 1998; Hoshino & Kroll, 2008; Pulvermüller, 2007; Rodriguez-Fornells et al., 2005).

Within this perspective, CLI is not resolved at *lexical retrieval*, but continues into subsequent phonological processing. In order to discriminate between these two accounts, in this study we focus on two linguistic phenomena representing CLI, the *gender congruency effect* and the *cognate facilitation effect*. These effects provide us with further insight into the underlying production stages and their inner mechanisms in non-native NP production, see the following sections.

The gender congruency effect

The gender congruency effect is reflected in faster processing of *congruent* vs. *incongruent* nouns, as reported in language production studies (Bordag & Pechmann, 2007; Lemhöfer et al., 2008; Morales et al., 2016; Paolieri et al., 2019, 2020; Schiller & Caramazza, 2003; Schiller, 2006; Schiller & Costa, 2006). Congruent nouns have similar gender values across languages; for example, the lexical items for the concept "arm" are masculine in German [der_M *Arm*] and in Spanish [el_M *brazo*]. In contrast, incongruent nouns have dissimilar gender values across languages; for example, the lexical items for "key" are masculine in German [der_M *Schlüssel*] and feminine in Spanish [la_F *llave*]. Gender systems can vary across languages. For example, German has three gender values, i.e. feminine, masculine and neuter, and their distribution is not equally distributed across all lexical items (Schiller & Caramazza, 2003). Spanish, however, is characterized by a feminine-masculine gender value distinction with an approximately balanced distribution (Bull, 1965; Eddington, 2002)[2].

As discussed above, gender processing in Romance languages was linked to *lexical retrieval* and *phonological encoding* (Alario & Caramazza, 2002; Badecker et al., 1995; Bürki & Laganaro, 2014; Miozzo & Caramazza, 1999). Subsequently, this links the gender

[2] Note that similar labels for gender values ("masculine", "feminine") can be found across different languages. However, we do not assume that these labels are conceptually identical (Lemhöfer et al., 2008) but merely utilise them for descriptive purposes.

congruency effect to these two production stages. Therefore, the gender congruency effect offers a gateway to the following three issues: first, it allows us to observe the mechanisms underlying CLI of the gender systems during *lexical retrieval*; second, it provides us with a way to study the implications of CLI of the gender systems on the time course of non-native NP production; and third, it allows us to explore the locus of target language selection with respect to the two accounts regarding target language selection (Hoshino & Thierry, 2011; Sá-Leite et al., 2019). If the target language was selected prior to *lexical retrieval* in multilingual language production, we would not observe a gender congruency effect during the subsequent production stages. Alternatively, if the target language was not selected before *lexical retrieval*, we would observe a gender congruency effect because the activated lexical entries from both languages would be subject to CLI of the gender systems during gender processing. As a result, CLI would facilitate the processing of congruent nouns compared to incongruent nouns. Current behavioural evidence supports this notion in late language learners with intermediate proficiency compared to early learners with high proficiency (Bordag & Pechmann, 2007; Costa et al., 2003; Sá-Leite et al., 2020).

Yet, few studies so far have investigated the gender congruency effect from a neural perspective (Heim, Friederici, Schiller, Rüschemeyer & Amunts, 2009). One ERP component associated with this effect is the N400 (Paolieri et al., 2020; Wicha, Bates, Moreno & Kutas, 2003). It is reflected in a negative voltage amplitude peak around 400 ms post-stimulus onset and was previously linked to lexical-semantic integration and lexical co-activation (P. Chen, Bobb, Hoshino & Marian, 2017; Hoshino & Thierry, 2011; Kutas & Federmeier, 2011; Lau, Phillips & Poeppel, 2008; Leckey & Federmeier, 2019). In relation to the gender congruency effect, less negative N400 amplitudes were linked to congruent trials compared to incongruent trials in the time window between 300 ms and 500 ms post-stimulus onset in a translation-recognition task (Paolieri et al., 2020; Wicha et al., 2003). Given the temporal characteristics of these neural correlates of the gender congruency effect, this suggests that both co-activation and CLI between the languages may remain unresolved until around 500 ms post-stimulus onset. In LRM terms, this time window coincides with *phonological encoding*. Therefore, the behavioural and ERP findings related to the gender congruency effect suggest that the target language is not selected prior to lexical retrieval. However, from these findings it remains unclear whether CLI is resolved upon termination of gender processing or whether it indeed continues into *phonological encoding* stages.

Regarding the specific question about the locus of target language selection, there are two possible scenarios. In the first scenario, the target language is selected after the completion of gender processing. Subsequently, only the lexical entry from the target language carries over to phonological encoding. In contrast, the second scenario postulates that the target language is selected during or after *phonological encoding*. Here, the lexical entries from both the target and non-target language are processed for *phonological encoding*. This implies that both languages remain active after the completion of gender processing and that CLI potentially results in further delays during later phonological processing. Evidence by Hoshino and Thierry (2011) preliminarily supported the latter notion. In a picture-word interference (PWI) task with highly proficient Spanish learners of English, the EEG signal was modulated during lexical retrieval for semantically and phonologically related trials compared to unrelated trials. However, they found no further modulation of the signal after 400 ms post-stimulus onset. The authors interpreted these results as showing that the target language had not been selected at lexical retrieval, but that the selection had taken place by 400 ms post-stimulus onset. In LRM terms, this time window coincides with *phonological encoding*. Therefore, these results supported the second scenario whereby first, lexical entries from both the target and non-target language were selected for phonological processing, and second, CLI continued beyond lexical retrieval (Christoffels et al., 2007; Colomé, 2001; Costa et al., 2000; Hoshino & Kroll, 2008; Hoshino & Thierry, 2011; Pulvermüller, Shtyrov & Hauk, 2009; Rodriguez-Fornells et al., 2005). Building on the work by Hoshino and Thierry (2011) to add further evidence for discriminating between the two scenarios outlined above, we also explored the *cognate facilitation effect*. Via the exploration of this effect, we probed whether or not CLI continued after gender processing in late language learners.

The cognate facilitation effect

In a broad sense, cognates are words with a large degree of phonological and orthographic overlap (C. Li & Gollan, 2018). For example, [*Melone*] and [*melón*] "melon" are examples for *cognates* in German and Spanish, respectively, whereas [*Arm*] and [*brazo*] "arm" are *non-cognates*. It has previously been shown that cognates are processed faster compared to non-cognates (Bosma et al., 2019; Casaponsa, Antón, Pérez & Duñabeitia, 2015; C. Li & Gollan, 2018). This is a critical finding as it suggests that this *cognate facilitation effect* can be linked to the CLI during *phonological encoding* in production (Christoffels et al., 2007; Costa et al., 2000, 2005; Hoshino & Kroll, 2008). In order for this effect to occur, the lexical entries from both the target and non-target language need to be subject to subsequent *phonological encoding*. Here, we used the cognate facilitation effect to first, study CLI during this particular production stage with respect to the overall time course of non-native production; and second, to add to the discussion of whether or not the target language is selected after gender processing.

While neural correlates of the cognate facilitation effect have been scarcely researched in non-native production until now, evidence from non-native comprehension links the N400 to this specific effect (Midgley et al., 2011; Peeters, Dijkstra & Grainger, 2013; Xiong, Verdonschot & Tamaoka, 2020). For example, Peeters et al. (2013) found faster latencies and smaller N400 amplitudes for cognates compared to non-cognates between 400 ms and 500 ms in French late learners of English in a lexical decision task. Contrastingly, Christoffels et al. (2007) found faster naming latencies and more negative amplitudes for cognates compared to non-cognates in fronto-central regions between 275 ms and 475 ms post-stimulus onset in unbalanced German-Dutch speakers in an overt picture naming task, without linking it to the N400. Further, studies showed that the size of the cognate facilitation effect decreased as non-native proficiency increased (Bultena, Dijkstra & van Hell, 2014; Casaponsa et al., 2015). Therefore, ERP evidence from the cognate facilitation effect suggests that first, the target and the non-target language are co-activated, which in turn leads to CLI of the phonological systems; second, that lexical entries from both languages are subject to *phonological encoding*; and finally, that the non-target language is not inhibited during *lexical retrieval*, particularly at lower proficiency levels. Instead, the cognate facilitation effect suggests that CLI continues beyond lexical retrieval into *phonological encoding* stages, as supported by the literature (Christoffels et al., 2007; Colomé, 2001; Costa et al., 2000; Hoshino & Kroll, 2008; Hoshino & Thierry, 2011).

Combining evidence from the gender congruency effect and the cognate facilitation effect, the findings presented above suggest a modulating influence of CLI on the time course of non-native production that continues beyond gender processing until phonological processing stages. However, these interpretations are debatable given the scarcity of research on CLI, in particular in terms of the neural correlates of the gender congruency effect and the cognate facilitation effect. Therefore, aside from exploring CLI effects in late language learners, we also focused on the neural underpinnings of CLI to characterise the time course of non-native production and the locus of target language selection.

3.1.2 Electrophysiological correlates of CLI

As discussed above, the N400 was linked to CLI effects such as the gender congruency effect and the cognate facilitation effect, but also the co-activation of languages (P. Chen et al., 2017; Paolieri et al., 2020; Peeters et al., 2013). In this study, we used the N400 to capture the co-activation and the linguistic aspects of CLI. Importantly, studies suggested that the N400 onset may be delayed in late language learners and that overall N400 amplitudes in these learners decrease compared to those of native speakers (Midgley, Holcomb

& Grainger, 2009; Weber-Fox & Neville, 1996). Some studies further suggested that N400 amplitudes were modulated by non-native proficiency, i.e., the N400 became more native-like with increasing proficiency (Midgley et al., 2009; Newman, Tremblay, Nichols, Neville & Ullman, 2012; White et al., 2017), but see Wood Bowden, Steinhauer, Sanz and Ullman (2013). For example, in a phoneme discrimination task with French low and high proficient late language learners of English, Heidlmayr et al. (2021) found a smaller N400 effect for the low compared to the high proficient group of late language learners. Therefore, with respect to our group of late language learners, the N400 is likely to be delayed, or smaller in size, compared to that in Paolieri et al. (2020) and (Peeters et al., 2013).

In this study, we also focused on the P300 component to capture the cognitive mechanisms underlying the successful mitigation of CLI and the selection of the target language. The P300 is a positive-going deflection of the EEG signal with a peak around 300 ms post-stimulus onset. Early studies found that the P300 was elicited at different topographical sites (Ritter & Vaughan, 1969), leading to suggest separate P300 subcomponents. These subcomponents include the P3a, and the P3b (Barry et al., 2020; Polich, 2007). The P3a component is a positive-going wave with a fronto-central distribution which occurs around 200 ms–300 ms post-stimulus onset. In contrast, the P3b component was found in later 300 ms–400 ms time windows at centro-parietal electrodes (Hruby & Marsalek, 2003; Squires, Squires & Hillyard, 1975). Relevant to this study, the P300 has been previously linked to cognitive processes such as cognitive interference, cognitive control, working memory load and inhibition (Barker & Bialystok, 2019; Luck, 1998; Neuhaus, Trempler et al., 2010; Polich, 2007). More recently, the P300 was linked to the allocation of attentional resources (Barker & Bialystok, 2019; González Alonso et al., 2020). It is typically found with inhibitory paradigms such as the Flanker task or the Oddball paradigm (Eriksen & Eriksen, 1974; Pereira Soares et al., 2019). More relevant to our study, it was also reported in paradigms which included a Flanker task preceded by a linguistic task such as code-switching or picture-naming, e.g., in Bosma and Pablos (2020) and in Jiao, Grundy, Liu and Chen (2020). These studies highlighted the critical role of the P300 in inhibition and in regulating native and non-native language use. Our experimental paradigm relies on the successful mitigation of CLI effects and the inhibition of the non-target language. Therefore, the P300 is a critical component to consider in this study alongside the N400.

3.1.3 The current study

In the current study, we explore the effect of CLI on the time course of non-native NP production from a behavioural and neural perspective. The goals of the present study are twofold. First, we investigate how CLI affects behavioural measures and the EEG signal in non-native NP production. On the basis of the LRM model, we use two CLI effects to characterise the unfolding and neural signatures of individual production stages: *lexical retrieval* via the *gender congruency effect*, and *phonological encoding* via the *cognate facilitation effect*. Our second goal is to gain further insight into the process of target language selection. More specifically, we probe the locus of target language selection by investigating the two CLI effects with respect to the individual production stages. Therefore, our research questions are: first, are there traceable effects of gender congruency (congruent vs. incongruent) and cognate status (cognate vs. non-cognate) at the behavioural and neural processing level in non-native NP production? Second, what can the temporal unfolding of processing gender congruency and cognate status tell us about the time course of non-native NP production? Finally, during which processing stage is the target language selected during non-native NP production?

We study non-native NP production in German late intermediate learners of Spanish with a B1/B2 proficiency level by employing an overt picture-naming task. We combine behavioural measures of naming accuracy and naming latencies with EEG recordings. We are particularly interested in the modulation of the P300 as an index for inhibitory control, and the N400 as an index for co-activation and CLI. To obtain information about the linguistic background of our late language learners, we combine the Language Experience and Proficiency Questionnaire, LEAP-Q (Marian et al., 2007) with a Spanish vocabulary size test, the LexTALE-Esp (Izura et al., 2014). To formulate the hypotheses for this study, we rely both on the time estimates proposed by Indefrey and Levelt (2004), Indefrey (2011), Bürki and Laganaro (2014) and Hoshino and Thierry (2011), and on the theoretical framework and discussion of ERP components outlined in the previous sections.

Hypotheses

Behavioural hypotheses. We predict effects of gender congruency and cognate status on behavioural measures of naming accuracy and naming latencies. For congruent and cognate nouns, we predict a facilitatory CLI effect, reflected in higher naming accuracy and shorter naming latencies compared to incongruent and non-cognate nouns. In turn, this has direct implications for the time course of non-native NP production. For congruent non-cognates and incongruent cognates, we predict more subtle CLI effects. In concrete behavioural terms, we expect lower naming accuracy and longer naming latencies for congruent non-cognates compared to congruent cognates, and higher naming accuracy and shorter naming latencies compared to incongruent non-cognates. On the other hand, we anticipate the reverse pattern for incongruent cognates: CLI would hinder gender processing, but act as a facilitator during phonological processing and influence the time course of non-native NP production.

EEG hypotheses. We first probe the presence of a P300 or an N400 effect, as existing research remains inconclusive about whether or not we can expect both components to be elicited in our experimental paradigm. Further, we predict a modulation of P300 and N400 as a function of condition. We expect smaller P300 and N400 amplitudes for producing congruent cognates compared to incongruent non-cognates. This would reflect higher processing costs and more involvement of the inhibitory control system for the latter. For congruent non-cognates and incongruent cognates, we predict a similar degree of processing costs. These trials are subject to both CLI facilitation and hindrance. Therefore, we do not expect significant differences between congruent non-cognates and incongruent cognates in terms of P300 and N400 amplitudes. However, we do expect the P300 and N400 amplitudes for these particular trials to be significantly larger compared to congruent cognates, and to be significantly smaller for incongruent non-cognates. Therefore, we expect the smallest P300 and N400 amplitudes for congruent cognates, followed by larger amplitudes for congruent non-cognates and incongruent cognates, and finally, the largest amplitudes for incongruent non-cognates.

3.2 Methods

3.2.1 Participants

Thirty-three (twenty-seven females) healthy, right-handed native German speakers with a B1/B2 level of Spanish were recruited from the campus of the University of Konstanz ($M_{age} = 23.06$ years, $SD_{age} = 2.47$ years). At the time of testing, participants did not report any psychological or language disorders, nor visual and hearing impairments. Prior to the experiment, we provided all participants with an information sheet. Next, they signed an informed consent form before the experiment in compliance with the Ethics Code for linguistic research in the Faculty of Humanities at Leiden University. Upon termination of all tasks, participants received a debrief form, signed the final consent form and received a monetary compensation.

LEAP-Q: linguistic profile of participants

Prior to the experimental session, the linguistic profile of participants and their experience with Spanish was assessed using LEAP-Q (Marian et al., 2007); see Appendix 3.A for details. We opted for a home-based administration of the questionnaire to minimise any self-report biases often induced by laboratory environment (Rosenman et al., 2011). The majority of participants (n = 31) reported English as their first non-native language (M_{AOA} = 8.90, SD_{AOA} = 1.90), while two participants learnt French as their first foreign language ($M_{AOA} = 8.5$, $SD_{AOA} = 2.5$). A total of sixteen participants learnt Spanish as second non-native language. Further, Spanish was disclosed as third non-native language by fifteen participants, and as fourth non-native language by two participants. The mean AoA of Spanish was $M_{AOA} = 16.29$ years ($SD_{AOA} = 2.39$). Participants reported to be fluent in Spanish on average at M = 18.53 years of age ($SD = 2.29$). They started to read in Spanish at $M = 17.27$ years of age ($SD = 3.03$). Before the time of testing, almost all participants (n = 31) spent some time in a Spanish-speaking country ($M = 0.96$ years, $SD = 0.69$). On a scale from one to ten (ten being maximally proficient), participants reported a current speaking proficiency of $M = 6.76$ ($SD = 1.00$) in Spanish. Further, they classified their comprehension proficiency with M = 7.34 ($SD = 0.92$) and finally, reading proficiency with $M = 7.18$ ($SD = 1.07$). On a scale from zero to ten (ten being maximally exposed), participants quantified their exposure to Spanish at the time of testing with $M = 5.20$ ($SD = 2.48$). This compares to an exposure of $M = 3.12$ ($SD = 2.31$) to their first foreign language. On a daily basis, this corresponded to an exposure to Spanish of M = 10.03% ($SD = 9.48$) compared to the other languages. Exposure to Spanish occurred via the following contexts: interaction with Spanish native speakers, listening to Spanish radio shows, watching Spanish television, reading or self-instruction in Spanish. At the time of testing, six participants reported a self-perceived proficiency of Spanish as first non-native language, twenty-six participants as second non-native language, and one participant as third non-native language. We used this metric as a proxy for moderate confidence levels with Spanish.

Noting here that most participants acquired one or more foreign languages prior to acquiring Spanish is important. Research on CLI effects in L3 (L4, L5, etc.) language processing has demonstrated that all languages within a multilingual system might affect processing in the target language (Lago et al., 2021; Lemhöfer & Dijkstra, 2004; Rothman, 2015). Here, language dominance was found to be a driving factor of CLI, in that more dominant languages are linked to stronger interference, compared to less dominant languages (Francis & Gallard, 2005; Lago et al., 2021). In our study, a total of eighteen participants reported that they had acquired French, of which fourteen acquired it prior to Spanish. AoA of French was $M = 11.38$ years of age ($SD = 1.98$). Accordingly, speakers of French reported a speaking proficiency of $M = 2.85$ ($SD = 0.87$), a comprehension proficiency of $M = 4.15$ ($SD = 1.46$) and finally, a reading proficiency of $M = 4.85$ ($SD = 1.72$) on a scale from one to ten. At the time of testing, exposure to French was reported as $M = 0.56$ ($SD = 2.72$) on a scale from zero to ten. Compared to the other languages, participants were exposed to French $M = 1.11\%$ ($SD = 1.41$) on a daily basis. All participants who had acquired French claimed a higher self-reported proficiency for Spanish compared to French, which was reported as third nonnative language following Spanish. Therefore, on the basis of previous research (Lago et al., 2021), we predict only a limited influence of French on CLI effects due to the low dominance and proficiency of speakers in this language at the time of testing. Nevertheless, we included the acquisition of French as a covariate in our analysis to see whether this had an effect on our results. As discussed in section 3.3, we did not find an effect on our outcome variables.

3.2.2 Materials and design

Prior to the experimental session, we asked participants to complete the LEAP-Q. In the laboratory, they completed the LexTALE-Esp vocabulary size test and an overt picture-naming task. We measured EEG during the picture-naming task.

Tasks and stimuli

The LexTALE-Esp and the picture-naming task were both programmed in E-prime 2 (Schneider, Eschman & Zuccolotto, 2002) and administered on a Windows 10 computer.

LexTALE-Esp. We created an E-prime version of the LexTALE -Esp task with identical instructions and stimuli. This task was used to complement self-reported measures from the LEAP-Q and the vocabulary size score was added as a covariate to subsequent statistical analyses.

Picture-naming task. Our picture stimuli were obtained from the MultiPic picture database (Duñabeitia et al., 2018). We selected the picture stimuli according to two criteria: those with the highest percentage of valid responses given by participants and those with the highest percentage of participants giving the object's exact name. Then, each picture was assigned a gender congruency type (*congruent* versus *incongruent* across German and Spanish) and a cognate status (*cognate* versus *non-cognate* across German and Spanish). The latter was based on the degree of semantic, phonological and orthographic overlap. We excluded identical cognates [*die Kiwi*] – [*el kiwi*] "the kiwi", plural forms [*die Brille*] – [*las gafas*] "the glasses", professions [*die Sängerin*] – [*la cantante*] "the (female) singer", English loanwords [*der Boomerang*] – [*el boomerang*] "the boomerang", and translation equivalents of opposing genders [*der Esel*] – [la_F $mula_F/el_M$ $burro_M$] "the donkey". To increase ecological validity of our stimuli, we modelled the distribution of terminal morphemes in Spanish according to previous work by Clegg (2011). Further, we included terminal phoneme of the target noun as a covariate and item (i.e., the individual picture) as a random effect in the statistical analyses (see sections 3.3.2. and 3.3.5. for more details).

EEG recordings

EEG data were collected via the BrainVision Recorder software (Version 1.23.0001) by Brain Products GmbH. We used an EasyCap electrode cap following a standard 10/20 montage (Appendix 3.B). Data were measured at thirty-two channel locations via passive electrodes. We recorded the horizontal electrooculogram (HEOG) from two electrodes at the outer canthus of the left and right eye. We recorded the vertical electrooculogram (VEOG) from an electrode placed below the left eye. All electrodes were initially referenced to channel Cz, which we later reused as a data channel during rereferencing. The ground electrode was placed on the right cheek of the participant. Impedances of the electrodes were checked and configured using actiCAP Control Software (Version 1.2.5.3) by Brain Products GmbH. We kept impedances below 5 kΩ for the reference and ground electrode. For the remaining channels, impedances were below 10 kΩ. The sampling rate was 500 Hz.

3.2.3 Procedure

LexTALE-Esp

Participants first completed the LexTALE-Esp task. During this task, we presented them with a fixation cross at the centre of the screen for 1,000 ms. This was followed by the visual presentation of a letter string on the horizontal midline of the screen which corresponded to either a Spanish *word*, or a pronounceable *pseudo-word*. Participants were then asked to indicate via a button-press whether or not the letter string corresponded to a Spanish word. The letter string remained on the screen until the participant responded. Each letter string was only shown once. The total number of trials was 87, because we excluded three trials due to an overlap with the experimental stimuli from the picture-naming task prior to the experiment. Offline, we calculated the vocabulary size score by subtracting the percentage of incorrectly identified pseudo-words from the percentage of correctly identified words for each participant (Izura et al., 2014). The maximum score was 100, whereas the minimum score varied as a function of false positives.

Picture-naming task

For the picture-naming task, we followed a 2 x 2 fully factorial within-subjects design with two main manipulations: *gender congruency* and *cognate status*. Half of the trials were *congruent* in that the gender was similar across German and Spanish. The other half of trials were *incongruent*, characterised by a dissimilarity in gender across German and Spanish. Further, half of congruent and incongruent pictures were *cognate* words, and the other half were *non-cognate* words (Table 3.2.1). There were 24 stimuli per condition, resulting in a total of 96 stimuli.

Table 3.2.1: *Sample set of stimuli for the picture-naming task.*

		congruent	**incongruent**
cognate	**German**	die$_F$ Giraffe$_F$	die$_F$ Melone$_F$
	Spanish	la$_F$ jirafa$_F$	el$_M$ melón$_M$
		the giraffe	*the melon*
non-cognate	**German**	der$_M$ Arm$_M$	der$_M$ Schlüssel$_M$
	Spanish	el$_M$ brazo$_M$	la$_F$ llave$_F$
		the arm	*the key*

Following the standard procedure in the field of speech production, the task was divided into a familiarisation phase and an experimental phase, with a total duration of 30–40 min. The familiarisation phase consisted of three rounds. In each round, participants were exposed to all 96 stimuli pictures and were instructed to overtly name each picture in Spanish using an NP construction with the correct determiner and noun (e.g., [*el brazo*] "the arm"). During this phase, the experimenter provided oral feedback on the accuracy of the NP production by the participant whenever necessary. Specifically, the correct determiner or noun was provided in Spanish for cases where either the determiner or the noun, or both, were incorrectly produced by the participant. In the experimental phase, participants named the objects as fast and accurately as possible using a Spanish NP. Participants' EEG and voice were recorded exclusively during the experimental phase. A typical trial was initiated with the display of a fixation cross for 1,000 ms, followed by the display of the picture for 2700 ms in the centre of the screen. Each picture was shown only once during the experimental phase, resulting in a total of 96 trials. Trial order was randomized. Participants were reminded throughout the experimental phase to name the object as fast and accurately as possible and to reduce all unnecessary movement. There was a short break after 50 trials to minimise participants' fatigue.

3.3 Results

3.3.1 Behavioural data exclusion

Naming latencies and EEG data for one participant were lost due to a malfunctioning microphone and a subsequent failure during the EEG recording. Further two participants were excluded to match the datasets included in the EEG analysis (see section 3.3.5 for details). In total, we included 30 datasets in this analysis.

3.3.2 Behavioural data analysis

We used Praat (Broersma & Weenink, 2019) to calculate naming accuracy and naming latencies for each trial for the picture naming task. Next, we analysed our behavioural data in RStudio Version 1.3.959 (R Core Team, 2020). We employed a single-trial modelling approach using the *lme4* package (Bates et al., 2020) to model our two behavioural outcome variables, *naming accuracy* and *naming latencies.* We modelled naming accuracy using generalised linear mixed models (GLMM) and the *glmer()* function with a binomial distribution. Next, we modelled positively skewed naming latencies using the *glmer()* function in combination with a gamma distribution and the identity link function. Only correct trials were included in our analysis of the naming latencies. For both outcome variables, we generated the most theoretically plausible maximal model on the basis of our hypotheses and our two main manipulations, *gender congruency* and *cognate status.* To preserve statistical power and to control for potential confounds, we added *familiarisation phase performance* as a covariate to our statistical analysis, rather than excluding trials where errors were made before the experimental phase. Similarly, we added *LexTALE-Esp score, target noun gender, word length, order of acquisition of Spanish, acquisition of French* and *terminal phoneme* as covariates. Further, we included *subject* and *item* as random effects. To establish the model of best fit for the picture-naming task, we followed a top-down model selection procedure by testing for the significance of each factor (Barr, 2013; Bates et al., 2018). In order to balance Type I error and power, random effects were chosen as maximal as possible while avoiding over-fitting (Matuschek et al., 2017). In the case of non-convergence or singular fit, we simplified our model structure by removing first interactions for random slopes; second, correlations between random slopes; and finally, interactions between fixed effects and covariates. We used treatment coding as our contrast, which defaulted to congruent trials and cognates as the reference level. Absolute t-values greater than 1.96 were interpreted as statistically significant at $\alpha = 0.05$ (Alday et al., 2017). Next, we performed model comparisons using the *anova()* function based on the Akaike's Information Criterion, AIC (Akaike, 1974), the Bayesian Information Criterion, BIC (Neath & Cavanaugh, 2012) and the log-likelihood ratio. To perform model diagnostics, we checked the model fit by plotting the model residuals against the predicted values.

3.3.3 Behavioural data results

LexTALE-Esp

The mean LexTALE-Esp score was $M = 18.45$ ($SD = 20.52$). Scores were highly variable and ranged between -23 and 60, with 100 being the maximum score. Vocabulary scores of 60–80 on this task were previously associated with C1–C2 proficiency levels (Lemhöfer & Broersma, 2012), therefore all of our speakers fell below the B2 proficiency range.

Picture-naming task

We first calculated descriptive statistics for naming accuracy and naming latencies for each condition (Table 3.3.1).

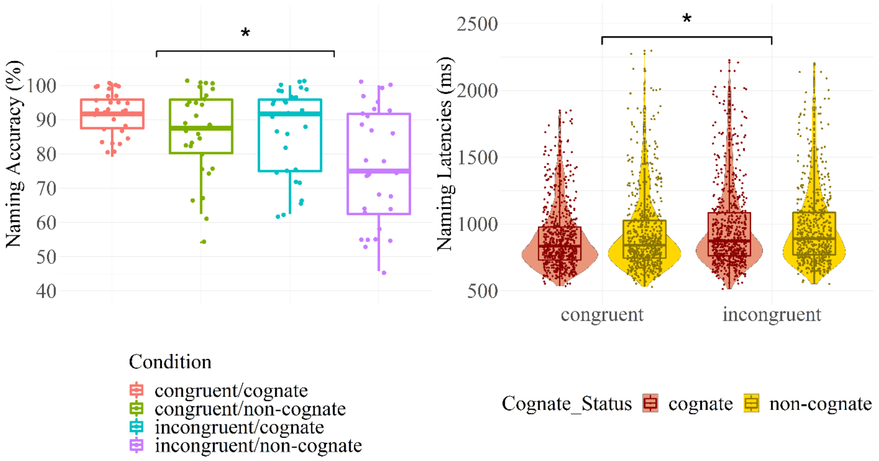
Table 3.3.1: *Mean naming accuracy and naming latencies (only correct trials included) for each condition (n = 30).*

	Naming accuracy (%)		**Naming latency (ms)**	
Condition	**Mean**	**SD**	**Mean**	**SD**
congruent/cognate	92.08	27.02	891	237
congruent/non-cognate	86.39	34.31	933	294
incongruent/cognate	87.22	33.41	971	313
incongruent/non-cognate	75.83	42.84	978	303

Naming accuracy. For naming accuracy, our model of best fit included the main effects for *gender congruency* and *cognate status*, and *subject* and *item* as random effects. Moreover, the covariates *LexTALE-Esp score* and *familiarisation phase performance* were included in the final model (Appendix 3.C). The remaining covariates target *noun gender, word length, order of acquisition, acquisition of French* and *terminal phoneme* resulted in non-convergence or singular fit and were therefore excluded from the model fitting procedure. The model of best fit was the following: naming accuracy $\sim$ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + LexTALE-Esp score + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (1|subject) + (1|item). As predicted, participants were marginally more accurate for congruent trials compared to incongruent trials with β = 0.636, *95% CI*[0.403, 1.00], z = -1.95, p = 0.052. Despite being included in the model of best fit, cognates were not significantly different from non-cognates with β = 0.757, *95% CI*[0.479, 1.20], z = -1.19, p = 0.233 (Figure 3.3.1).

Naming latencies. For naming latencies, our model of best fit included a main effect for *gender congruency* as well as a random effect for *subject* and *item* and a by-*subject* random slope for *gender congruency*. *Cognate status* did not significantly improve the model fit and was dropped from the model fitting procedure. Further, *familiarisation phase performance* was included as a covariate (Appendix 3.D). The best-fitting model was: naming latency $\sim$ gender congruency (congruent vs. incongruent) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency|subject) + (1|item). *LexTALE-Esp score, target noun gender, word length, order of acquisition, acquisition of French* and *terminal phoneme* resulted in non-convergence or singular fit and were not included in the subsequent model fitting procedure. Participants were significantly faster in naming congruent items compared to incongruent items with $\beta = 0.059$, *95% CI*[0.002, 0.116], $t = 2.04$, $p = 0.041$ (Figure 3.3.1).

Figure 3.3.1: *Mean naming accuracy by subject and condition (left) and naming latencies by condition (right) for the picture-naming task (n = 30).*



3.3.4 EEG data exclusion

The EEG data from the same participant where we lost the voice recordings was also lost during the EEG data acquisition process. Further, we determined a set of criteria to include data in the EEG analyses. First, we only included trials where the correct NP was produced. Second, only correct trials not contaminated by artefacts (valid trials) were analysed. Finally, we set the inclusion threshold for correct and valid trials at 60%. As a result, two additional data sets were excluded due to excess artefact contamination.

See Appendix 3.E for rejection rates by condition.

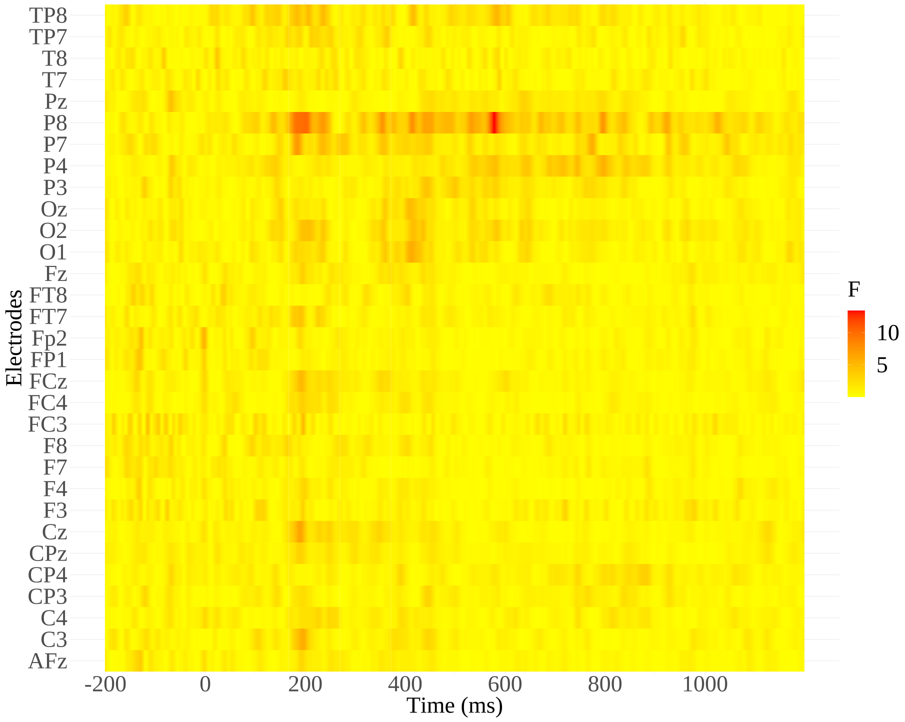
3.3.5 EEG data pre-processing

Articulatory artefacts pose a serious challenge when examining EEG data from word production tasks since they may contaminate the signal (Ganushchak, Christoffels & Schiller, 2011; Grözinger, Kornhuber & Kriebel, 1975; Porcaro, Medaglia & Krott, 2015). Therefore, we applied a vigorous pre-processing procedure to separate the signal from artefacts using BrainVision Analyser 2.2. The pre-processing procedure included the following steps: visual inspection of the raw data, re-referencing from Cz to the average mastoid electrodes (TP9 and TP10) and reusing Cz as a data channel, filtering between 0.1 Hz and 30 Hz, linear derivation of the two HEOG electrodes to form a combined channel for horizontal eye movements, interpolation of noisy channels, ocular correction ICA using VEOG and HEOG parameters, and finally, artefact rejection. After pre-processing our EEG data, we added a unique voice onset (VO) marker to every correct trial to mark the articulation onset for each participant. We then generated segments around the picture onset markers and the VO markers for each participant from -200 ms prior to picture onset to 1,200 ms after picture onset. Following segmentation, we applied baseline correction using the 200 ms prior to picture onset until picture onset. A novelty of our statistical analysis was the implementation of single-trial linear mixed effects models (LMM) for our EEG data (Frömer et al., 2018). For this, we exported all available voltage samples from valid segments for statistical analysis in RStudio (R Core Team, 2020). In contrast to more traditional EEG analyses involving ANOVAs, the assumptions for single-trial LMM do not include equal number of observations for each participant or uniform effects for each participant. Instead, single-trial LMM capture by-subject and by-item variance and therefore have superior explanatory power over more traditional ANOVAs when modelling EEG data (Baayen et al., 2008; Fröber et al., 2017).

3.3.6 EEG data analysis

After exporting our EEG data, we performed a permutation test to tentatively explore the locus of the effect of *gender congruency* and *cognate status* (collapsed into the variable *condition*) on voltage amplitudes. We used the *permutes* package (Voeten, 2019) to calculate F-values across all electrodes and the entire available time window between -200 ms and 1,200 ms with respect to stimulus onset. Visual inspection of the outcome of the permutation test revealed potential modulatory effects of condition in centro-parietal areas between 350 ms and 600 ms post stimulus onset (Figure 3.3.2). Previous literature on the distribution and time correlates of both the P300 and the N400 support this outcome (Barry et al., 2020; Koester & Schiller, 2008; Paolieri et al., 2020; Peeters et al., 2013; Polich, 2007; Roelofs, Piai, Garrido Rodriguez & Chwilla, 2016). Due to increased articulatory artefacts in EEG data closer to the participant's articulatory onset, we only explored the EEG data up to a maximum of 600 ms post-stimulus onset (Porcaro et al., 2015). On the basis of the outcomes of the permutation test and previous literature, we defined nine topographic areas for our data channels. Along the anterior-posterior axis, we defined anterior, central and posterior regions. Each region was further divided into three smaller regions: anterior left, anterior midline and anterior right regions; central left, central midline and central right regions; and finally, posterior left, posterior midline and posterior right regions. In line with previous research on the P300 and the N400, we were particularly interested in a broader topography including the posterior left, posterior midline and posterior right regions between 350 ms and 600 ms post-picture onset. These channels included TP8, P8, O2, Oz, O1, P7, TP7, CP4, P4, Pz, P3, CP3 and CPz (Appendix 3.B). In the statistical analysis, we modelled modulating effects of *condition* on *voltage amplitudes* between 350 ms and 600 ms post-stimulus onset. Furthermore, we controlled for confounding effects of *hemisphere, LexTALE-Esp score, familiarisation phase performance, target noun gender, word length, order of acquisition, acquisition of French* and *terminal phoneme.*

Figure 3.3.2: *Permutation test across all data electrodes for the time window between -200 and 1,200 ms post-stimulus onset (n = 30). Larger F-values are shown in darker colours and denote an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes.*



3.3.7 EEG data results

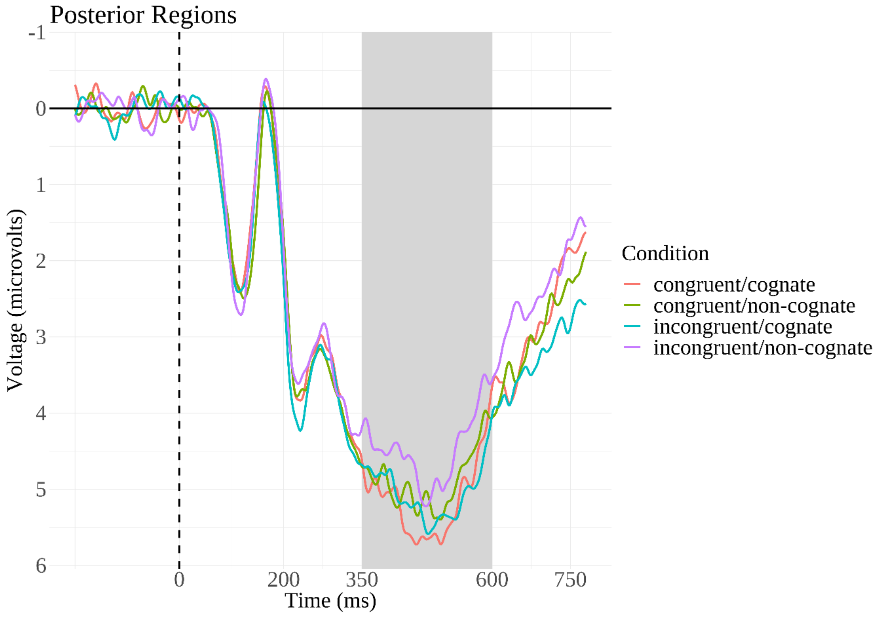
Visual inspection of the voltage amplitudes for the selected channels revealed the characteristic P1/N2 complex for early visual processing (X. Cheng et al., 2010; Eulitz et al., 2000; Misra et al., 2012; Schendan & Kutas, 2003). Further, visual inspection also revealed a positive-going wave between 350 ms and 600 ms, consistent with the topographic distribution of a P300 (Barry et al., 2020). Next, Figure 3.3.3 shows a by-condition modulation of the EEG signal between 350 and 600 ms, as tentatively suggested in the permutation test (Figure 3.3.2). Descriptively speaking, we saw the largest amplitudes for congruent cognates with $M = 5.14$ μV (*SD*

= 9.29), followed by incongruent cognates with M = 5.05 μV (SD = 9.24), congruent non-cognates with M = 4.88 μV (SD = 8.94), and finally, incongruent non-cognates with M = 4.47 μV (SD = 9.04) in the 350 ms–600 ms time window. We found no indication for an N400 effect prior to the 600 ms, after which the signal becomes increasingly noisy due to the proximity to the articulatory onset. See Figure 3.3.4 for a visualisation of the individual channels included in this analysis.

The model of best fit for voltage amplitudes included an interaction effect of *gender congruency* and *cognate status*. Further, *hemisphere* and *familiarisation phase performance* were included as covariates (Appendix 3.F). *LexTALE-Esp score, target noun gender, word length, order of acquisition, acquisition of French* and *terminal phoneme* did not significantly improve the model fit or led to overfitting. More specific to the covariate of *acquisition of French*, the model comparison between the model with and without this particular covariate yielded χ^2(1, n = 30) = 0.018, p = 0.893. We therefore dropped *acquisition of French* from the model selection procedure. In the model of best fit, *item* and *subject* emerged as random effects, with a by-*subject* random slope for the interaction effect of *gender congruency* and *cognate status*. The final model was: voltage amplitudes ~ gender congruency (congruent vs. incongruent) * cognate status (cognate vs. non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency * cognate status|subject) + (1|item). Voltage amplitudes were more positive for *congruent cognate* nouns compared to *incongruent non-cognates* with β = -0.684, *95% CI*[-1.354, -0.014], t = -2.002, p = 0.045. The difference in amplitude between the remaining conditions was not significant[3].

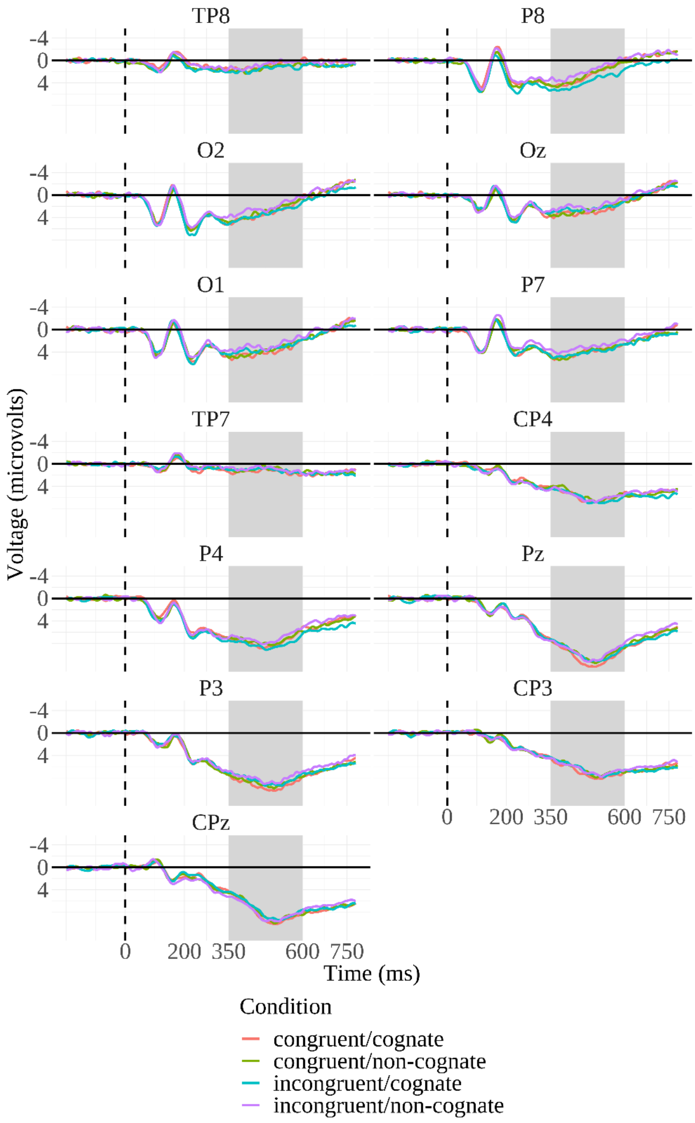
[3] As per suggestion of a reviewer, we also explored left anterior negativity (LAN) effects as a function of condition. Based on previous literature (Barber & Carreiras, 2005; Friederici et al., 1999; Hahne & Friederici, 2001; Steinhauer et al., 2009; Valente, Pinet, Alario & Laganaro, 2016; Weber-Fox & Neville, 1996), we determined channels Fp1, F3, F7, FC3 and FT7 in left anterior regions as our ROI, and the time window of interest between 300 ms and 500 ms post

Figure 3.3.3: *Voltage amplitudes by condition over time for channels TP8, P8, O2, Oz, O1, P7, TP7, CP4, P4, Pz, P3, CP3 and CPz for the picture-naming task (n = 30). The time window of interest from 350 ms to 600 ms is highlighted in grey. Negativity is plotted up.*



stimulus onset. The data show a negative-going wave peaking at around 450 ms post-stimulus onset, consistent with the topography of a delayed LAN. However, there seemed to be little difference between the conditions in terms of LAN voltage amplitudes. This was confirmed in our statistical analysis: The model that contained condition as fixed effect was not significantly better than the model that did not contain condition ($\chi^2(3, \text{n} = 30) = 0.072$, $p = 0.995$). We therefore found no evidence for a by-condition modulation of the LAN in this particular study.

Figure 3.3.4: *Voltage amplitudes by condition over time for each individual electrode included in the analysis of the picture-naming task (n = 30). The time window of interest from 350 ms to 600 ms is highlighted in grey. Negativity is plotted up.*



3.4 Discussion

The aim of this study was twofold: first, we examined CLI of the gender systems and phonological systems to obtain a better characterisation of the time course of non-native NP production. Secondly, we explored which production stage was associated with the selection of the target language in a multilingual language production configuration. We studied the gender congruency effect to highlight CLI of the gender systems during *lexical retrieval.* We predicted higher naming accuracy, shorter naming latencies, less positive P300 amplitudes and less negative N400 amplitudes for congruent compared to incongruent nouns. Critically, this would indicate that the target language was not selected prior to lexical retrieval. We also explored the cognate facilitation effect to illustrate CLI during *phonological encoding* and expected higher naming accuracy, shorter naming latencies, less positive P300 amplitudes and less negative N400 voltage amplitudes for cognates compared to non-cognates. The presence of a cognate facilitation effect would imply that the lexical entries from both the target and non-target language actively competed during *phonological encoding*, placing the locus of target language selection beyond *lexical retrieval* (Christoffels et al., 2007; Colomé, 2001; Hoshino & Kroll, 2008; Hoshino & Thierry, 2011; Pulvermüller et al., 2009; Rodriguez-Fornells et al., 2005).

In line with our predictions, we found that participants were significantly more accurate and faster at naming congruent nouns compared to incongruent nouns. These behavioural findings are important for two reasons: First, the presence of the gender congruency effect suggests CLI during gender processing. Second, the gender congruency effect also implies that the target language was not selected before *lexical retrieval.* Yet, results from the gender congruency effect alone cannot clarify whether CLI continued beyond gender processing. Therefore, we complemented these findings with results from the cognate facilitation effect. Despite a clear descriptive trend, we found no evidence for a cognate facilitation effect at the behavioural level – in contrast to previous research on the cognate status in non-native production (Acheson, Ganushchak, Christoffels & Hagoort, 2012; Christoffels et al., 2007; Peeters et al., 2013). There are two possible interpretations of this outcome: first, our late language learners did not face CLI during *phonological encoding*. As a result, there were no detectable processing differences between cognates and non-cognates. Critically, this would imply that CLI may be resolved prior to *phonological encoding* and that only the lexical entry from the target language is phonologically encoded. The second interpretation is that the behavioural measures lacked the power to pick up on a fine-grained modulation based on cognate status. Our EEG data are able to discriminate between these two possible interpretations.

Despite clear evidence for a P300 effect, we did not find evidence for an N400 effect in the time window of interest. This is somewhat surprising given that previous research linked the N400 to language co-activation, and to the neural correlates of the gender congruency effect and the cognate facilitation effect (P. Chen et al., 2017; Paolieri et al., 2020). However, studies also showed a reduced or delayed N400 in speakers with lower proficiency levels (Heidlmayr et al., 2021; Midgley et al., 2009; Weber-Fox & Neville, 1996). Therefore, the N400 effect may have been absent, or delayed and masked by articulatory artefacts.

Regarding the P300, its topographic characteristics are in line with a P300 component in the time window between 350 ms and 600 ms, more specifically a P3b component (Barry et al., 2020; Hruby & Marsalek, 2003; Polich, 2007; Squires et al., 1975). The P300, in particular the P3b, has been linked to classical inhibitory tasks as well as inhibitory tasks combined with a linguistic task (Bosma & Pablos, 2020; Eriksen & Eriksen, 1974; Jiao et al., 2020; Pereira Soares et al., 2019). Critically, it was proposed to reflect general cognitive mechanisms such as inhibition, conflict resolution and cognitive interference, and more recently the recruitment and allocation of attentional resources and working memory load (Barker & Bialystok, 2019; González Alonso et al., 2020; Neuhaus, Urbanek et al., 2010; Polich, 2007, 2012; Wu & Thierry, 2013). In order to successfully produce the correct NP in the target language, speakers not only had to go through the multi-stage process of language production, but had to simultaneously mitigate CLI effects between the target and non-target language. Here, we argue that the P300 directly taps into this latter notion and that it provides an index for this ongoing conflict between the target and the non-target language. Our EEG data revealed a small, but robust by-condition modulation of P300 voltage amplitudes, reflected in the interaction effect of gender congruency and cognate status. As predicted, P300 amplitudes were significantly different for trials with high processing costs and a larger involvement of the inhibitory control system, i.e., incongruent non-cognate trials compared to congruent cognate trials. Therefore, our results suggest quantitatively different neural patterns for producing NPs subject to differential processing costs and inhibitory demands. More importantly, this modulation of P300 amplitudes appeared to last until 600 ms post-stimulus onset (and possibly beyond). This notion has direct implications for the time course of non-native production because it provides a clear time frame for the cognitive mechanisms underlying the mitigation of CLI.

Our EEG results were indicative of the following: first, our speakers faced CLI both during the processing of gender and cognate status. Secondly, in line with the findings by Hoshino and Thierry (2011) on the locus of target language selection, CLI appears to continue beyond gender processing until at least *phonological encoding* in late language learners. This finding favours the interpretation that target language selection takes place after gender processing. To the best of our knowledge, this is the first study to report a P300 effect during overt non-native NP production in a paradigm that was not explicitly about inhibitory control, but instead included an implicit inhibitory control component. Similar tentative EEG results are reported by González Alonso et al. (2020) within the framework of third language acquisition of an artificial mini-grammar.

An interesting feature of the P300 effect was the elicitation of more positive amplitudes for congruent and cognate nouns compared to incongruent and non-cognate nouns. This is in contrast to our original hypothesis, where we predicted less positive amplitudes for congruent and cognate nouns compared to incongruent and non-cognate. Notably, this particular pattern of behavioural and EEG results has been previously reported in the literature in connection to the cognate facilitation effect (Acheson et al., 2012; Christoffels et al., 2007). For example, Acheson et al. (2012) used an overt picture-naming task with unbalanced German-Dutch speakers to study conflict monitoring during bilingual language production with respect to the Error-Related Negativity (ERN). They found faster naming latencies for cognates compared to non-cognates. However, this was linked to more negative amplitudes for cognates from about 150 ms post-stimulus onset at the FCz electrode. Furthermore, Jiao et al. (2020) measured EEG during a picture-naming task combined with a flanker task in unbalanced Chinese-English bilinguals. Their ERP results showed more positive P300 amplitudes for congruent compared to incongruent flankers in centro-parietal regions, while response times for congruent flanker trials were shorter compared to incongruent flanker trials; see also Bosma and Pablos (2020). These results mirror those from our study. On the other hand, studies have also supported the more traditional notion of faster response times or shorter naming latencies in combination with smaller ERP amplitudes for cognates compared to non-cognates (Comesaña et al., 2012; Peeters et al., 2013; Strijkers et al., 2010; Xiong et al., 2020) and smaller P300 amplitudes for congruent trials in the flanker task (Wu & Thierry, 2013). Therefore, our behavioural and EEG patterns are by no means unusual. Instead, they suggest a clear involvement of inhibitory control and acutely reflect the critical processes linked to successful non-native NP production. We propose that this study highlights the significance of the P300 as an index for the cognitive processes underlying the mitigation of CLI and the selection of the target language. Nevertheless, given the scarcity in terms of research, the directionality of the P300 effect elicited during non-native NP production warrants closer inspection in the future.

Taken together, we found traceable effects of CLI both at the behavioural and at the neural level, establishing CLI as a significant modulator of the time course of non-native NP production. This has implications for the time course of the production processes in non-native NP production, as reflected in naming accuracy, naming latencies and ERP patterns with respect to gender processing and phonological processing. CLI acts both as a facilitator and a hindrance during the production process: on one hand, there is a processing advantage for congruent and cognate nouns. On the other hand, this appears to be less the case for incongruent and non-cognate nouns. Moreover, our findings suggest that late language learners not only face CLI during early production stages and *lexical retrieval*, but possibly also during later phonological processing stages of *phonological encoding*. In turn, this implies that lexical entries from both the target and non-target language are selected for phonological processing, thereby shifting the locus of target language selection until *phonological encoding* or after it. Arguably, this highlights the complexity of non-native production processes compared to the production process in native-like speakers. Given the design of our study, we cannot exclude the possibility that our speakers resolve CLI between target and non-target language at even later production stages, e.g., the *phonetic encoding* stage. Yet, our findings have important implications for characterising the theoretical and neural underpinnings of the time course of non-native production processes, in particular for speakers with intermediate proficiency levels. Further, our findings add novel evidence to the debate about the locus of target language selection in late language learners.

3.4.1 Conclusions and future directions

In this study, we found traceable CLI effects at the behavioural and the neural level. More specifically, speakers faced CLI during gender processing and during phonological processing, which in turn impacted the time course of non-native production. In terms of the locus of target language selection, our findings suggested that the target and non-target language remained active at least until *phon-*

ological encoding. Our findings have important theoretical implications for the conceptualisation of non-native production mechanisms, and warrant further exploration with regard to subsequent production stages and the exact involvement of inhibitory control in non-native NP production. Finally, we argue that there should be an increased focus on both the P300 component as an index of CLI and lower proficiency levels in studies on non-native NP production.

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 are grateful to the members of the neurolinguistics lab at the University of Konstanz, in particular Carsten Eulitz, Oleksiy Bobrow, and the research assistants Charlotte Englert, Khrystyna Oliynyk, Zeynep Dogan, Fangming Zhang and Anna-Maria Waibel. We also thank Theo Marinis and Maria del Carmen Parafita Couto for their feedback at the initial stages of the study, the audience of SNL2020 and COM2021, Jason Rothman and the POLAR lab members, and Julian Karch for his consultation on the statistical analyses. The statistical analyses were performed using the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. Finally, we are thankful to our participants for dedicating their time to our 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/j9ys7/?view_only=ae027ce0d2194899bdab249b1725211b`

Citation diversity statement

To shed light on the systematic under-representation of work by female scientists and scientists identifying as members of a minority compared to the papers published in the field, we included a citation diversity statement (Dworkin et al., 2020; Rust & Mehrpour, 2020; Torres et al., 2020; Zurn et al., 2020). For this, we classified the first and last author in each paper from our reference list based on their preferred gender (wherever this information was available). Our reference list consisted of 26% woman/ woman authors, 38% man/ man, 21% woman/ man and finally, 13% man/ woman authors. 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 (Dworkin et al., 2020). A clear limitation of this classification is the broad binary woman/man distinction. However, with the routine addition of the preferred gender to personal and academic websites, we are confident that this is a temporary limitation with a tangible solution.

Appendix

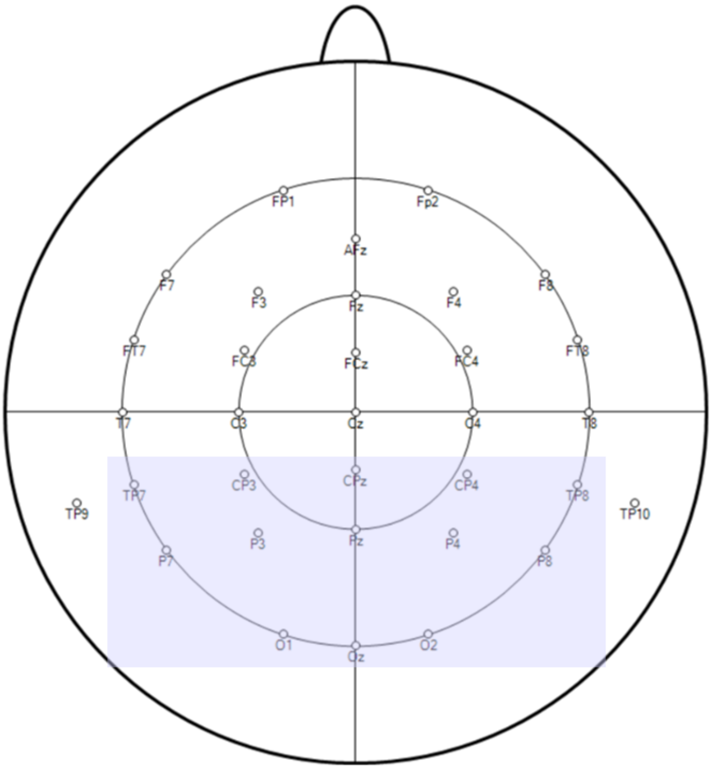
3.A Linguistic profile: German-Spanish group

Table 3.A.1: *Overview of the native and non-native languages acquired by the participants of the current study (N = 33) according to the LEAP-Q (Marian et al., 2007).*

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

3.B EEG montage

Figure 3.B.1: *Electrode positions following a 10/20 montage. Electrodes included in the analysis are highlighted in purple.*



3.C Model parameters: naming accuracy

Table 3.C.1: *Model of best fit for naming accuracy, including estimated means, confidence intervals errors and z-values (n = 30). Note that effect estimates are reported as log odds.*

Formula: naming accuracy ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + LexTALE-Esp score + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (1|subject) + (1|item)

Term	**Odds ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	0.497 [0.250, 0.990]	-1.99	0.047
Gender congruency [incongruent]	0.636 [0.403, 1.00]	-1.95	**0.052**
Cognate status [non-cognate]	0.757 [0.479, 1.20]	-1.19	0.233
LexTALE-Esp score	1.02 [1.01, 1.04]	2.72	0.007
Familiarisation phase performance [one correct]	7.26 [4.58, 11.50]	8.44	< 0.001
Familiarisation phase performance [two correct]	30.50 [18.85, 49.37]	13.91	< 0.001
Familiarisation phase performance [three correct]	71.01 [42.32, 119.15]	16.14	< 0.001
Random effects			
σ^2	3.29		
$\tau_{00 Item}$	0.66		
$\tau_{00 Subject}$	0.53		
ICC	0.27		
$N_{Subject}$	30		

N_{Item}	96
Observations	2,880
Marginal R^2/ Conditional R^2	0.341/0.516

3.D Model parameters: naming latencies

Table 3.D.1: *Model of best fit for naming latencies, including estimated means, confidence intervals errors and z-values (n = 30). Note that effect estimates are reported in seconds.*

Formula: naming latency ∼ gender congruency (congruent vs. incongruent) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency|subject) + (1|item)

Term	Estimate [95% CI]	t-value	p-value
(Intercept)	1.43 [1.32, 1.55]	25.04	< 0.001
Gender congruency [incongruent]	0.059 [0.002, 0.116]	2.04	**0.041**
Familiarisation phase performance [one correct]	-0.180 [-0.267, -0.092]	-4.02	< 0.001
Familiarisation phase performance [two correct]	-0.419 [-0.501, -0.337]	-9.98	< 0.001
Familiarisation phase performance [three correct]	-0.495 [-0.577, -0.412]	-11.76	< 0.001
Random effects			
σ^2	0.05		
$\tau_{00 Item}$	0.00		
$\tau_{00 Subject}$	0.00		
$\tau_{11 Subject[incongruent]}$	0.00		
$\rho_{01 Subject}$	-0.14		
ICC	0.16		
$N_{Subject}$	30		
N_{Item}	96		

Observations	2,459
Marginal R^2/ Conditional R^2	0.184/0.312

3.E EEG data: by-condition trial rejection rates

Table 3.E.1: *Rejection rates for each condition for the EEG data of the picture-naming task (n = 30).*

Condition	**Rejection rate (%)**	**Rejected trials**	**Total valid trials**
congruent/cognate	3.47	23	663
congruent/non-cognate	4.02	25	622
incongruent/cognate	3.34	21	628
incongruent/non-cognate	4.95	27	546
Average per condition	3.94	24	614.75
Total across conditions		96	2459

3.F Model parameters: P300 component

Table 3.F.1: *Model of best fit for voltage amplitudes, including estimated means, confidence intervals and t-values (n = 30). Note that effect estimates are reported in μV.*

Formula: voltage amplitudes ~ gender congruency (congruent vs. incongruent) * cognate status (cognate vs. non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency * cognate status|subject) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	5.27 [4.32, 6.23]	10.84	< 0.001
Condition [incongruent/cognate]	-0.075 [-0.802, 0.652]	-0.203	0.839
Condition [congruent/non-cognate]	-0.254 [-0.916, 0.409]	-0.751	0.453
Condition [incongruent/ non-cognate]	-0.684 [-1.36, -0.014]	-2.00	**0.045**
Hemisphere [midline]	1.98 [1.96, 2.01]	170.06	< 0.001
Hemisphere [right]	-0.710 [-0.730, -0.691]	-70.30	< 0.001
Familiarisation phase performance [one correct]	-0.106 [-0.180, -0.033]	-2.85	< 0.001
Familiarisation phase performance [two correct]	0.026 [-0.043, 0.095]	0.742	0.458
Familiarisation phase performance [three correct]	-0.333 [-0.402, -0.263]	-9.35	< 0.001
Random effects			
σ^2	76.06		
$\tau_{00 Item}$	1.01		

$\tau_{00 Subject}$	5.79
$\tau_{11 Subject[incongr/cogn]}$	1.60
$\tau_{11 Subject[congr/non-cogn]}$	0.90
$\tau_{11 Subject[incongr/non-cogn]}$	0.97
$\rho_{01 Subject[incongr/cogn]}$	-0.33
$\rho_{01 Subject[congr/non-cogn]}$	-0.39
$\rho_{01 Subject[incongr/non-cogn]}$	-0.42
ICC	0.08
$N_{Subject}$	30
N_{Item}	96
Observations	3,873,870
Marginal R^2/ Conditional R^2	0.014/0.089

CHAPTER 4

Cross-language effects in comprehension and production: the case of Italian-Spanish speakers

4.1 Introduction

In Chapters 2 and 3 of this thesis, we explored cross-linguistic influence (CLI) between German and Spanish in non-native comprehension and production in German late language learners of Spanish with intermediate proficiency levels (Von Grebmer Zu Wolfsthurn et al., 2021a; Von Grebmer Zu Wolfsthurn, Pablos-Robles & Schiller, 2021b). In this chapter, we expanded on this previous work and applied the experimental design to a linguistically more similar language pair: Italian and Spanish. Both of these languages show a significant overlap in terms of morphosyntax, cognates and phonology (Serratrice, Sorace, Filiaci & Baldo, 2012; Schepens, Dijkstra, Grootjen & Van Heuven, 2013; Van der Slik, 2010). Therefore, they can be considered as linguistically more similar compared to German and Spanish. In line with the previous chapters, the primary aim of the present chapter was the following: we examined how congruency type (i.e., gender congruent or incongruent across Italian and Spanish) and cognate status (i.e., cognate or non-cognate across Italian and Spanish) influenced behavioural and electrophysiological measures of non-native comprehension and production in a syntactic violation task and in a picture-naming task. We focused on the gender congruency effect (Klassen, 2016; Morales et al., 2016) and the cognate facilitation effect (Costa et al., 2005). As discussed in the previous chapters, the gender congruency effect represents CLI at the level of gender and is typically reflected in more accurate and faster processing of gender congruent compared to incongruent nouns, e.g., [il$_M$ cane$_M$ - el$_M$ perro$_M$] *"the dog"* vs. [il$_M$ tavolo$_M$ - la$_F$ mesa$_F$] *"the table"*. On the other hand, the cognate facilitation effect represents CLI at the orthographic and the phonological level and manifests itself in more accurate and faster processing of cognates compared to non-cognates, e.g., [trattore - tractor] *"tractor"* vs. [viso - cara] *"face"* (Costa et al., 2005; Lemhöfer et al., 2008, 2004; Paolieri et al., 2020). In the current chapter, we subsequently explored the interplay between Italian and Spanish at the level of gender, and at the level of orthography and phonology.

From the perspective of non-native comprehension, we embedded our study within the broader theoretical framework of how multilingual speakers represent gender. On the one hand, the gender-integrated representation hypothesis predicts CLI of the gender systems (Bordag & Pechmann, 2007; Morales et al., 2016; Salamoura & Williams, 2007) due to shared gender systems across languages. On the other hand, the gender-autonomous representation hypothesis does not predict such an interplay between languages and proposes independent gender systems for each language (Costa et al., 2003). Similar to Chapter 2, in this chapter we studied how gender congruency influenced non-native syntactic processing in noun phrases such as [el$_M$ perro$_M$] *"the dog"*. We additionally explored cognate status as a potential modulator of syntactic processing in order to examine a possible interaction effect alongside gender congruency. Moreover, we aimed to characterise the neural correlates of syntactic processing in late language learners: some studies suggested that language learners may be less sensitive to syntactic violations compared to more proficient speakers, as reflected in smaller event-related potentials or ERPs (Foucart & Frenck-Mestre, 2012; Gillon-Dowens et al., 2010; Hahne, 2001; Tokowicz & MacWhinney, 2005; Weber-Fox & Neville, 1996; Zawiszewski & Laka, 2020), but see Von Grebmer Zu Wolfsthurn et al. (2021a) in Chapter 2 for contrasting results. Results from the study conducted in Chapter 2 suggested a separate influence of gender congruency and cognate status in German-Spanish and also provided evidence for the gender-integrated representation hypothesis (Bordag & Pechmann, 2007; Salamoura & Williams, 2007). Critically, those previous results for German-Spanish speakers showed the ERP effect typically linked to syntactic violation processing, namely the P600 component (Steinhauer et al., 2009; Von Grebmer Zu Wolfsthurn et al., 2021a). However, it is unclear whether these results are also applicable to a linguistically more similar language pair such as Italian and Spanish.

From the perspective of non-native production, the primary motivation behind both the current study and previous work in Chapter 3 (Von Grebmer Zu Wolfsthurn et al., 2021b) was to examine the effect of CLI on the time course of non-native production. We again focused on the gender congruency effect and the cognate facilitation effect. Both features were shown to significantly modulate non-native production (Lemhöfer et al., 2008; Midgley et al., 2011; Paolieri et al., 2020; Peeters et al., 2013). By extension, this implied that speakers experienced CLI during non-native production. Similar to Chapter 3, the current chapter was concerned with the modulation of the time course of non-native production in light of CLI and the locus of target language selection. We used the LRM model (Levelt et al., 1999) of single word production to test two contrasting accounts of target language selection: one theoretical account previously suggested that the target language was selected before or upon lexical retrieval (Hermans et al., 1998; Lee & Williams, 2001). In contrast, a second account postulated that the target language was selected after lexical retrieval (Christoffels et al., 2007; Colomé, 2001; Hoshino & Thierry, 2011). To discriminate between these two accounts, in this chapter we used the gender congruency effect to examine the lexical retrieval stage, and the cognate facilitation effect to explore the phonological encoding stage described in the LRM model (Levelt et al., 1999). Results from the

German-Spanish speakers from Chapter 3 suggested first, that CLI was traceable at the level of gender and cognates; and second, that CLI continued beyond lexical retrieval into phonological encoding. Moreover, we provided evidence for the P300 ERP component as an index for the mitigation of CLI. Yet, it remains unclear whether these findings were applicable to speakers of highly similar languages (e.g., Italian and Spanish).

Taking previous chapters from this thesis as its starting point, the current chapter extends on the findings reported in Chapters 2 and Chapter 3 and applies an almost identical theoretical framework and methodology to a linguistically highly similar language pair, namely Italian and Spanish. The tasks used in this study were modelled after previous work by Von Grebmer Zu Wolfsthurn et al. (2021a) for the syntactic violation task, and after Von Grebmer Zu Wolfsthurn et al. (2021b) for the picture-naming task. In the next sections, we separately discuss the syntactic violation task and the picture-naming task, followed by a more general discussion.

4.2 Syntactic violation task

4.2.1 Research questions

For the syntactic violation task, the research questions were identical to the ones outlined in Von Grebmer Zu Wolfsthurn et al. (2021a). The questions were as follows: first, whether there was an effect of gender congruency (congruent vs. incongruent) and cognate status (cognate vs. non-cognate) on processing syntactic violations; second, whether there was a P600 effect in late language learners; and finally, whether the P600 effect was modulated by gender congruency and cognate status. Therefore, our focus in terms of the neural correlates of CLI in non-native comprehension was the P600 effect, see Chapter 2 and Von Grebmer Zu Wolfsthurn et al. (2021a).

Hypotheses

We first predicted that participants would be more accurate and faster during non-violation trials compared to violation trials. Second, we also predicted participants to be more accurate and faster for congruent and cognate trials compared to incongruent and non-cognate trials. Critically, we expected *gender congruency* and *cognate status* to have a joint effect on syntactic violation processing: we expected participants to be most accurate and fastest for congruent cognates compared to incongruent non-cognates, in line with the hypotheses in Von Grebmer Zu Wolfsthurn et al. (2021a). We empirically tested this by aggregating *gender congruency* and *cognate status* into the variable *condition* and by including an interaction term for *violation type* and *condition* in the statistical model.

For the EEG data for the syntactic violation task, we expected larger voltage amplitudes in centro-parietal regions around 500 ms to 900 ms post-stimulus onset for violation trials compared to non-violation trials. This would be evidence for a classical P600 effect (Von Grebmer Zu Wolfsthurn et al., 2021a). Moreover, we expected an interactive effect of *gender congruency* and *cognate status* on P600 effect amplitudes. More specifically, we predicted a larger P600 effect for congruent cognates compared to incongruent non-cognates. This would not only be evidence for CLI at the level of gender and cognates, but also that these two linguistic features have a joint influence on the neural correlates underlying syntactic violation processing. Similar to the behavioural analysis, we tested for this by including an interaction effect for *violation type* and *condition* in our statistical analysis.

4.2.2 Methods

Prior to the experiment, participants were asked to complete the Language Proficiency and Experience Questionnaire, LEAP-Q (Kaushanskaya, Blumenfeld & Marian, 2020; Marian et al., 2007). This questionnaire was designed to establish a detailed picture about the linguistic profile of each participant with respect to the acquired languages. During the experiment, participants completed the LexTALE-Esp (Izura et al., 2014), a lexical decision task to measure vocabulary size in Spanish. Participants then alternated between completing the syntactic violation task, as described in Von Grebmer Zu Wolfsthurn et al. (2021a), and the picture-naming task, as described in Von Grebmer Zu Wolfsthurn et al. (2021b). The picture-naming task is described separately in section 4.3. We recorded participants' EEG during both of these latter tasks.

Participants

Participants were 33 native Italian late learners of Spanish living in Barcelona and tested at Pompeu Fabra University. Twenty-four of our participants were female, and participants' mean age was 27.12 years (*SD* = 4.08). Recruitment criteria were the following: right-handedness, no language, reading or psychological impairments, no second language learnt before five years of age, between 18 and 35 years old and acquisition of Spanish after fourteen years of age. Moreover, participants who had lived in a Spanish-speaking country for more than one year were not included in this study. Critically, participants had to have a B1/B2 level of Spanish (Council of Europe, 2001). This proficiency level was established first, by recruiting participants directly from Spanish language courses for this specific level; and second, by using the measures obtained from the LEAP-Q and the LexTALE-Esp as an additional proxy indicator for their proficiency in Spanish. Participants acquired Spanish at the age of M = 23.94 (*SD* = 5.07). They reported oral fluency in Spanish at the age of M = 24.89 (*SD* = 4.48). Reading onset age was M = 24.36 (*SD* = 4.91) and reading fluency was reached at the age of M = 24.24 (*SD* = 4.82). On average, participants had spent M = 0.46 years (*SD* = 0.34) in a Spanish-speaking country. Their self-rated mean speaking proficiency was M = 6.09 (*SD* = 1.76), their comprehension proficiency M = 7.26 (*SD* = 1.67) and their reading proficiency M = 7.36 (*SD* = 1.48) on a scale from one to ten, with ten being maximally proficient. Two participants acquired Spanish as their first foreign language, eighteen participants as their second, ten participants as their third and three participants as their fourth foreign language. Further, thirteen participants reported Spanish as their current second most dominant language after Italian, fourteen as their third, five as their fourth and one participant as their fifth most dominant language. See Appendix 4.A for an overview of the other languages participants reported in the LEAP-Q.

Tasks and stimuli

We used the original LexTALE-Esp by Izura et al. (2014), but excluded three stimuli words for both groups due to overlap with the stimuli from the syntactic violation task. The critical manipulation in this task was *condition* (word vs. pseudoword). We then selected stimuli from the MultiPic database (Duñabeitia et al., 2018) and the Spanish Frequency Dictionary (Davies & Davies, 2017). Both databases included common nouns and pictures of objects in Spanish. We chose highly frequent nouns and pictures where the highest percentage of the correct name of the object was provided in the norming phase. Next, each selected noun was assigned a congruency type (i.e., either *congruent* or *incongruent* across Italian and Spanish), and a cognate status (i.e., either a *cognate* or a *non-cognate* in Italian and Spanish). Cognate status was defined based on orthographic and phonological overlap, and only recognisable cognates were included as stimuli in the respective tasks. Importantly, we did not include identical cognates (e.g., il taxi - el taxi [the taxi]), plural forms of nouns (e.g., gafas [glasses]), professions, English loan words or words with multiple translation equivalents (e.g., el asno/el burro [the donkey]). We modelled the distribution of terminal phonemes of the nouns according to the natural terminal phoneme distribution in Spanish to increase the ecological validity of our stimuli (Clegg, 2011). Further, we included a balanced ratio of feminine-to-masculine nouns, and controlled for syllable length in our stimuli. See Table 4.2.1 for an example set of stimuli for the syntactic violation task. The design was a 2 x 2 x 2 fully factorial within-subjects design with *violation type* (non-violation vs. violation), *congruency type* (congruent vs. incongruent) and *cognate status* (cognate vs. non-cognate) as our three critical manipulations.

Table 4.2.1: *Example stimuli for the syntactic violation task, illustrating the three manipulations of violation type, gender congruency and cognate status.*

		non-violation	
		congruent	**incongruent**
cognate	**Italian**	il$_M$ trattore$_M$	il$_M$ grasso$_M$
	Spanish	el$_M$ tractor$_M$	la$_F$ grasa$_F$
		the tractor	*the fat*
non-cognate	**Italian**	il$_M$ cane$_M$	il$_M$ tavolo$_M$
	Spanish	el$_M$ perro$_M$	la$_F$ mesa$_F$
		the dog	*the table*
		violation	
		congruent	**incongruent**
cognate	**Italian**	il$_M$ pane$_M$	la$_F$ labbra$_F$
	Spanish	*la$_F$ pan$_M$	*la$_F$ labio$_M$
		the bread	*the lip*
non-cognate	**Italian**	il$_M$ fiume$_M$	il$_M$ viso$_M$
	Spanish	*la$_F$ río$_M$	*el$_M$ cara$_F$
		the river	*the face*

Procedure

During the experiment, participants were comfortably seated in front of a computer screen in an experimental booth. They were provided with an information sheet and gave informed consent before proceeding to the tasks, in line with the ethics guidelines at the Faculty of Humanities at Leiden University. Participants completed the LexTALE-Esp before the syntactic violation task. Both tasks were programmed in E-prime2 (Psychology Software Tools, Inc).

The procedure for the LexTALE-Esp and the syntactic violation task for each task was identical to Von Grebmer Zu Wolfsthurn et al. (2021a). For the LexTALE-Esp, participants made a lexical decision on whether or not the string on the screen was a Spanish word while accuracy was measured. Post-task, we calculated a LexTALE-Esp vocabulary scores by subtracting the percentage of yes-answers to pseudowords from the percentage of yes-answers to words. The resulting vocabulary size score (*LexTALE-Esp score*) was included as a covariate in the analysis of the syntactic violation task. For the syntactic violation task, participants were instructed to use keyboard buttons to indicate their familiarity with the noun, and to then provide a judgement as accurately and fast as possible of whether or not the presented noun phrase was correct. during this task, we measured both accuracy and response times (RTs), as well as voltage amplitudes. The two major differences to the procedure in Chapter 2 were first, that the stimuli differed significantly for the Italian-Spanish group due to constraints by gender congruency type and cognate status; and second, that the information sheet, the consent form, the oral and written task instructions and the debrief were provided in the participants' native language Italian.

EEG recordings

EEG data were collected from 32 active Ag/AgCl channels at 500 Hz configured in a 10/20 montage from BrainProducts. Channel FT9 was placed underneath the left eye to record the vertical electrooculogram (VEOG), and channel FP10 on the outer canthus of the left eye to record the horizontal electrooculogram (HEOG). We positioned the ground electrode on the participants' right cheek. The original reference channel was FCz, and the impedances for all channels were configured to be below 10kΩ for optimal EEG signal conductivity. EEG data was recorded using BrainVision Recorder (BrainProducts GmbH).

4.2.3 Results

Behavioural data analysis

The behavioural data analysis procedure was identical as in Von Grebmer Zu Wolfsthurn et al. (2021a). Moreover, we included the same participants in the behavioural analyses and in the EEG analyses, see the next section. Subsequently, 29 of the 33 participants were included for the analysis of this task. To model accuracy and

RTs, we employed a linear mixed effects model (LMM) approach (Baayen et al., 2008) in R via RStudio (R Core Team, 2020) using the *lme4* package (Bates et al., 2020). Our model selection procedure was as follows: first, we separately specified a theoretically plausible maximal model for accuracy and for RTs. More concretely, we specified a generalised linear mixed effects model (GLMM) with a binomial distribution to model accuracy for familiar trials, and a GLMM with a gamma distribution and the identity link function to model positively skewed RTs for familiar and correct trials (Lo & Andrews, 2015). Each maximal model included the interaction term for *violation type* and *condition*, the variable aggregating both *congruency type* and *cognate status*. Further, we included several covariates, such as *LexTALE-Esp score*, *terminal phoneme* of the target word, *target noun gender* and *order of acquisition of Spanish*. Moreover, we included random intercepts for each *participant* and individual *item*, as well as by-participant random slopes for the effect of *condition*. In a second step, in the case of non-convergence or singular fit of the model, we first simplified the random effects structure. We then tested for the statistical relevance of the covariates and interaction effects by systematically comparing models with an without a particular term by using the *anova()* function. For each model, we performed model diagnostics to assess the goodness of fit via the *DHARMa* package (Hartig, 2020). Absolute test-statistics larger than 1.96 were interpreted as being statistically significant at $\alpha = 0.05$ (Alday et al., 2017). Treatment coding was our default contrast.

EEG data exclusion

Our inclusion criteria for the EEG analysis for this task were the following: we only included trials where participants had indicated familiarity with the noun prior to the experimental trial. Second, trials which were incorrectly identified as violations or non-violations were excluded, as were trials containing artefacts. Taking these criteria together, we only included participants with more than 60% of (valid) trials left in their data. Subsequently, we excluded four participants from the EEG analysis, adding to a total of 29 included datasets. The same datasets were included in the behavioural data analysis.

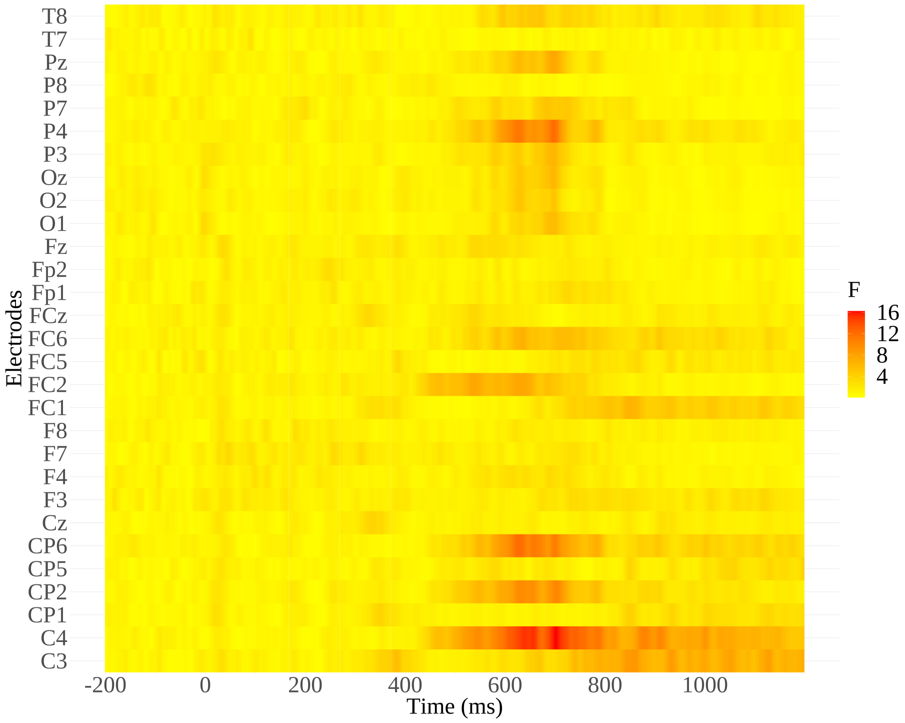
EEG data pre-processing

We thoroughly pre-processed the EEG data to increase the signal-to-noise ratio and to minimise noise related to artefacts, for example jaw muscle movement, eye blinks, or other external interferences (Ganushchak, Christoffels & Schiller, 2011; Porcaro et al., 2015). For this, we used BrainVision Analyzer (BrainProducts GmbH). As the first pre-processing step, we applied the average of the mastoid electrodes TP9 and TP10 as the new references. In this, FCz was reused as a regular data channel. Next, we filtered the data using a high-pass filter of 0.1 Hz, and a low-pass filter of 30 Hz. Channel interpolation was performed if deemed appropriate given the quality of the surrounding channels. We then performed residual drift detection in preparation for ocular independent component analysis (ICA) for blink correction. After this step, we performed an artefact search across all data channels. The criteria for artefact detection were the following: for gradient, the maximal voltage step was defined as 50 μV/ms, the maximal difference in 100 ms - intervals as 200 μV, the maximal amplitude as ± 200 μV, and the lowest allowable amplitudes in 100 ms - intervals as 0.5 μV. In a final step, we segmented our EEG signal on the basis of the stimulus onset, thereby generating segments between -200 ms prior and 1,200 ms post-stimulus. Segments were then baseline-corrected using the activity in the 200 ms prior to the stimulus onset. We exported all available valid segments for each channel and participant. As described above, we defined valid trials as those trials where participants had indicated familiarity with the noun, provided a correct response and were artefact-free for the statistical analysis (Christoffels et al., 2007).

EEG data analysis

We performed a cluster-based permutation analysis on the exported EEG data to determine our region of interest (ROI). For this, we used the *permutes* package (Voeten, 2019) in R. This analysis is particularly powerful because it reveals potentially significant differences in voltage amplitudes by condition for each channel. The outcome is measured in F-values, with larger F-values indicating an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes. The permutation analysis suggested channels *C4, CP2, CP6, P3, P4, P7, P8* and *Pz* in centro-parietal regions as ROI in the time-window between 500 ms and 800 ms. We then performed the statistical analysis using these channels as our ROI. See Figure 4.2.1 for the permutation analysis outcome for this task.

Figure 4.2.1: *Permutation test outcome for the syntactic violation task (n = 29). Larger F-values are shown in darker colours and denote an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes.*



Similar to the behavioural analyses, we followed a single-trial LMM approach for the EEG data analysis using the *lme4* package (Bates et al., 2020). We based this approach on work by Frömer et al. (2018). In this, we examined voltage amplitudes from all exported valid segments. The modelling procedure was as follows: first, we specified a theoretically feasible maximal LMM model using a Gaussian distribution. The maximal model consisted of the interaction between *violation type* and *condition* (which combined *congruency type* and *cognate status* into a single variable), the covariates *channel*, *LexTALE-Esp score*, *terminal phoneme* of the target word, *target noun gender* and *order of acquisition of Spanish*, random intercepts for each participant and individual item, and finally, correlated by-participant random slopes for the effect of *violation type* and *condition.* We did not specify an interaction random slope with *violation type* to avoid over-parametrisation (Matuschek et al., 2017). Next, we thoroughly evaluated the goodness of fit of this model using the *DHARMa* package (Hartig, 2020). In the event of non-convergence, we simplified the random effects structure, followed by the systematic evaluation of the contribution of the covariates in the fixed effects structure. We then performed model comparisons using the *anova()* function to establish our model of best fit. We again used treatment coding as our contrast, and absolute test-statistics larger than 1.96 were interpreted as showing a significant effect on voltage amplitudes at $\alpha = 0.05$ (Alday et al., 2017).

Data results

We first calculated descriptive statistics for mean accuracy and RTs. See Table 4.2.2 for mean accuracy per condition and Table 4.2.3 for mean RTs per condition.

Table 4.2.2: *Mean accuracy for each condition for the syntactic violation task (n = 29).*

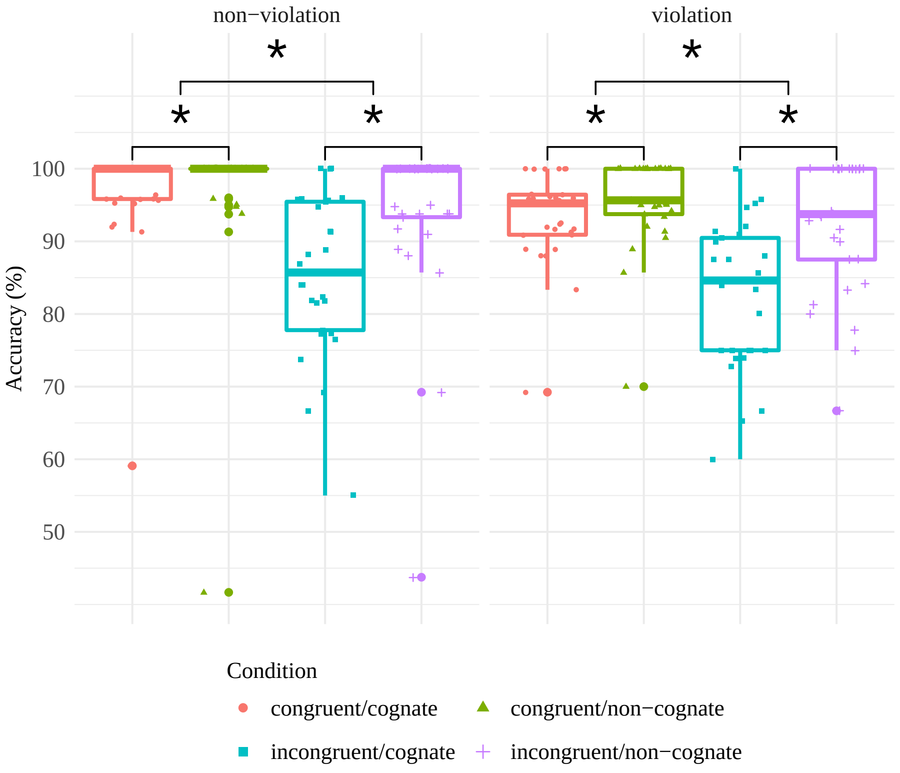
Condition	**Mean accuracy (%)**	**SD**
non-violation/congruent/cognate	96.65	18.01
non-violation/congruent/non-cognate	96.47	18.47
non-violation/incongruent/cognate	85.97	34.76
non-violation/incongruent/non-cognate	93.72	24.28
violation/congruent/cognate	93.17	25.25
violation/congruent/non-cognate	95.56	20.63
violation/incongruent/cognate	82.75	37.81
violation/incongruent/non-cognate	91.99	27.17

Table 4.2.3: *Mean RTs for each condition for the syntactic violation task (n = 29).*

Condition	**Mean RTs (ms)**	**SD**
non-violation/congruent/cognate	816.76	328.44
non-violation/congruent/non-cognate	817.31	347.17
non-violation/incongruent/cognate	931.11	452.49
non-violation/incongruent/non-cognate	854.38	377.15
violation/congruent/cognate	953.55	401.33
violation/congruent/non-cognate	929.73	356.10
violation/incongruent/cognate	1049.37	489.65
violation/incongruent/non-cognate	991.48	439.02

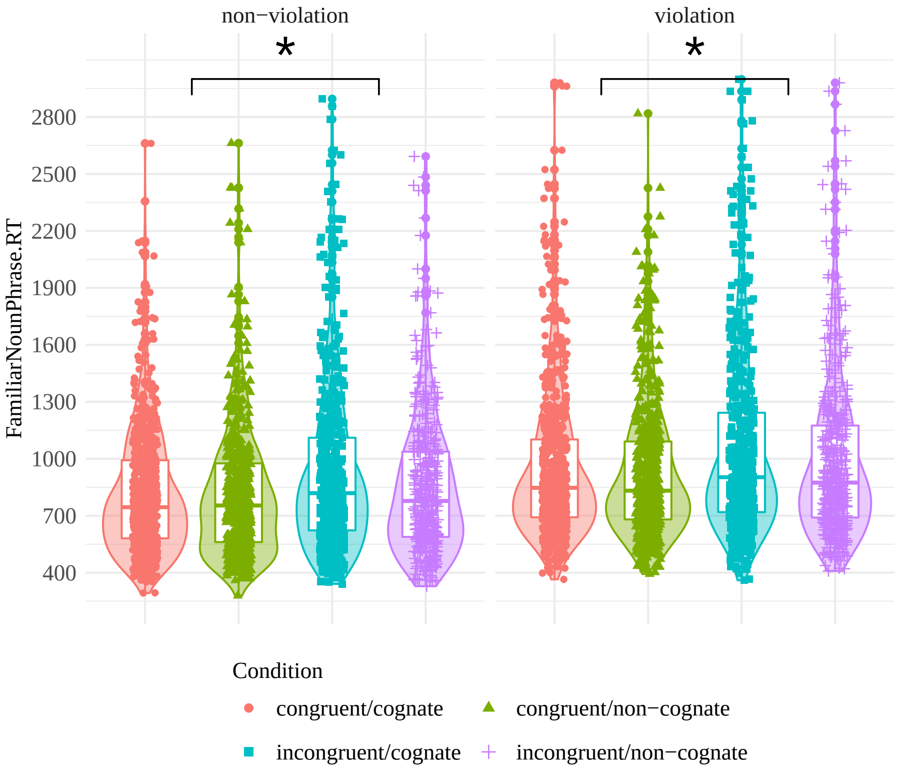
Accuracy. For accuracy, the maximal model did not converge and was subsequently simplified. The simplified model, which included the interaction between *violation type* and *condition*, did not yield a better model fit compared to a model without the interaction $\chi^2(1, \text{n} = 29) = 0.894$, $p = 0.827$. We then only included a main effect for *violation type* and an interaction effect between *gender congruency* and *cognate status* in the subsequent model and compared it with a model without this interaction term but only main effects. This comparison yielded a better model fit for the latter model, with $\chi^2(1, \text{n} = 29) = 1.28$, $p = 0.258$. Therefore, our model of best fit included a main effect of *violation type*, *gender congruency* and *cognate status*, as well as correlated random slopes for *gender congruency* and *cognate status* for the random effect of *participant.* Moreover, *item* was included as a random effect. None of the covariates significantly contributed to a better model fit and were therefore not included in the best-fitting model. Taken together, the model of best fit was as follows: accuracy ∼ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + (gender congruency + cognate status|participant) + (1|item). Participants were more accurate for *congruent* compared to *incongruent* trials with $\beta =$ 0.321, *95% CI*[0.195, 0.531], z = -4.44, $p < 0.001$, and for *non-cognates* compared to *cognates* with $\beta = 2.11$, *95% CI*[1.31, 3.40], $z = 3.07$, $p = 0.002$. Contrary to our predictions, we did not find evidence that our participants were more accurate for *non-violation* trials compared to *violation* trials with $\beta = 0.679$, *95% CI*[0.449, 1.03], z = -1.84, $p = 0.066$. See Appendix 4.B for model parameters and Figure 4.2.2 for a visualisation of the accuracy results. Note that model parameters are reported as odds ratios.

Figure 4.2.2: *Visualisation of mean accuracy for each condition for the syntactic violation task (n = 29). The brackets indicate statistical significance between conditions.*



Response times. For RTs, the maximal model yielded non-convergence and was therefore simplified. Our best-fitting model for RTs included main effects for *violation type*, *gender congruency* and *cognate status*, and an interaction effect for *gender congruency* and *cognate status*. Further, the model also included correlated by-participant random slopes for *gender congruency* and *cognate status* in addition to a random effect for *item*. The best-fitting model was the following: RTs ~ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) * cognate status (cognate vs. non-cognate) + (gender congruency + cognate status|participant) + (1|item). Participants were significantly faster for *non-violation* trials compared to *violation trials* with $\beta = 112.80$, *95% CI*[102.54, 123.06], $t = 21.55$, $p < 0.001$. Further, the main effect for *gender congruency* was significant with $\beta = 123.42$, *95% CI*[110.16, 136.69], $t = 18.24$, $p < 0.001$ for *congruent* compared to *incongruent* trials. The main effect of *cognate status* was not significant with $\beta = -6.83$, *95% CI*[-15.95, 2.29], $t = -1.47$, $p = 0.142$. The interaction effect for *gender congruency* and *cognate status* was significant, with $\beta = -68.33$, *95% CI*[-79.05, -57.60], $t = -12.49$, $p < 0.001$. See Appendix 4.C for model parameters and Figure 4.2.3 for a visualisation of the RT results.

Figure 4.2.3: *Visualisation of mean RTs for each condition for the syntactic violation task (n = 29). The brackets indicate statistical significance between conditions.*



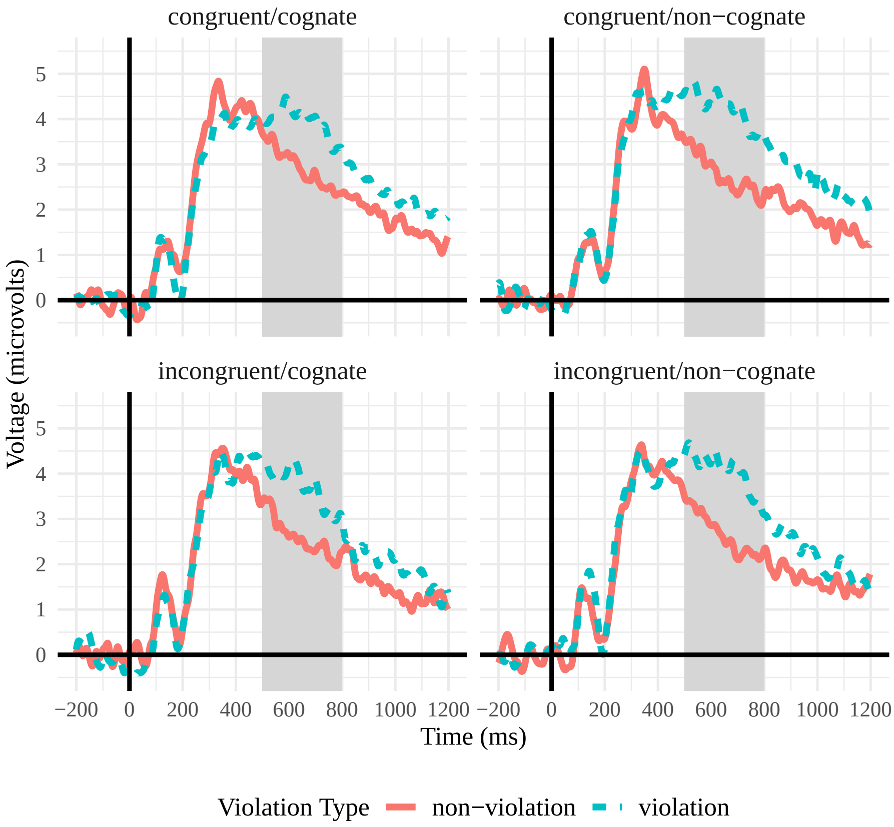
Voltage amplitudes. We calculated mean voltage amplitudes for each condition for the time window between 500 ms and 800 ms post-stimulus onset for the selected channels in centro-parietal regions (Table 4.2.4).

Table 4.2.4: *Voltage amplitudes by condition for the time window of interest (500 ms - 800 ms) for channels C4, CP2, CP6, P3, P4, P7, P8 and Pz for the syntactic violation task (n = 29).*

Condition	**Mean voltage (μV)**	**SD**
non-violation/congruent/cognate	1.99	8.51
non-violation/congruent/non-cognate	2.16	8.23
non-violation/incongruent/cognate	1.78	8.06
non-violation/incongruent/non-cognate	1.92	7.93
violation/congruent/cognate	2.61	8.27
violation/congruent/non-cognate	2.81	8.83
violation/incongruent/cognate	2.24	8.06
violation/incongruent/non-cognate	2.47	8.37

Following the analysis procedure outlined above, the model of best fit for voltage amplitudes included the main effects for *violation type*, *gender congruency* and *cognate status*, the covariates *channel*, *terminal phoneme* and *LexTALE-Esp score*, correlated random slopes for by-participant effects of *violation type* and *condition* and random intercepts for both *participant* and *item*. Therefore, the final model was: voltage amplitudes ∼ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + channel + terminal phoneme + LexTALE-Esp score + (violation type + gender congruency * cognate status|participant) + (1|item). See Appendix 4.D for the full model parameters. The interaction effect between *violation type* and *condition* on P600 voltage amplitudes was neither significant nor did it significantly improve the model fit. It was therefore dropped from the model-fitting procedure. Subsequently, these results did not provide evidence for a significant modulation of P600 voltage amplitudes as a function of the CLI effects. In addition, the main effects for *gender congruency* and *cognate status* in the best-fitting model were insignificant with β = -0.072, *95% CI*[-0.551, 0.406], t = -0.296, p = 0.767 for congruent compared to incongruent trials, and with β = 0.108, *95% CI*[-0.330, 0.545], t = 0.483, p = 0.629 for cognates compared to non-cognates. Nevertheless, voltage amplitudes were significantly higher for violation trials compared to non-violation trials with β = 1.48, *95% CI*[0.893, 2.08], t = 4.92, p < 0.001 (Appendix 4.D). Figure 4.2.4 visualises mean voltage amplitudes over time for each condition for our ROI.

Figure 4.2.4: *Visualisation of voltage amplitudes for each condition for channels C4, CP2, CP6, P3, P4, P7, P8 and Pz (n = 29). The time window of interest is highlighted in grey.*



4.2.4 Discussion

The aim of this study was to examine CLI in non-native comprehension in Italian late learners of Spanish. First, we explored whether and how gender congruency and cognate status affected non-native language comprehension in the context of a syntactic violation task; second, we examined whether there was evidence for a sensitivity to syntactic errors in the form of a P600 effect in Italian late language learners of Spanish; and finally, we studied whether P600 amplitudes were modulated by gender congruency and cognate status. We predicted that participants would be more accurate and faster for non-violation compared to violation trials. We also predicted participants to be most accurate and fastest at detecting congruent cognates compared to incongruent non-cognates for syntactic violations. Finally, we predicted larger P600 voltage amplitudes for violation compared to non-violation trials, as well as larger amplitudes for congruent cognates compared to incongruent non-cognates in an interaction effect with violation type.

With respect to our first research question, our behavioural results suggested that participants were significantly faster, but not more accurate for non-violation compared to violation trials. Critically, participants were more accurate and faster for congruent compared to incongruent items. This reflects the classical gender congruency effect (Klassen, 2016; Lemhöfer et al., 2008). Interestingly, participants were also more accurate for non-cognates compared to cognates. This reflects a reverse cognate facilitation effect, with higher accuracy for non-cognates instead of cognates. In contrast, we found no evidence for an effect of cognate status on RTs. Importantly, we also did not find evidence for an interaction effect between gender congruency and cognate status with violation type. This suggested that the performance in detecting syntactic violations was not significantly modulated by a joint effect of gender congruency and cognate status. Therefore, with respect to our first research question, we found some evidence for differential processing of violation vs. non-violation trials and congruent vs. incongruent trials, thereby reflecting the effects of violation and gender congruency on behavioural measures of non-native comprehension. However, as we had previously predicted, we did not find evidence that gender congruency and cognate status had a joint effect on detecting syntactic violations. In addition, we found a significant interaction effect of gender congruency and cognate status on RTs, suggesting that the effect of one factor was dependent on the other and vice versa.

Comparing the current findings with the results from Chapter 2 on German-Spanish speakers (Von Grebmer Zu Wolfsthurn et al., 2021a), the results from both groups are highly fascinating for several reasons: first, the findings across both groups are compatible in that both participant groups were faster for non-violation trials compared to violation trials. This reflects differential processing of NPs with a syntactic error vs. syntactically correct NPs. Secondly, both the Italian-Spanish speakers and the German-Spanish speakers displayed the classical gender congruency effect, with more accurate and faster processing for congruent compared to incongruent items (Bordag & Pechmann, 2007; Lemhöfer et al., 2008). In turn, this particular finding supports the gender-integrated representation hypothesis (Bordag & Pechmann, 2007; Morales et al., 2016; Salamoura & Williams, 2007). In other words, the results from the current study support a theoretical framework whereby the gender systems are shared between two languages, irrespective of the linguistic similarity of the languages. Thirdly, we found comparable effects of cognate status in both the Italian-Spanish speakers and the German-Spanish speakers: the former were more accurate for non-cognates compared to cognates, while the latter were faster for non-cognates compared to cognates. This is an interesting finding as it suggests that cognate status may have a similar reverse effect on non-native comprehension across two different language pairs in late learners. Finally, in line with findings from the German-Spanish speakers, we did not find evidence for an interaction effect of gender congruency and cognate status on detecting syntactic violations in the Italian-Spanish speakers. Overall, the behavioural results from both studies show that participants were more sensitive to an overlap in gender rather than to cognate status. They also showed that gender congruency and cognate status together had a limited effect on detecting syntactic violations in this task. However, one significant difference to the German-Spanish speakers was that gender congruency and cognate status did emerge as having an overall interaction effect on RTs in the Italian-Spanish speakers, while there was no interaction of gender congruency and cognate status with violation type in either group. Subsequently, this indicates a small joint influence of these two linguistic features on RTs in the syntactic violation task.

With respect to our second and third research question regarding the P600 effect, results from the EEG data suggested that violation trials elicited larger voltage amplitudes compared to non-violation trials. This reflected the classical P600 effect (Hahne & Friederici, 2001; Steinhauer et al., 2009) and was in line with our original hypotheses. Therefore, our Italian-Spanish speakers were indeed sensitive to syntactic violations even at moderate proficiency levels. This contrasts with previous research reporting no P600 effect for late language learners (Hahne & Friederici, 2001; Weber-Fox & Neville, 1996), but is consistent with more recent research presenting evidence for a P600 effect in late learners (S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005; Von Grebmer Zu Wolfsthurn et al., 2021a). Critically, matching the behavioural results, we did not find differential P600 effects as a function of the interaction effect between gender congruency and cognate status. In other words, we found no evidence that CLI effects modulated P600 effect sizes. This is in contrast to our predictions and suggests that the P600 effect was comparable in size across our experimental conditions.

The EEG findings from the Italian-Spanish speakers in this chapter are highly similar to the findings reported for the German-Spanish speakers. First, we again found evidence for a P600 effect for syntactic violations in our Italian-Spanish speakers. Secondly, we similarly did not find evidence that CLI effects modulated P600 effects and instead found similar P600 amplitudes across our conditions. In this, our results were compatible with our previous work because they suggested, on the one hand, a sensitivity to syntactic violations even at moderate proficiency levels. On the other hand, they did not show that CLI for gender or cognates had a significant impact on these processes per se. Therefore, our results support the gender-integrated representation hypothesis (Salamoura & Williams, 2007). Importantly, our results add to the findings from Chapter 2: independently of whether the languages were linguistically highly similar (Italian-Spanish) or less similar (German-Spanish), speakers displayed a P600 effect despite their early acquisition stages and intermediate proficiency levels. The P600 effect, in turn, was largely unaffected by the linguistic features of congruency type and cognate status we put to test in Chapter 2 and in the current chapter. Questions remain as to whether or not there was a statistical difference in terms of the P600 effect sizes across the Italian-Spanish and the German-Spanish group, and whether CLI effects differed across these two groups. These issues were the scope of Chapter 5 of this thesis, which includes a direct comparison of the data from the Italian-Spanish and German-Spanish speakers in light of language similarity.

Summary and conclusions

In this study, we explored non-native comprehension in the context of a linguistically similar language pair, Italian and Spanish. Therefore, this study represented an extension to Chapter 2, where we studied the linguistically less similar language pair German and Spanish. Behavioural results from the syntactic violation task in the current chapter showed that participants were sensitive to gender congruency across Italian and Spanish. This was reflected in a processing advantage for congruent items compared to incongruent items. In contrast, the feature of cognate status was less salient in the context of non-native comprehension. Further, we found ERP evidence for a P600 effect in our late language learners. This P600 effect, however, did not appear to be modulated by neither gender congruency nor cognate status. Generally speaking, we provided support for the gender-integrated representation hypothesis (Salamoura & Williams, 2007), and demonstrated that language learners with limited non-native proficiency displayed a clear sensitivity to syntactic violations. Our results are also highly similar to the findings reported in Chapter 2 and in Von Grebmer Zu Wolfsthurn et al. (2021a).

4.3 Picture-naming task

In this task, we tested the same participants outlined in section 4.2 on non-native production. As described above, participants completed the LexTALE-Esp (Izura et al., 2014) before the syntactic violation task and the picture-naming task. Importantly, participants alternated between first completing the syntactic violation task vs. the picture-naming task.

4.3.1 Research questions

We asked the question whether gender congruency (congruent vs. incongruent) and cognate status (cognate vs. non-cognate) modulated the behavioural and neural correlates of non-native production. Here, we focused specifically on P300 amplitudes, see Von Grebmer Zu Wolfsthurn et al. (2021b). Second, we used the gender congruency and the cognate facilitation effect to examine *when* during non-native production speakers experienced CLI between the native and the non-native language. By extension, our third research question was concerned with the locus of target language selection: when during non-native production is the target language selected? These questions are identical to the research questions in Chapter 3 and in Von Grebmer Zu Wolfsthurn et al. (2021b).

Hypotheses

Behaviourally, we were interested in whether gender congruency and cognate status would impact naming accuracy and naming latencies. More specifically, we predicted more accurate and faster naming of congruent cognate nouns compared to incongruent non-cognate nouns, in line with previous findings from Chapter 3 and Von Grebmer Zu Wolfsthurn et al. (2021b). In contrast, we did not expect significant behavioural differences in naming accuracy and latencies between congruent non-cognates and incongruent cognates.

For the EEG data, we first examined whether a P300 effect would be also elicited in the Italian-Spanish group. Moreover, we predicted P300 amplitudes to be modulated by gender congruency and cognate status: in line with the results reported in Chapter 3, we hypothesised larger P300 amplitudes for congruent cognate nouns compared to incongruent non-cognate nouns. In contrast, we expected similar amplitudes for congruent non-cognates and incongruent cognates. Therefore, we predicted the smallest P300 amplitudes for incongruent non-cognates, and largest amplitudes for congruent cognates.

4.3.2 Methods

Participants

The participants were the same as the ones described in section 4.2.2.

Tasks and stimuli

The stimuli selection procedure for the picture-naming task was identical to the procedure described for the syntactic violation task in section 4.2.2. Critically however, the stimuli differed from the stimuli used in the previous task and in Chapter 3. The design for the picture-naming task was a 2 x 2 fully factorial within-subjects design with *gender congruency* (congruent vs. incongruent) and *cognate status* (cognate vs. non-cognate) as our critical manipulations. See Table 4.3.1 for an example stimuli set.

Table 4.3.1: *Example picture stimuli for the picture-naming task, illustrating the two manipulations of gender congruency and cognate status.*

Condition	**Noun phrase**	**Italian translation**	**English translation**
congruent/ cognate	la$_F$ llave$_F$	la$_F$ chiave$_F$	*the key*
congruent/ non-cognate	la$_F$ gonna$_F$	la$_F$ falda$_F$	*the skirt*
incongruent/ cognate	el$_M$ bolso$_M$	la$_F$ borsa$_F$	*the handbag*
incongruent/ non-cognate	el$_M$ caracol$_M$	la$_F$ lumaca$_F$	*the slug*

Procedure

The task procedure was identical to Chapter 3 and Von Grebmer Zu Wolfsthurn et al. (2021b). Participants were placed in an experimental booth in front of a computer screen and presented with the stimuli pictures. During the task, we recorded participants' voice using a built-in microphone while they produced the name of the object together with the corresponding determiner as accurately and fast as possible. In this task, we measured naming accuracy, naming latencies and voltage amplitudes.

EEG recordings

The EEG recording set-up was identical to the syntactic violation task, see section 4.2.2.

4.3.3 Results

We closely followed the behavioural data analysis procedure described in section 4.2.3 and in Chapter 3, see also Von Grebmer Zu Wolfsthurn et al. (2021b). Here we also included the same participants in both the behavioural analysis and in the EEG analysis, see the next section. Therefore, we included 28 participants in the behavioural analysis. We followed a LMM approach (Baayen et al.,

2008) in R via RStudio (R Core Team, 2020) using the *lme4* package (Bates et al., 2020) to model naming accuracy and naming latencies. We used a GLMM with a binomial distribution to model naming accuracy, and a GLMM with a gamma distribution and the identity link function (Lo & Andrews, 2015) to model correctly named stimuli pictures. The model selection procedure was identical to section 4.2.3, with the following exceptions: first, the fixed effects structure of the maximal models for naming accuracy and naming latencies included an interaction effect between *gender congruency* and *cognate status* and the covariates *LexTALE-Esp score, familiarisation phase performance, order of acquisition of Spanish, target noun gender, word length* and *terminal phoneme.* Second, the random effects consisted of random slopes for the interaction effect between *gender congruency* and *cognate status* for each participant, as well as random intercepts for each *participant* and *item.*

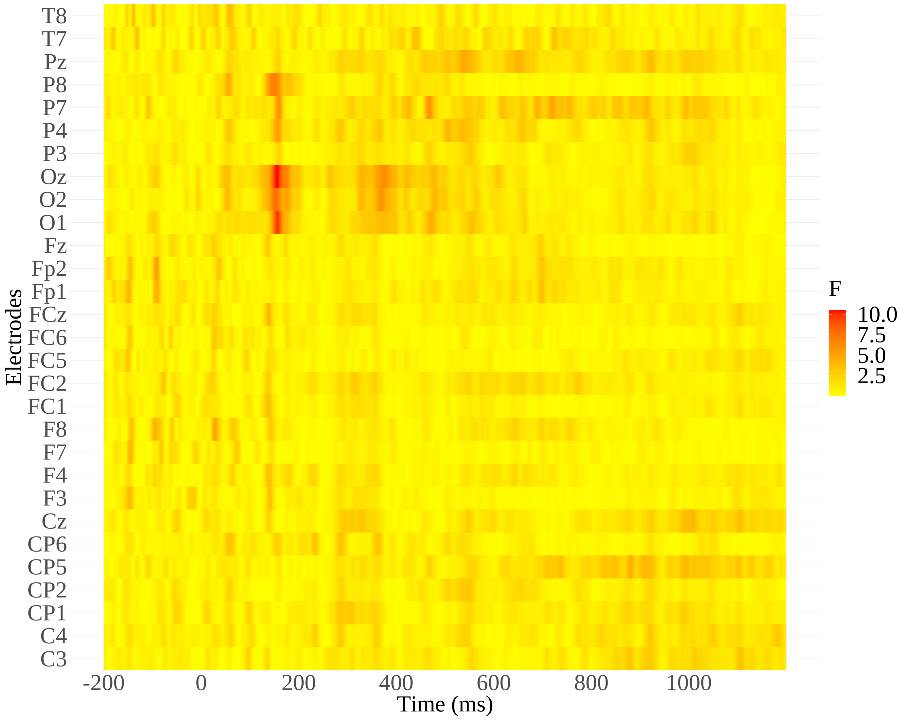
EEG data exclusion

For the EEG analysis, the inclusion criteria consisted of correctly named trials, as well as artefact-free trials and a sufficiently high data quality in terms of artefacts. In this, we excluded participants with less than 60% of (valid) trials left after the application of these criteria. Subsequently, five participants were excluded for the picture-naming task, thereby including 28 datasets in the behavioural and in the EEG analysis of this task.

EEG data analysis

The EEG data analysis procedure was modelled after section 4.2.3. Data pre-processing was especially critical as production data is often characterised by large articulatory artefacts (Grözinger et al., 1975). Therefore, we set an upper threshold of 600 ms post-stimulus onset to avoid those artefacts in our signal as much as possible. We again used a cluster-based permutation test to determine our ROI. This permutation analysis outcome suggested channels *P3, P4, P7, P8, Pz, O1, O2* and *Oz* in centro-parietal regions as ROI in the time window between 350 to 600 ms for the picture-naming task, see Figure 4.3.1.

Figure 4.3.1: *Permutation test outcome for the picture-naming task (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.*



The EEG data analysis procedure was identical as in section 4.2.3. We used a LMM approach to model voltage amplitudes. The maximal model included an interaction term for *gender congruency* and *cognate status*, as well as the covariates *hemisphere*, *LexTALE-Esp score*, *terminal phoneme* of the target word, *target noun gender*, *order of acquisition of Spanish* and *familiarisation phase performance*. Finally, we also included random slopes for the interaction effect of *gender congruency* and *cognate status* for each participant, and random intercepts for each *participant* and *item*.

Data results

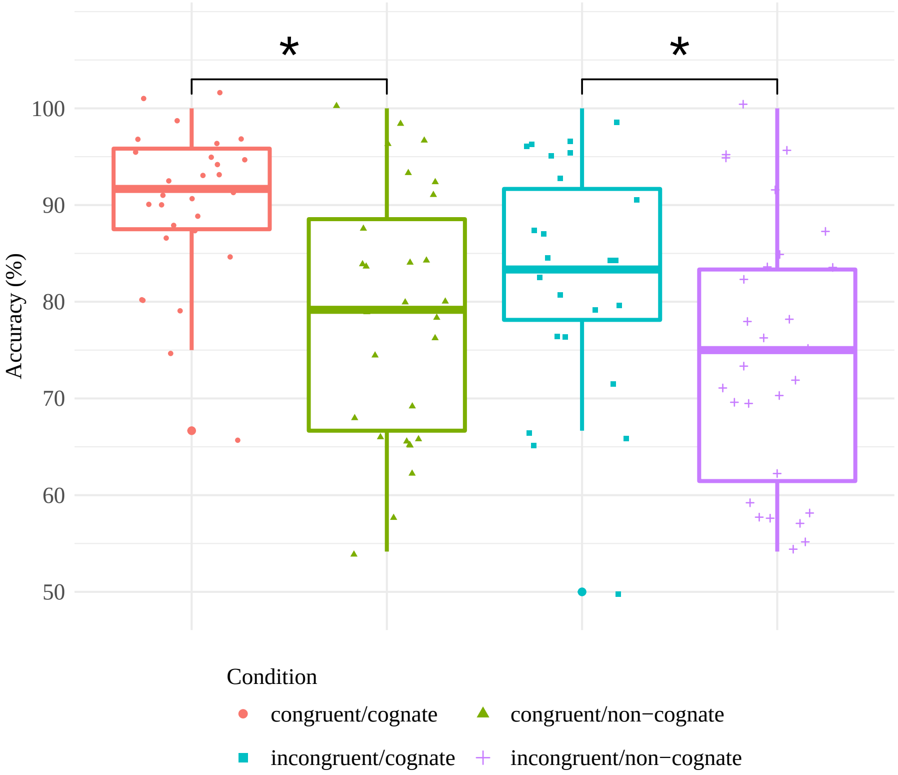
We first computed descriptive statistics for naming accuracy and naming latencies. Mean naming accuracy and mean naming latencies for each condition are shown in Table 4.3.2.

Table 4.3.2: *Mean naming accuracy and latencies by condition for the picture-naming task (n = 28).*

Condition	**Mean naming accuracy (%)**	**SD**	**Mean naming latencies (ms)**	**SD**
congruent/ cognate	89.88	30.18	911.12	253.91
congruent/ non-cognate	78.57	41.06	1011.90	297.67
incongruent/ cognate	82.59	37.95	1027.61	305.52
incongruent/ non-cognate	75.00	43.33	1068.59	281.19

Naming accuracy. The model of best fit included the following terms: main effects for *gender congruency* and *cognate status*, as well as *LexTALE-Esp score* and *familiarisation phase performance* as covariates. The random slopes for our main manipulations for each participant led to singular fit. Similarly, the model with included the interaction effect between *gender congruency* and *cognate status* did not yield a better model fit with $\chi^2(1, \text{n} = 28) = 0.494$, $p = 0.482$. We subsequently simplified our fixed effects and random effects structures and included random intercepts for *participant* and *item*. Our best-fitting model was therefore as follows: naming accuracy ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + LexTALE-Esp score + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (1|participant) + (1|item). Participants were significantly more accurate for cognates over non-cognates with $\beta = 0.647$, *95% CI*[0.443, 0.946], $z = -2.24$, $p = 0.025$. In contrast, participants were not more accurate for congruent items compared to incongruent items with $\beta = 0.807$, *95% CI*[0.550, 1.18], $z = -1.10$, $p = 0.270$. See Appendix 4.E for details about the model parameters, and Table 4.3.2 for mean naming accuracy across the conditions.

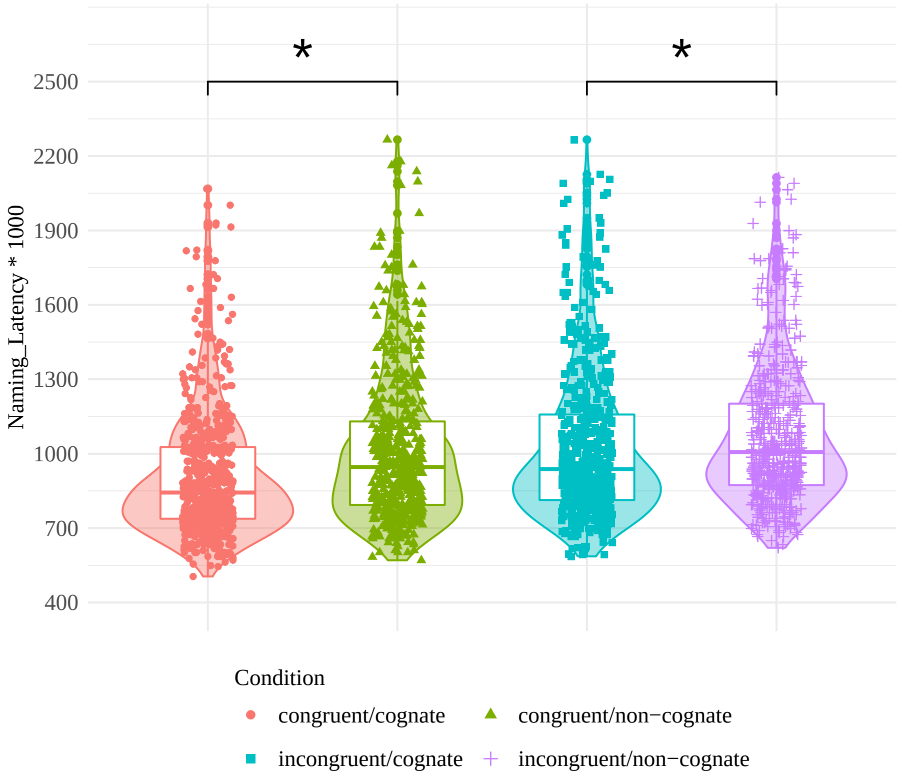
Figure 4.3.2: *Visualisation of naming accuracy for each condition for the picture-naming task (n = 28). The brackets indicate statistical differences across conditions.*



Naming latencies. Our best-fitting model included a main effect for both *gender congruency* and *cognate status*, *familiarisation phase performance* as covariate, correlated random slopes for *gender congruency* and *cognate status* for each participant, and random intercepts for *participant* and *item*. The model containing the interaction effect between *gender congruency* and *cognate status* was not statistically better compared to the model without with χ^2(1, n = 28) = 0.256, p = 0.613. Therefore, the best-fitting model was: naming latencies ∼ gender congruency (congruent vs.

incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|participant) + (1|item). Critically, participants were faster at naming cognates compared to non-cognates, with $\beta = 0.071$, *95% CI*[0.007, 0.135], t = 2.18, p = 0.030. Participants were marginally not faster at naming congruent compared to non-congruent items, with $\beta = 0.067$, *95% CI*[-0.001, 0.135], t = 1.94, p = 0.052. See Appendix 4.F for the parameters of the best-fitting model, and Figure 4.3.3 for a visualisation of mean naming latencies across conditions. Note that model estimates are reported in seconds. Taken together, *cognate status*, but not *gender congruency*, significantly influenced both naming accuracy and naming latencies in this study.

Figure 4.3.3: *Visualisation of naming latencies for each condition for the picture-naming task (n = 28). The brackets indicate statistical differences across conditions.*



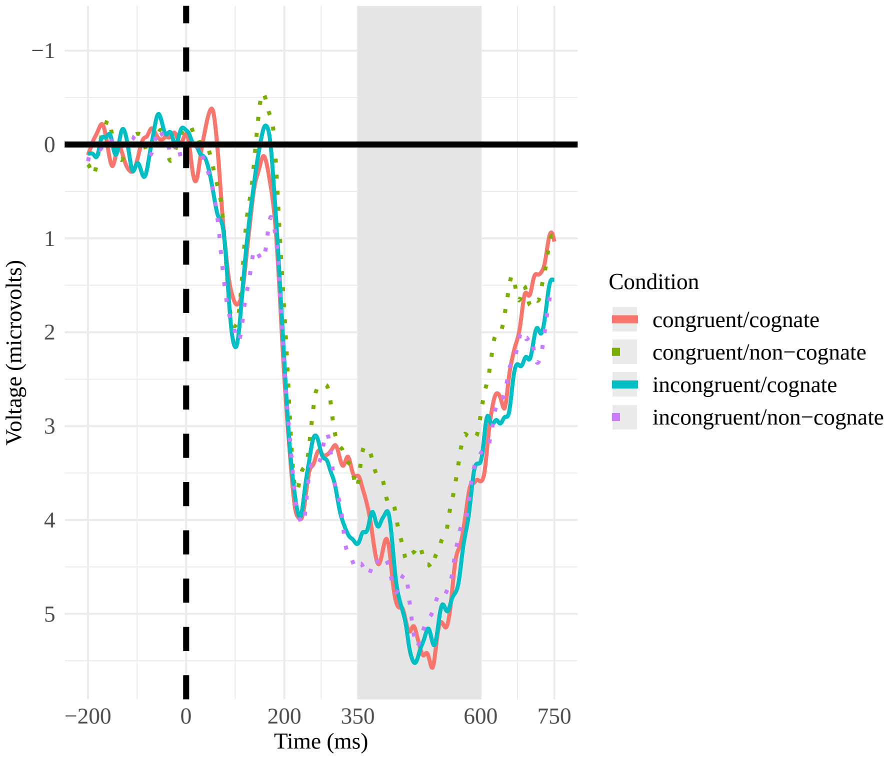
Voltage amplitudes. Visual inspection of the EEG data across the entire segment showed the classical N1/P2/N2 ERP complex for early visual processing (Eulitz et al., 2000) in our ROI (Figure 4.3.4). Voltage amplitudes reached a peak around 550 ms, which was followed by a downward trend back to baseline. Descriptive statistics for the time window between 350 ms and 550 ms for channels P3, P4, P7, P8, Pz, O1, O2 and Oz can be found in Table 4.3.3. Descriptively, we found the highest voltage amplitudes in this particular time window for congruent cognates, followed by incongruent non-cognates, incongruent cognates and finally, congruent non-cognates.

Table 4.3.3: *Voltage amplitudes by condition for the time window of interest (350 ms - 550 ms) for channels P3, P4, P7, P8, Pz, O1, O2 and Oz for the picture-naming task (n = 28).*

Condition	**Mean voltage(μV)**	**SD**
congruent/cognate	4.77	8.71
congruent/non-cognate	3.96	8.91
incongruent/cognate	4.72	8.16
incongruent/non-cognate	4.74	8.35

Our maximal model as outlined in section 4.3.3 failed to converge. We subsequently simplified our fixed and random effects structures. The best-fitting model for voltage amplitudes included *condition* as fixed effect, as well as the covariate *hemisphere* and *familiarisation phase performance.* The best-fitting model was as follows: voltage amplitudes ∼ condition (congruent cognate vs. congruent non-cognate vs. incongruent cognate vs. incongruent non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (condition|participant) + (1|item). Despite being included in the final model, there was no statistical difference between the condition levels; see Appendix 4.G for the exact model parameters. Therefore, statistically speaking, there was no significant modulation of voltage amplitudes as a function of *condition*, and voltage amplitudes were statistically comparable across *congruency type* and *cognate status.* In addition, the model included a random slope for the effect of *condition* for each random intercept of *participant*, as well as a random intercept for *item.*

Figure 4.3.4: *Visualisation of voltage amplitudes for each condition for channels P3, P4, P7, P8, Pz, O1, O2 and Oz (n = 28).*



4.3.4 Discussion

In this study, we conducted an almost identical experiment to Von Grebmer Zu Wolfsthurn et al. (2021b), except that instead of German-Spanish speakers, we tested Italian-Spanish speakers and we used stimuli which would fit with the constraints of gender congruency and cognate status across Italian and Spanish. The aim was to examine whether CLI, in particular with respect to congruency type (congruent vs. incongruent) and cognate status (cognate vs. non-cognate), had an effect on non-native production. We were particularly interested in naming accuracy, naming latencies and P300 voltage amplitudes. Taking the LRM model (Levelt et al., 1999) as the theoretical basis, we investigated during which stage of non-native language production speakers would experience measurable CLI. By extension, the current study tapped directly into the question at which stage during non-native production the target language was selected over the non-target language. On the basis of this theoretical framework and the findings from Chapter 3 of this thesis, we predicted the following: higher naming accuracy, shorter naming latencies and larger P300 voltage amplitudes for congruent cognate items, followed by congruent non-cognates and incongruent cognates, and the lowest naming accuracy, longest RTs and smallest P300 voltage amplitudes for incongruent non-cognates.

Regarding the behavioural data, we found that gender congruency and cognate status did not yield an interaction effect on neither naming accuracy nor naming latencies, contrary to our hypotheses. However, in line with our predictions, we found a significant effect of cognate status on naming accuracy and naming latencies: participants were both more accurate and faster at naming cognates compared to non-cognates. This reflects the classical cognate facilitation effect (Christoffels et al., 2007; Peeters et al., 2013). From the perspective of target language selection, this indicates that the speaker experienced CLI at the orthographic and phonological level. In other words, lexical entries from both Italian and Spanish competed at the *phonological encoding* stage of production. In turn, this suggested that the target language was not selected when lexical retrieval was completed (Christoffels et al., 2007; Colomé, 2001). Interestingly, we found no effect of gender congruency on naming accuracy or naming latencies. Despite a statistical trend, the main effect of gender congruency marginally failed to reach statistical significance. Therefore, our results did not provide evidence for differential production of congruent vs. incongruent items, and therefore also not for a gender congruency effect.

The behavioural results from this study are highly relevant for multiple reasons: first, results from the previous study on German-Spanish speakers suggested that participants were more accurate and faster for congruent vs. incongruent nouns but did not yield an effect of cognate status (Von Grebmer Zu Wolfsthurn et al., 2021b). In contrast, the results of the present study reflect the opposite pattern, namely that participants were more accurate and faster for cognates compared to non-cognates, but not for congruent compared to incongruent nouns. This suggests that during non-native production, German-Spanish speakers were more sensitive to similarities and dissimilarities at the gender level, whereas Italian-Spanish speakers were more sensitive to similarities at the orthographic and phonological level. In turn, CLI effects were traceable primarily during the stage of *lexical retrieval* for the German-Spanish speakers, and during the *phonological encoding* stage for the Italian-Spanish speakers. In other words, our findings from both studies suggest that German-Spanish speakers battled CLI mostly when processing gender, whereas Italian-Spanish speakers faced CLI primarily when processing cognates at the behavioural level. Subsequently, this suggested qualitative and quantitative differences in the underlying non-native production mechanisms for the linguistically similar languages, i.e., Italian and Spanish, compared to linguistically less similar languages, i.e., German and Spanish. Importantly, the statistical comparison of the behavioural differences in non-native production between the German-Spanish speakers and the Italian-Spanish speakers can be found in Chapter 6.

As for the EEG results, the oscillatory pattern in centro-parietal regions between 350 ms and 600 ms post-stimulus onset was consistent with a P300 component, in line with our original hypothesis. Therefore, similar to the results reported in Chapter 3 and in Von Grebmer Zu Wolfsthurn et al. (2021b), the P300 emerged as a critical component during non-native language production. However, we did not find an effect of gender congruency or cognate status on P300 voltage amplitudes: the descriptive trend of smaller voltage amplitudes for congruent non-cognates compared to the other three conditions was not statistically significant. Therefore, we did not find neural evidence for a modulation of P300 amplitudes by gender congruency or cognate status. In turn, these results do not support the notion of CLI during non-native production, as was indicated in the behavioural modulation by cognate status in the behavioural results. A possible interpretation of this finding is that the modulation of P300 amplitudes was too subtle and potentially masked by the remaining noise in our data, despite meticulous pre-processing of the ERP data. Another interpretation is that effects of gender congruency and cognate status on P300 amplitudes had a counterbalancing effect; thereby effectively cancelling any neural effects but preserving a behavioural effect of cognate status. Taken together, we did not find traceable CLI effects at the level of P300 component amplitudes. In turn, this did not allow for the exploration of the locus of target language selection. Subsequently, we were unable to examine during which production stage our speakers experienced CLI from either gender congruency or cognate status. This particular notion therefore remains an open issue for future research. Nevertheless, our results again highlight the relevance of the P300 component during non-native production.

Comparing these EEG results with those from Chapter 3 and Von Grebmer Zu Wolfsthurn et al. (2021b), where incongruent non-cognates descriptively elicited the smallest voltage amplitudes for the German-Spanish speakers, the present study linked congruent non-cognates to the smallest elicited voltage amplitudes. However, unlike the German-Spanish speakers, the Italian-Spanish speakers did not display a significant difference in voltage amplitudes as a function of condition. In other words, ERP results from the German-Spanish speakers indicated that both the native and non-native language were active beyond the stage of lexical retrieval until at least the stage of phonological encoding. Conversely, we did not find any evidence for this in the current study. These are relevant findings because they indicate a potentially different neural signature of non-native production for German-Spanish speakers compared to Italian-Spanish speakers. The contrast in voltage amplitudes across the two groups is examined in more detail in Chapter 6 of this thesis, which describes a direct comparison of the EEG data for the German-Spanish speakers and the Italian-Spanish speakers. Finally, results from both studies provide evidence for the critical role of the P300 component in non-native production, both in linguistically similar and less similar language combinations. Future research should therefore examine the exact characteristics and involvement of the P300 component in non-native production in a more nuanced manner.

Summary and conclusions

In this chapter, we examined the effect of gender congruency (congruent vs. incongruent) and cognate status (cognate vs. non-cognate) on non-native production. Within the framework of the LRM model (Levelt et al., 1999), we probed the effect of CLI on non-native production and the locus of target language selection. We used a picture-naming task in native Italian late learners of Spanish. We were particularly focused on the impact of CLI on naming accuracy, naming latencies and P300 voltage amplitudes. Our behavioural results showed that participants were more accurate and faster at naming cognates compared to non-cognates. In contrast, they did not show an effect of gender congruency. From a neural perspective, we found no evidence that gender congruency or cognate status modulated P300 voltage amplitudes. Behaviourally, our results therefore suggested that participants faced CLI until the *phonological encoding* stage, which is consistent with Chapter 3 and Von Grebmer Zu Wolfsthurn et al. (2021b). However, we did not find complementary evidence in the EEG data. We were therefore unable to examine the locus of target language selection in non-native production in our Italian late learners of Spanish. Importantly, this notion therefore warrants a closer investigation in future studies.

4.4 General discussion

The aim of this chapter was to expand on the cross-linguistic evidence from Chapters 2 and 3 and to investigate the gender congruency effect and the cognate facilitation effect in a linguistically highly similar language pair, namely Italian-Spanish. More specifically, we used a syntactic violation paradigm to quantify CLI effects in non-native comprehension, and a picture-naming task to investigate CLI effects in non-native production. We explored several different issues: in terms of non-native comprehension, we asked the question whether and how gender congruency (congruent vs. incongruent) and cognate status (cognate vs. non-cognate) influenced syntactic violation processing. Moreover, we characterised the corresponding neural correlates and investigated whether our late language learners with moderate proficiency levels would show the P600 effect, which is typically reported for syntactic violations (Steinhauer et al., 2009). Finally, we were interested in whether P600 effect sizes would be modulated by gender congruency and cognate status as representatives for cross-linguistic influence (CLI) effects in non-native comprehension. In contrast, for non-native production, our goals were to explore whether and how gender congruency and cognate status modulated the non-native production process. Further, we also studied until when speakers experienced CLI effects related to gender congruency and cognate status and when the target language was selected over the non-target language during non-native production. Here, our main ERP component of interest was the P300 component.

The general picture emerging from our findings in this chapter is the following: first, in non-native comprehension, gender congruency emerged as the primary salient linguistic feature to impact performance on a syntactic violation task. This suggests that there is interaction between the Italian and Spanish gender systems, resulting in measurable CLI effects at the behavioural level in non-native comprehension. Next, we provided evidence for a P600 effect in our Italian late learners of Spanish with moderate proficiency levels. This is a critical finding because it suggests that there are distinct neural signatures for processing syntactically correct vs. incorrect structures. In turn, this has implications for the characterisation of the comprehension processes in late language learners. Finally, the P600 effect was statistically comparable independently of any potential influences from gender congruency and cognate status. This suggests that at earlier acquisition stages, we are not yet able to describe distinct neural signatures as a function of CLI effects. Second, in non-native production, behavioural results showed that cognate status was the more salient cue during the production process. This suggested that both languages were active until the later stages of non-native production, and that CLI persisted at least until the phonological encoding stage. This is consistent with previous research suggesting that the target language is selected after lexical retrieval (Christoffels et al., 2007; Colomé, 2001; Hoshino & Thierry, 2011). By extension, this finding implied that Italian late learners of Spanish with moderate proficiency levels may have faced CLI for a large part of the production process. However, in the absence of complementary evidence at the neural level, further research is needed to corroborate these findings.

Comparing these findings from both experiments on non-native comprehension and production in the Italian-Spanish speakers, the striking difference was that gender congruency was a modulator for non-native comprehension, but that cognate status played a more significant role in non-native production. This is a critical finding because it speaks directly to the respective relevance of linguistic features such as gender congruency and cognate status across the linguistic domains of comprehension vs. production. In turn, this suggests that speakers use different linguistic cues to successfully manage non-native comprehension and non-native production. The systematic comparison of the CLI effects across the two domains is beyond the scope of this chapter (but see Chapter 5 and Chapter 6). However, we argue that future research should investigate this particular notion more closely to provide a more detailed and nuanced picture of CLI effects across comprehension and production.

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 the CDC lab group at University of Pompeu Fabra, in particular Nuría Sebastián-Gallés and Ege Ekin Özer. We are also grateful to Julian Karch for his support in the statistical analyses. The analyses were performed using the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. A special thanks goes to all of our participants for their time and participation.

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/tjea9/?view_only=cc2a2daf17bd4cf38a42ec8fc06b20ce`

Citation diversity statement

This citation diversity statement serves the purpose of highlighting the systematic bias in academia against citing from women and members of minorities (Dworkin et al., 2020; Zurn et al., 2020). We therefore calculated the distribution of gender in terms of the first and last author in our reference list. We included papers written by 17.6% woman/woman authors, 31.4% man/man, 23.5% woman/man and finally, 15.7% man/woman authors. For lack of a direct comparison, in the field neuroscience the reference lists typically contain 6.7% woman/woman authors, 58.4% man/man, 25.5% woman/man and finally, 9.4% man/woman authors (Dworkin et al., 2020).

Appendix

4.A Linguistic profile: Italian-Spanish group

Table 4.A.1: *Overview of the native and non-native languages acquired by the Italian-Spanish speakers (N = 33).*

	L1	L2	L3	L4	L5	**Total**
Italian	n = 33					**33**
Spanish		n = 2	n = 18	n = 10	n = 3	**33**
English		n = 27	n = 5			**32**
French		n = 4	n = 8	n = 3		**15**
German			n = 1	n = 2		**3**
Portuguese					n = 3	**3**
Catalan				n = 1	n = 1	**2**
Total	**33**	**33**	**32**	**16**	**7**	

4.B Model parameters: accuracy

Table 4.B.1: *Specification of model of best fit for accuracy for the syntactic violation task (n = 29). Note that estimates are reported as odds ratios.*

Formula: accuracy ∼ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + (gender congruency + cognate status|participant) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	42.78 [24.70, 74.01]	13.40	< 0.001
Violation type [violation]	0.679 [0.449, 1.03]	-1.84	0.066
Gender congruency [incongruent]	0.321 [0.195, 0.531]	-4.44	$<$ **0.001**
Cognate status [non-cognate]	2.11 [1.31, 3.40]	3.07	**0.002**
Random effects			
σ^2	3.29		
$\tau_{00 Item}$	1.22		
$\tau_{00 Participant}$	0.63		
$\tau_{11 Participant[incongruent]}$	0.30		
$\tau_{11 Participant[non-cognate]}$	0.11		
$\rho_{01 Participant[incongruent]}$	-0.49		
$\rho_{01 Participant[non-cognate]}$	0.97		
ICC	0.38		
$N_{Participant}$	29		
N_{Item}	224		
Observations	4,754		
Marginal R^2/ Conditional R^2	0.087 / 0.435		

4.C Model parameters: response times

Table 4.C.1: *Specification of model of best fit for response times (RTs) for the syntactic violation task (n = 29). Note that estimates are reported in milliseconds.*

Formula: RTs ∼ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) * cognate status (cognate vs. non-cognate) + (gender congruency + cognate status|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	864.35 [844.89, 883.81]	87.08	< 0.001
Violation type [violation]	112.80 [102.54, 123.06]	21.55	< **0.001**
Gender congruency [incongruent]	123.42 [110.12, 136.69]	18.24	< **0.001**
Cognate status [non-cognate]	-6.83 [-15.95, 2.29]	-1.47	0.142
Gender congruency [incongruent] * cognate status [non-cognate]	-68.33 [-79.05, -57.60]	-12.49	< **0.001**
Random effects			
σ^2	0.14		
τ_{00Item}	5098.76		
$\tau_{00Subject}$	7983.03		
$\tau_{11Subject[incongruent]}$	3701.49		
$\tau_{11Subject[non-cognate]}$	1095.85		
$\rho_{01Subject[incongruent]}$	-0.04		
$\rho_{01Subject[non-cognate]}$	-0.22		
ICC	1.00		
$N_{Subject}$	29		
N_{Item}	224		
Observations	4,374		
Marginal R^2/ Conditional R^2	0.293 / 1.000		

4.D Model parameters: P600 component

Table 4.D.1: *Specification of model of best fit for voltage amplitudes for the syntactic violation task (n = 29). Note that estimates are reported in microvolts.*

Formula: voltage amplitudes ∼ violation type (violation vs. non-violation) + gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + channel + terminal phoneme + LexTALE-Esp score + (violation type + gender congruency * cognate status|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	3.03 [1.86, 4.20]	5.09	< 0.001
Violation type [violation]	1.48 [0.893, 2.08]	4.92	< **0.001**
Gender congruency [incongruent]	-0.072 [-0.551, 0.406]	-0.296	0.767
Cognate status [non-cognate]	0.108 [-0.330, 0.545]	0.483	0.629
Channel [CP2]	0.693 [0.667, 0.720]	51.11	< 0.001
Channel [CP6]	0.173 [0.146, 0.199]	12.75	< 0.001
Channel [P3]	0.549 [0.523, 0.576]	40.50	< 0.001
Channel [P4]	0.906 [0.879, 0.932]	66.78	< 0.001
Channel [P7]	-2.08 [-2.10, -2.05]	-153.14	< 0.001
Channel [P8]	-1.44 [-1.47, -1.41]	-106.21	< 0.001
Channel [Pz]	1.154 [1.13, 1.18]	85.10	< 0.001
Terminal phoneme [d]	2.94 [0.256, 5.63]	2.15	0.032
Terminal phoneme [e]	-0.896 [-1.54, -0.251]	-2.72	0.006
Terminal phoneme [ión]	-0.547 [-2.12, 1.03]	-0.682	0.495
Terminal phoneme [l]	0.318 [-0.604, 1.24]	0.676	0.499
Terminal phoneme [n]	0.020 [-0.627, 0.667]	0.060	0.952
Terminal phoneme [o]	-0.308 [-0.735, 0.118]	-1.42	0.157
Terminal phoneme [r]	-0.459 [-1.50, 0.580]	-0.866	0.387

Terminal phoneme [s]	0.968 [-0.914, 2.85]	1.01	0.314
Terminal phoneme [z]	-0.967 [-3.62, 1.69]	-0.713	0.476
LexTALE-Esp score	-0.012 [-0.046, 0.023]	-0.657	0.511
Random effects			
σ^2	59.10		
τ_{00Item}	1.78		
$\tau_{00Participant}$	2.89		
$\tau_{11Participant[violation]}$	1.71		
$\tau_{11Participant[congr/non-cogn]}$	1.01		
$\tau_{11Participant[incongr/cogn]}$	1.53		
$\tau_{11Participant[incongr/non-cogn]}$	1.61		
$\rho_{01Participant[violation]}$	-0.03		
$\rho_{01Participant[congr/non-cogn]}$	-0.53		
$\rho_{01Participant[incongr/cogn]}$	-0.61		
$\rho_{01Participant[incongr/non-cogn]}$	-0.45		
ICC	0.08		
$N_{Participant}$	29		
N_{Item}	224		
Observations	5,141,248		
Marginal R^2/ Conditional R^2	0.027 / 0.110		

4.E Model parameters: naming accuracy

Table 4.E.1: *Specification of model of best fit for naming accuracy for the picture-naming task (n = 28). Note that estimates are reported as odds ratios.*

Formula: naming accuracy ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + LexTALE-Esp score + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (1|participant) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	0.326 [0.159, 0.672]	-3.04	0.002
Gender congruency [incongruent]	0.807 [0.550, 1.18]	-1.10	0.270
Cognate Status [non-cognate]	0.647 [0.443, 0.946]	-2.24	**0.025**
LexTALE-Esp score	1.02 [1.00, 1.05]	2.35	0.019
Familiarisation phase performance [one correct]	5.54 [3.67, 8.35]	8.17	< 0.001
Familiarisation phase performance [two correct]	22.63 [14.91, 34.33]	14.66	< 0.001
Familiarisation phase performance [three correct]	43.57 [28.32, 67.03]	17.18	< 0.001
Random effects			
σ^2	3.29		
$\tau_{00 Item}$	0.47		
$\tau_{00 Subject}$	0.38		
ICC	0.20		
$N_{Subject}$	28		
N_{Item}	96		

Observations	2,688
Marginal R^2/ Conditional R^2	0.307/0.448

4.F Model parameters: naming latencies

Table 4.F.1: *Specification of model of best fit for naming latency for the picture-naming task (n = 28). Note that estimates are reported in seconds.*

Formula: naming latencies $\sim$ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	1.36 [1.25, 1.47]	24.98	< 0.001
Gender congruency [incongruent]	0.067 [-0.001, 0.135]	1.94	0.052
Cognate Status [non-cognate]	0.071 [0.007, 0.135]	2.18	**0.030**
Familiarisation phase performance [one correct]	-0.200 [-0.276, -0.124]	-5.16	< 0.001
Familiarisation phase performance [two correct]	-0.338 [-0.408, -0.268]	-9.47	< 0.001
Familiarisation phase performance [three correct]	-0.412 [-0.482, -0.341]	-11.46	< 0.001
Random effects			
σ^2	0.05		
$\tau_{00 Item}$	0.00		
$\tau_{00 Subject}$	0.00		
$\tau_{11 Subject[incongruent]}$	0.00		
$\tau_{11 Subject[non-cognate]}$	0.00		
$\rho_{01 Subject[incongruent]}$	-0.20		
$\rho_{01 Subject[non-cognate]}$	-0.39		
ICC	0.17		

$N_{Subject}$	28
N_{Item}	96
Observations	2,191
Marginal R^2/ Conditional R^2	0.168/0.306

4.G Model parameters: P300 component

Table 4.G.1: *Specification of model of best fit for voltage amplitudes for the picture-naming task (n = 28). Note that estimates are reported in microvolts.*

Formula: voltage amplitudes ∼ Condition (congruent cognate vs. congruent non-cognate vs. incongruent cognate vs. incongruent non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (none correct vs. one correct vs. two correct vs. three correct) + (condition|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	3.48 [2.14, 4.81]	5.10	< 0.001
Condition [congruent/cognate]	0.657 [-0.347, 1.66]	1.28	0.200
Condition [incongruent/cognate]	0.781 [-0.324, 1.89]	1.39	0.166
Condition [incongruent/ non-cognate]	0.652 [-0.446, 1.75]	1.16	0.244
Hemisphere [midline]	0.243 [0.211, 0.274]	15.16	< 0.001
Hemisphere [right]	0.083 [0.055, 0.111]	5.84	< 0.001
Familiarisation phase performance [one correct]	0.231 [0.144, 0.318]	5.21	< 0.001
Familiarisation phase performance [two correct]	0.164 [0.084, 0.245]	4.004	< 0.001
Familiarisation phase performance [three correct]	0.684 [0.603, 0.766]	16.38	< 0.001
Random effects			
σ^2	64.64		
$\tau_{00 Item}$	1.70		

$\tau_{00 Participant}$	10.96
$\tau_{11 Participant[congr/cogn]}$	3.39
$\tau_{11 Participant[incongr/cogn]}$	4.92
$\tau_{11 Participant[incongr/non-cogn]}$	4.82
$\rho_{01 Participant[congr/cogn]}$	-0.60
$\rho_{01 Participant[incongr/cogn]}$	-0.79
$\rho_{01 Participant[incongr/non-cogn]}$	-0.62
ICC	0.12
$N_{Participant}$	28
N_{Item}	96
Observations	1,696,396
Marginal R^2/ Conditional R^2	0.002/0.123

CHAPTER 5

Processing non-native syntactic violations: different ERP correlates as a function of typological similarity

This article was submitted for review as: Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L., & Schiller, N. O. (2022). Processing syntactic violations in the non-native language: different ERP correlates as a function of typological similarity. Neuropsychologia.

Abstract: Despite often featured in theoretical accounts, the exact impact of typological similarity on non-native language comprehension and its corresponding neural correlates remain unclear. Here, we examined the modulatory role of typological similarity in syntactic violation processing, e.g., [el volcán] (the volcano) vs. [*la volcán] in the non-native language Spanish, as well as in cross-linguistic influence. Participants were either Italian late learners of Spanish (highly similar language pair) or German late learners of Spanish (less similar language pair). We measured P600 component amplitudes, accuracy and response times. In line with our predictions, we found a larger P600 effect and differential CLI effects for Italian-Spanish speakers compared to German-Spanish speakers. Interestingly, Italian-Spanish speakers responded overall more slowly compared to German-Spanish speakers. Taken together, the results reflect a typological similarity effect in non-native comprehension in the form of a processing advantage for typologically similar languages, but only at the neural level. These findings have critical implications for the interplay of different languages in the multilingual brain.

Keywords: *typological similarity, non-native comprehension, cross-linguistic influence, gender congruency effect, cognate facilitation effect, EEG, ERPs, P600 effect, generalised additive mixed models*

5.1 Introduction

A fundamental characteristic of multilingual language comprehension is *cross-linguistic influence* (CLI) between the native language (L1) and the non-native language (Kroll et al., 2015; Lago et al., 2021; Lemhöfer et al., 2008). In this study, we considered individuals who were able to communicate in two or more languages as multilinguals (Cenoz, 2013). In language comprehension, CLI is often conceptualised as the parallel activation of both the L1 and the non-native language (Hamers & Lambert, 1972; Lago et al., 2021), even when the circumstances only require the use of one language (Blumenfeld & Marian, 2013; Lago et al., 2021; Marian & Spivey, 2003b; Nozari & Pinet, 2020). CLI was demonstrated at the level of (morpho)syntax (Grüter, Lew-Williams & Fernald, 2012; Lemhöfer et al., 2008; Tolentino & Tokowicz, 2011; Zawiszewski et al., 2011), for grammatical gender (Lemhöfer et al., 2008; Paolieri et al., 2020) and for cognate processing (Midgley et al., 2011; Peeters et al., 2013). Moreover, CLI was reported for different ages of non-native acquisition (AoA), with some evidence suggesting that CLI may be more pronounced in early acquisition stages (Gillon-Dowens et al., 2010; Ringbom, 1987; Sunderman & Kroll, 2006). One important question is whether CLI is modulated by the *typological similarity*, that is, the syntactic and structural similarities between the L1 and the non-native language (Foote, 2009; Putnam, Carlson & Reitter, 2018; Tolentino & Tokowicz, 2011). In other words, does similarity at the level of, for example, grammatical gender or orthographic and phonological form overlap have an impact on non-native language processing? This is a critical issue because it is intimately linked to the functional organisation of multilinguals' languages and the question of how cross-language similarities can facilitate or hinder non-native processing (Tolentino & Tokowicz, 2011). As will be discussed below, it has long been proposed that typological similarity is a crucial factor in multilingual language processing (Casaponsa & Duñabeitia, 2016; MacWhinney, 2005; Odlin, 1989; Sabourin & Stowe, 2008; Tolentino & Tokowicz, 2011; Weinreich, 1953; Zawiszewski & Laka, 2020). Yet, there is a distinct lack of studies directly tackling the impact of typological similarity on some of the most fundamental cognitive aspects of multilingual language processing such as CLI.

This study focused on examining the role of typological similarity via two CLI effects. The first CLI effect we investigated was the *gender congruency effect*, which reflects CLI at the level of grammatical gender (hereafter gender). Gender refers to a noun classification system which is featured in several Indo-European languages (Corbett, 1991). Among those languages are Italian, German and Spanish, which are the languages of interest in this study. The gender systems of both Italian and Spanish feature a feminine and masculine gender value, marked by [la$_F$] and [il$_M$], and [la$_F$] and [el$_M$], respectively. In contrast, German has a three-way gender system characterised by a feminine, masculine and neuter gender value marked by [der$_M$], [die$_F$] and [das$_N$], respectively (Schiller & Caramazza, 2003; Schiller & Costa, 2006). The so-called gender congruency effect manifests itself in more accurate and faster processing of gender congruent items, e.g., [il$_M$ cane$_M$] and [el$_M$ perro$_M$] *"the dog"* compared to incongruent items, e.g., [il$_M$ latte$_M$] and [la$_F$ leche$_F$] *"the milk"* in Italian and Spanish (Lemhöfer et al., 2008; Paolieri et al., 2019; Sá-Leite et al., 2020). In other words, similarity at the level of gender results in a measurable processing advantage for gender congruent items vs. incongruent items across the L1 and the non-native language.

The second CLI effect we examined in this study was the *cognate facilitation effect*. It reflects CLI at the level of orthographic and phonological overlap, i.e., cognates. More specifically, this effect entails more accurate and faster processing of cognates, i.e., words with a significant overlap in terms of orthographic and phonological word form, e.g., [vulcano] and [volcán] *"volcano"*; compared to non-cognates, e.g., [viso] and [cara] *"face"* in Italian and Spanish (Comesaña et al., 2014; Costa et al., 2005; Marian, Blumenfeld & Boukrina, 2008; Midgley et al., 2011; Lemhöfer et al., 2008). With respect to typological similarity, Marian et al. (2008) showed that a larger phonological overlap for native Russian speakers with high proficiency in English was linked to higher performance and shorter response times (RTs) in an auditory lexical decision task. In turn, this particular effect highlights the processing advantage for orthographically and phonologically similar word forms, i.e., cognates compared to non-cognates. Taking both effects together, the gender congruency effect and the cognate facilitation effect tentatively indicate a processing advantage for typologically more similar structures compared to less similar structures, as reflected by higher accuracy and faster RTs for congruent items and cognates compared to incongruent items and non-cognates.

In this study, we used both effects to closely examine the impact of typological similarity on non-native comprehension, specifically in terms of gender similarity and orthographic and phonological word form overlap between the L1 and the non-native language. Directly relevant to this study is the *Language Distance Hypothesis*, LDH (Zawiszewski & Laka, 2020), which provides a theoretical account of the interaction between typological similarity and CLI effects. The core prediction of this account is the modulation of CLI on the basis of (morpho)syntactic similarity between the L1 and the non-native language. Concretely, the LDH predicts more native-like behavioural patterns and event-related components (ERPs) emerging in the non-native language for highly morphologically similar structures across the L1 and the non-native language compared to less similar structures. This would be reflected in higher accuracy, shorter RTs and larger (more native-like) ERP components for morphologically similar structures across languages.

Zawiszewski and Laka (2020) systematically tested this account in a recent experiment on morphological processing in grammatical and ungrammatical sentences in highly proficient Basque-Spanish speakers and Spanish-Basque speakers. The critical manipulation was the presence or absence of a particular morphological feature in the non-native language compared to the L1. Consistent with the LDH, their results indicated a link between shorter RTs and larger ERP effects (i.e., native-like ERP effects) in the non-native language for some morphologically similar structures compared to less similar structures. In turn, this suggested an overall processing advantage in the non-native language for morphologically similar structures. Critically, the authors also acknowledged that AoA and non-native proficiency could modulate typological similarity effects. This is in line with previous studies which have highlighted the impact of non-native proficiency on typological similarity effects (Gillon-Dowens et al., 2010; Ringbom, 1987; Tokowicz & MacWhinney, 2005; Weber-Fox & Neville, 1996). For example, Tokowicz and MacWhinney (2005) examined low proficient and highly proficient English-Spanish speakers and their sensitivity to the correctness of syntactic structures. In addition to non-native proficiency, the second critical manipulation was that some syntactic structures were similar across the languages (auxiliary marking), whereas the other structures were not (gender and number agreement). Results demonstrated that increased typological similarity was linked to shorter RTs, in particular for lower proficient speakers. In contrast, highly proficient speakers in this study appeared to remain largely unaffected by typological similarity. This finding suggests that typological similarity effects may be more pronounced in earlier acquisition stages (Sunderman & Kroll, 2006; Zawiszewski & Laka, 2020). Therefore, in this study we focused on late language learners to examine typological similarity effects more closely, i.e., individuals who acquired a non-native language later during development after the age of fourteen (S. Rossi et al., 2006).

Before the formulation of the LDH, earlier work by Sabourin and Stowe (2008) examined the impact of typological similarity on gender processing. In their study, they compared gender agreement processing in Dutch across native Dutch speakers vs. native German and Romance language speakers, who were all late learners of Dutch (AoA $>$ 14 years of age). In terms of typological similarity, German and Dutch have a greater linguistic overlap compared to Romance languages and Dutch (Schepens et al., 2013; Van der Slik, 2010). Therefore, in their study, the German-Dutch speakers represented the typologically similar language pair, and the Romance language-Dutch speakers the typologically less similar language pair. Importantly, the authors also explored the effects of typological similarity on neural correlates of gender processing, with a specific focus on P600 component amplitudes. The P600 component is an event-related brain potential (ERP) and is characterised as a positive-going waveform reaching its peak approximately 600 ms post-stimulus onset in centro-parietal regions (Friederici et al., 1999; Friederici, Hahne & Saddy, 2002; Swaab et al., 2011). The so-called P600 effect has been reported in the context of higher voltage amplitudes for syntactic violations such as [*la$_F$ volcán$_M$] vs. syntactically correct structures such as [el$_M$ volcán$_M$] *"the volcano"* (Hagoort et al., 1993; Friederici, Gunter, Hahne & Mauth, 2004; Hahne, 2001; Weber-Fox & Neville, 1996). Critically, Sabourin and Stowe (2008) found that P600 effects were modulated by syntactic similarity between the L1 and Dutch: only native German speakers showed a clear P600 effect for syntactic violations in Dutch, whereas the native Romance language speakers did not. The results suggested that typologically similar languages (e.g., German-Dutch) were linked to an enhanced sensitivity to gender violations in comparison to less typologically similar languages (e.g., Romance language-Dutch) and a larger P600 effect. Behaviourally, the German-Dutch speakers outperformed the Romance language-Dutch speakers in terms of accuracy in gender assignment, which indicates differential CLI effects as a function of typological similarity. These results are in line with the predictions by the LDH (Zawiszewski & Laka, 2020) and are also compatible with studies linking increased CLI to typologically similar languages compared to typologically less similar languages (Mosca, 2017; Tolentino & Tokowicz, 2011).

In sum, current research strongly suggests that typological similarity plays a significant role in modulating both behavioural and neural measures of non-native language comprehension. More specifically, typological similarity was shown to influence non-native gender processing as well as orthographic and phonological processing. In this, previous studies suggest the following: first, behavioural effects of typological similarity were found for cross-linguistic gender processing, suggesting a gender processing advantage for typologically similar languages compared to less similar languages (Paolieri et al., 2020). Secondly, typological similarity effects were also found for cross-linguistic cognate processing, whereby a higher typological similarity was linked to more efficient and faster processing of orthographically and phonologically similar structures (Costa et al., 2005; Comesaña et al., 2014; Lemhöfer et al., 2008). Critically, this suggests that CLI is influenced by typological similarity, with more pronounced CLI for typologically similar language combinations compared to less similar combinations (Sabourin & Stowe, 2008; Tolentino & Tokowicz, 2011). Third, studies have also reported a typological similarity effect on the neural correlates of cross-linguistic non-native gender processing (Sabourin & Stowe, 2008). Specifically, larger, more native-like P600 effects were linked to a higher typological similarity (Sabourin & Stowe, 2008).

5.1.1 The current study

The aim of the current study was to systematically investigate the effect of typological similarity on syntactic violation processing and on CLI in non-native comprehension in late language learners using behavioural measures (accuracy and RTs) and ERP measures (P600 component voltage amplitudes). For this, we tested two groups of late learners of Spanish speakers with a varying degree of typological similarity: representing the typologically similar group, we tested native Italian speakers; and representing the typologically less similar group, we tested native German speakers (Schepens, Dijkstra & Grootjen, 2012; Schepens et al., 2013). Further, we focused on two CLI effects: the gender congruency effect, which reflects CLI of the gender systems (Lemhöfer et al., 2008; Paolieri et al., 2019; Sá-Leite et al., 2020), and the cognate facilitation effect, reflecting CLI of the orthographic and phonological systems (Costa et al., 2005; Comesaña et al., 2014; Lemhöfer et al., 2008). To test these typological similarity effects, we employed a syntactic violation paradigm, whereby participants judged the grammatical correctness of noun phrases such as *el volcán* [the volcano] (non-violation trial) vs. **la volcán* (violation trial) while we recorded their ERPs.

Research questions

The research questions we sought to answer in this study were the following: first, is there a P600 effect (i.e., a difference between non-violation and violation trials) for both the Italian-Spanish and the German-Spanish group? Second, is the P600 effect larger for one group compared to the other? Third, do CLI effects of gender congruency and cognate status vary across the two groups? This would reflect a typological similarity effect at the neural level, as well as a typological similarity effect on CLI between the native and the non-native language. Taking the LDH (Zawiszewski & Laka, 2020) as our theoretical basis, we predicted that speakers of typologically similar languages would bear a processing advantage in the non-native language compared to speakers of typologically less similar languages.

Hypotheses

Behavioural hypotheses. With respect to our first research question, we expected participants to be significantly more accurate and faster for non-violation trials compared to violation trials. Critically, for our second research question, we predicted an interaction effect of *L1* (Italian vs. German) with *violation type* (non-violation vs. violation) to indicate a typological similarity effect on processing syntactic (non-)violations. In other words, consistent with the LDH, we predicted that the Italian-Spanish group would be more accurate and faster at processing non-violation trials vs. violation trials compared to the German-Spanish group. For our third research question, we first predicted CLI effects, as manifested in more accurate and faster processing of congruent and cognate items compared to incongruent and non-cognate items. Importantly, here we aggregated our two main manipulations *gender congruency* (congruent vs. incongruent) and *cognate status* (cognate vs. non-cognate) into the variable *condition* with four levels: congruent/cognate, congruent/non-cognate, incongruent/cognate and incongruent/non-cognate items. In this, we investigated an interaction effect of *L1* with *condition.* We hypothesised that the Italian-Spanish group would be statistically more accurate and faster at processing congruent and cognate items vs. incongruent and non-cognate items compared to the German-Spanish group.

ERP hypotheses. In terms of our first research question, we expected a P600 effect in both groups, as indicated by smaller voltage amplitudes for non-violation trials compared to violation trials. For our second research question, we predicted an interaction effect between *L1* and *violation type* to indicate a typological similarity effect on the P600 effect size. More specifically, in line with the LDH, we hypothesised a larger P600 effect for the Italian-Spanish group compared to the German-Spanish group. For our third research question, we predicted an interaction effect between *L1* and *condition*, indicating an effect of typological similarity on CLI. Specifically, we expected to observe larger voltage amplitudes connected to larger CLI for the Italian-Spanish group compared to the German-Spanish group. Taken together, these findings would indicate a general processing advantage for the Italian-Spanish group compared to the German-Spanish group, with overall higher accuracy, shorter RTs and larger P600 amplitudes for the typologically similar language combination.

5.2 Methods

Before the experiment, participants filled out the Language Experience and Proficiency Questionnaire, LEAP-Q (Kaushanskaya et al., 2020; Marian et al., 2007). The LEAP-Q was used to establish proficiency and experience measures for the participants' known languages. During the experimental session, participants completed the LexTALE-Esp task (Izura et al., 2014), a lexical decision task that provides a vocabulary size score (*LexTALE-Esp score*) in Spanish. LexTALE-Esp scores were previously found to be highly correlated with overall proficiency levels, see Lemhöfer and Broersma (2012). Subsequently, participants completed the syntactic violation paradigm. Note that the German-Spanish participants included in this study as well as the procedures used are identical to the ones reported in Von Grebmer Zu Wolfsthurn et al. (2021a).

5.2.1 Participants

We recruited and tested 33 native speakers of Italian (24 females) with $M = 27.12$ years of age ($SD = 4.08$). We also tested 33 native speakers of German (27 females) with $M = 23.06$ years of age ($SD = 2.47$), previously described in Von Grebmer Zu Wolfsthurn et al. (2021a). All participants had an intermediate B1/B2 proficiency level in Spanish according to the Common European Framework of Reference for Languages, *CEFR* (Council of Europe, 2001). We established this proficiency level using various linguistic variables of the LEAP-Q, the LexTALE-Esp score and by recruiting directly from foreign language courses aimed at the B1/B2 level. Participants had to meet the recruitment criteria to be eligible for the study: dominant right-handed, between 18 and 35 years of age, absence of psychological, reading or language impairments, no second language learnt before the age of five and an age of acquisition of Spanish of more than fourteen years. We imposed additional recruitment criteria for the Italian-Spanish group because we tested them in the non-native environment: participants had to have lived in a Spanish-speaking country for less than one year and started learning Spanish shortly before or upon their arrival to Spain. We combined these criteria with the information of the LEAP-Q to establish our speakers within the category of late language learners with intermediate B1/B2 proficiency levels (Kaushanskaya et al., 2020). Note that not all participants were included in the data analyses, see section 5.3.4 for details about data exclusion.

Linguistic profile of participants

Below, we summarised several key linguistic variables related to Spanish from the LEAP-Q and the LexTALE-Esp (Table 5.2.1). We limited these descriptions to the participants included in the statistical analyses (section 5.3.4). In the Italian-Spanish group, twelve participants stated they perceived Spanish as their current first foreign language in terms of dominance, thirteen participants stated Spanish as their second, three participants as their third and finally, one participant as their fourth foreign language. For the German-Spanish group, four participants self-reported Spanish as their perceived first foreign language, twenty-one participants as their second, and three as their third foreign language. See Appendix 5.A and Appendix 5.B for a more detailed linguistic profile of the two groups.

Table 5.2.1: *Linguistic profile of Spanish for the Italian-Spanish group (n = 29) and the German-Spanish group (n = 28), including the LexTALE-Esp score. Self-reported proficiency measures (speaking, comprehension, reading) were rated on a scale from zero to ten (ten being equal to maximal proficiency) and are highlighted in bold.*

Measure	**Italian-Spanish**	**German-Spanish**
LexTALE-Esp score mean	27.29 (*SD* = 14.01)	18.91 (*SD* = 20.45)
LexTALE-Esp score range	-7.37 - 49.30	-23.16 - 60.18
AoA Spanish (years)	23.31 (*SD* = 4.86)	16.46 (*SD* = 2.33)
Fluency age Spanish (years)	24.52 (*SD* = 4.45)	18.59 (*SD* = 2.13)
Reading onset age Spanish (years)	23.79 (*SD* = 4.74)	17.36 (*SD* = 2.88)
Fluent reading age Spanish (years)	23.92 (*SD* = 4.84)	18.50 (*SD* = 2.52)
Immersion in Spanish-speaking country (years)	0.48 (*SD* = 0.35)	1.04 (*SD* = 0.69)
Daily exposure (%)	41.38 (*SD* = 18.27)	9.86 (*SD* = 9.73)
Speaking proficiency	6.31 (*SD* = 1.73)	6.85 (*SD* = 0.93)
Comprehension proficiency	7.32 (*SD* = 1.76)	7.50 (*SD* = 0.88)
Reading proficiency	7.48 (*SD* = 1.48)	7.18 (*SD* = 1.12)

5.2.2 Materials and design

We used the Italian and the German version of the LEAP-Q for our two groups, respectively. Further, we generated E-prime (Version 2) scripts (Schneider et al., 2002) for the LexTALE-Esp and the syntactic violation paradigm.

Stimuli

LexTALE-Esp. In line with the original lexical decision task by Izura et al. (2014), stimuli consisted of 60 Spanish words varying in terms of frequency, as well as 30 pseudowords with different degrees of similarity to real Spanish words, for example [*alardio*]. Therefore, the critical manipulation was *condition* (word vs. pseudoword), and we measured accuracy during this task.

Syntactic violation paradigm. The stimuli selection procedure for the Italian-Spanish and the stimuli for the German-Spanish group were identical as outlined in Von Grebmer Zu Wolfsthurn et al. (2021a). However, the selected stimuli differed between the two groups due to the constraints by our main manipulations: stimuli were selected separately for each group based on their gender congruency and cognate status across Italian and Spanish, and across German and Spanish. As a result, the stimuli were different for the Italian-Spanish compared to the German-Spanish group. We selected a total of 224 stimuli for each group. We followed a 2 x 2 x 2 fully factorial design, with *violation type* (non-violation vs. violation), *gender congruency* (congruent vs. incongruent) and *cognate status* (cognate vs. non-cognate) as our critical manipulations. Half of all trials were violation trials, and the other half non-violation trials. Half of our stimuli were gender congruent, and half gender incongruent. In turn, half of the stimuli nouns were cognates, and the rest non-cognates. Therefore, each experimental condition contained 28 stimuli, adding to a total of 224 stimuli for each group. The task was a grammaticality judgment task embedded within a syntactic violation paradigm, whereby participants determined whether a noun phrase such as *el volcán* was grammatically correct. We recorded participants' EEG during this task, as well as accuracy and RTs.

EEG recordings

Italian-Spanish group. We used 32 active electrodes in a standard 10/20 montage to collect EEG data at a sampling rate of 500 Hz via the BrainVision Recorder software (Version 1.10) by BrainProducts. We placed one electrode (FT9) under the participant's left eye to record the vertical electrooculogram (VEOG), and one electrode (FT10) at the outer canthus of the left eye for the horizontal electrooculogram (HEOG). All electrodes were referenced to FCz. A ground electrode was positioned on the participant's right cheek. We used BrainVision Recorder to keep our impedances for each electrode below 10 kΩ for an enhanced signal.

German-Spanish group. We sampled the EEG data from 32 passive electrodes configured in a 10/20 montage at a rate of 500 Hz and again using the BrainVision Recorder software (Version 1.23.0001). We placed one VEOG electrode underneath the left eye, two HEOG electrodes at the outer canthus of each eye, and the ground electrode on the right cheek of the participant. The original reference electrode was Cz. We used the actiCAP ControlSoftware (Version 1.2.5.3) to ensure that impedances were below 5 kΩ for the reference and ground electrode, and below 10 kΩ for the remaining electrodes.

5.2.3 Procedure

The experimental session was carried out on a computer screen in an experimental booth and took place in the CBC Laboratories at the Pompeu Fabra University for the Italian-Spanish group, and in the Neurolinguistic Laboratories at the University of Konstanz for the German-Spanish group. Prior to the start of the experiment, we provided participants with an information sheet and a consent form in their L1, complying with the ethics code for neurolinguistic research in the Faculty of Humanities at Leiden University. During the experiment, participants completed both the LexTALE-Esp and the syntactic violation paradigm. Written instructions for each task were provided on the screen in black font on a white background. The procedure for each task was identical for both groups, with the exception that the oral and written instructions were given in Italian to the Italian-Spanish group, and in German to the German-Spanish group. After the experiment, participants received a written and oral debrief in their L1, as well as a monetary compensation for their participation.

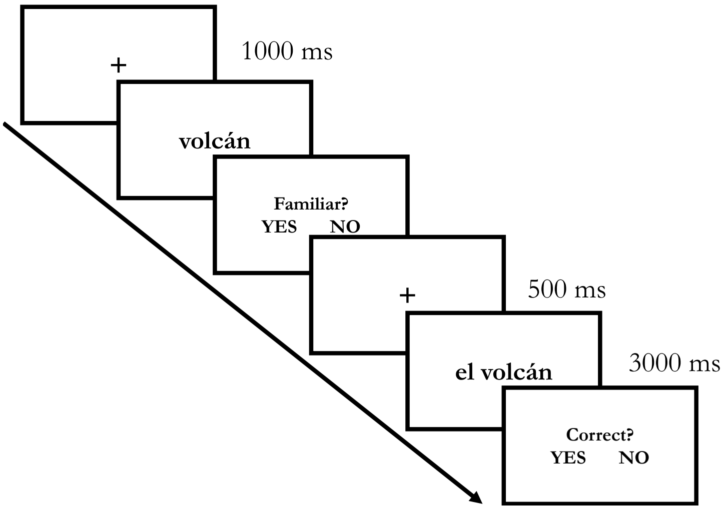
LexTALE-Esp

Participants were shown a fixation cross for 1,000 ms. Next, a letter string of either a Spanish word or pseudoword appeared on the screen. Participants decided whether or not the letter string was a Spanish word via a button press. The next trial was initiated following the participant's response. Prior to the experiment, we eliminated three word stimuli due to overlap with the stimuli from the syntactic violation paradigm. Therefore, we presented participants with 57 word stimuli, and 30 pseudoword stimuli, adding to a total of 87 trials. Each stimulus was only presented once, and trial order was fully randomised for each participant. In a final step, we calculated the LexTALE-Esp score in offline calculations by subtracting the percentage of incorrectly identified pseudowords from the correctly identified words for each participant (Izura et al., 2014).

Syntactic violation paradigm

The task procedure was identical for both groups, as is outlined in detail in Von Grebmer Zu Wolfsthurn et al. (2021a). It was as follows: participants were first presented with a fixation cross for 1,000 ms. Then, they were instructed that they would see a bare noun (e.g., *volcán* [volcano]) on the screen. Here they had to determine their familiarity with the noun by responding to a yes/no question during its presentation. This was followed by the display of a fixation cross for 500 ms. We then visually presented participants with determiner + noun constructions, e.g., *el volcán* [the volcano] for a maximum time of 3,000 ms and asked participants to determine the grammatical correctness of each noun phrase as accurately and fast as possible via a button press. The next trial was initiated upon participant's response. Each stimulus was only shown once within a noun phrase, adding to a total of 224 trials. Trial order was fully randomised, and we incorporated two self-paced breaks for our participants. At the beginning of the task, we included eight practise trials to familiarise participants with the trial procedure. Within-experiment instructions and prompts were displayed in Spanish. See Figure 5.2.1 for example trials of this task.

Figure 5.2.1: *Example trial for the syntactic violation paradigm. Within-trial prompts in the figure were translated to English for convenience. The final prompt was added to the figure for visualisation purposes only.*



5.3 Results

5.3.1 Behavioural data exclusion

We included the same participants in the behavioural analysis as in the EEG analysis (see section 5.3.4). This meant that we analysed data from 29 Italian-Spanish speakers after excluding four participants, and data from 28 German-Spanish speakers after excluding five participants, thereby analysing data from a total of 57 participants.

5.3.2 Behavioural data analysis

The behavioural data analysis procedure matched the analysis described in Von Grebmer Zu Wolfsthurn et al. (2021a), with the exception that our maximal model in this study reflected our research questions. Here, we used a generalised linear mixed effects modelling (GLMM) approach to model *accuracy* and *RTs* for the grammaticality judgement. All analyses were implemented in R, Version 4.1.2, and in RStudio, Version 2021.09.0 (R Core Team, 2021) using the *lme4* package (Bates et al., 2020). We specified a binomial distribution to model *accuracy*, and a gamma distribution with an identity link function to model positively skewed *RTs* from correct trials (Lo & Andrews, 2015). We initially built a theoretically plausible maximal model with an elaborate fixed effects structure. This included the interaction effect for *L1* (Italian vs. German) and *violation type* (violation vs. non-violation), as well as the interaction effect for *L1* and *condition* (congruent/cognate vs. congruent non-cognate vs. incongruent/cognate vs. incongruent non-cognate), representing the CLI effects. Next, our model further included the covariates *LexTALE-Esp score*, *order of acquisition of Spanish*, *terminal phoneme*, *target noun gender* and *word length*. Finally, we included random intercepts for *participant* and *item*, as well as random slopes for *violation type* and *condition* for a maximal random effects structure (Barr, 2013). Upon model non-convergence or singular fit, we simplified our random effects structure. We then tested for the relevance of each covariate and the significance of the other fixed effects terms by systematically examining their statistical significance in a model comparison approach using the *anova()* function. A significant χ^2-test indicated that a particular term significantly contributed to an improved goodness of fit and was subsequently kept in the model. For accuracy, the models were fitted with the Laplace approximation. For RTs, we used the default maximum likelihood estimation (Bates et al., 2020) for unbiased estimates for the model comparisons, but re-fitted the final model with the restricted maximum likelihood method (Mardia, Southworth & Taylor, 1999). We determined treatment coding as our default contrast, and vigorously checked the model diagnostics using the *DHARMa* package (Hartig, 2020). P-values were derived using the *lmerTest* package (Kuznetsova, Brockhoff, Christensen & Pødenphant-Jensen, 2020), and test statistics above ±1.96 were interpreted as significant at $\alpha = 0.05$ (Alday et al., 2017). Note that we report model parameters for accuracy as odds ratios.

5.3.3 Behavioural data results

We calculated mean accuracy and RTs for each condition and each group in Table 5.3.1.

Accuracy. The maximal model described above in section 5.3.2 did not converge and was subsequently simplified. The simplified model contained both interaction effects, but yielded an insignificant interaction effect for *L1* and *violation type* with β = 0.946, z = -0.145, p = 0.885. We therefore compared this model to a model which included only the interaction effect between *L1* and *condition*, but not *L1* and *violation type*. There was no significant difference in model fit between these two models with χ^2(1, n = 57) = 0.021, p = 0.885, and we subsequently selected the simpler model as our best-fitting model (Appendix 5.D). This best-fitting model included the interaction effect between *L1* and *condition*, and main effects for *L1* and *violation type*. Further, the model included *LexTALE-Esp score* and *target noun gender* as covariates, by-*participant* random slopes for *violation type*, and random intercepts for both *participant* and *item* (Appendix 5.D). Therefore, the final model was: accuracy ∼ L1 (Italian vs. German) + violation type (violation vs. non-violation) + L1 * condition (congruent/cognate vs. congruent/non-cognate vs. incongruent/cognate vs. incongruent/non-cognate) + LexTALE-Esp score + target noun gender (feminine vs. masculine) + (violation type|participant) + (1|item).

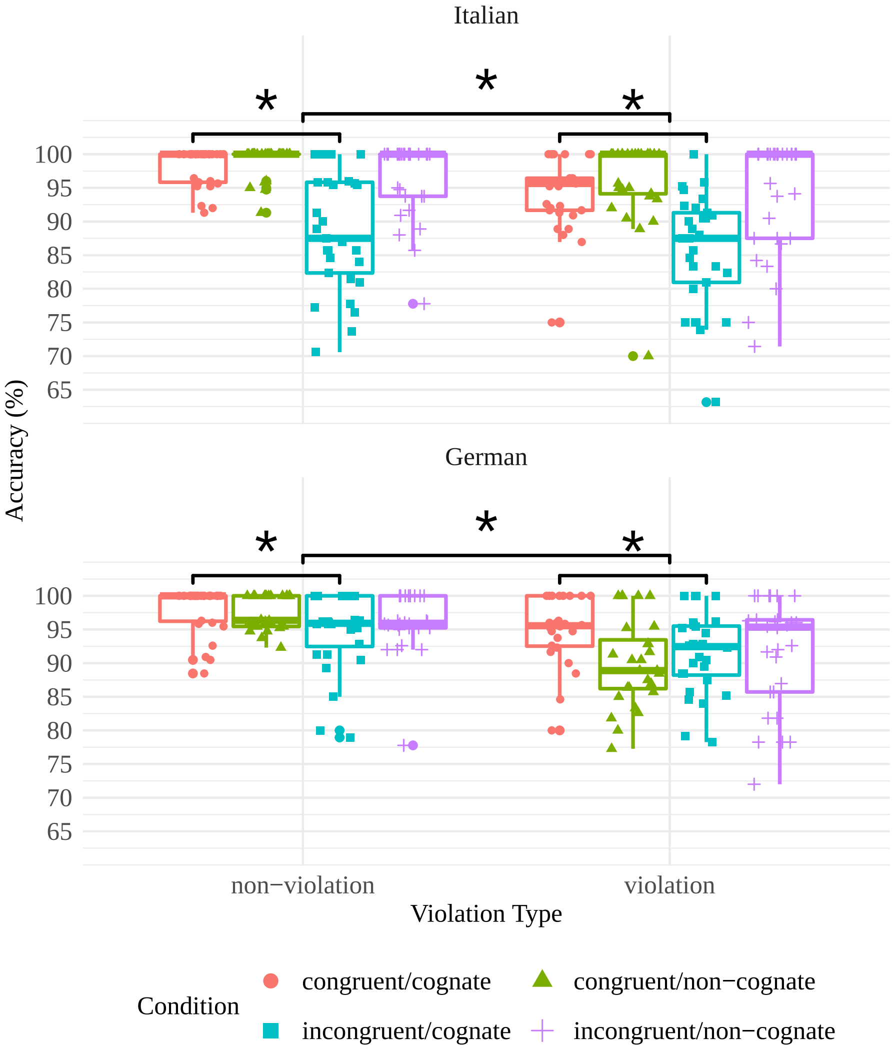
Participants were more accurate for non-violation trials compared to violation trials with β = 0.412, *95% CI*[0.279, 0.609], z = -4.45, p < 0.001. Further, there was a main effect of *condition* with participants being more accurate for congruent/cognate items compared to incongruent/cognate items with β = 0.258, *95% CI*[0.132, 0.504], z = -3.96, p < 0.001 (Figure 5.3.1). Despite being included in the final model, the main effect for *L1* was not significant with β = 1.50, *95% CI*[0.673, 3.35], z = 0.993, p = 0.321 for the Italian-Spanish group compared to German-Spanish group. Critically, the interaction effect between *L1* and *condition* was insigni-

Table 5.3.1: *Mean accuracy and mean RTs for each condition for each group (n = 57).*

L1	**Violation type**	**Condition**	**Mean (%)**	**SD**	**Mean (ms)**	**SD**
Italian	non-violation	congruent/cognate	98.08	13.74	816.76	328.44
		congruent/non-cognate	98.97	10.13	817.31	347.17
		incongruent/cognate	88.76	31.62	931.11	452.49
		incongruent/non-cognate	96.54	18.31	854.38	377.15
	violation	congruent/cognate	94.08	23.61	953.55	401.33
		congruent/non-cognate	96.38	18.70	929.73	356.10
		incongruent/cognate	85.60	35.14	1049.37	489.65
		incongruent/non-cognate	93.92	23.92	991.48	439.02
German	non-violation	congruent/cognate	98.08	13.72	721.69	320.48
		congruent/non-cognate	97.52	15.57	724.44	318.78
		incongruent/cognate	94.78	22.25	791.09	370.86
		incongruent/non-cognate	96.02	19.56	737.29	351.83
	violation	congruent/cognate	95.06	21.70	852.12	387.34
		congruent/non-cognate	90.23	29.71	833.22	368.91
		incongruent/cognate	91.47	27.95	875.94	386.87
		incongruent/non-cognate	91.92	27.27	859.32	406.40

ficant for all levels contrasted with the Italian-Spanish group and congruent/cognate items with β = 0.349, *95% CI*[0.121, 1.01], z = -1.94, p = 0.052 for the German-Spanish group and congruent/non-cognate items, β = 1.68, *95% CI*[0.631, 4.46], z = 1.04, p = 0.300 for the German-Spanish group and incongruent/cognate items, and finally, β = 0.661, *95% CI*[0.238, 1.84], z = -0.792, p = 0.428 for the German-Spanish group and incongruent/non-cognate items. Taken together, we found a main effect of *violation type* and a small main effect for *condition* on accuracy. However, we found neither a significant interaction effect of *L1* and *violation type*, nor of *L1* and *condition*. We also did not find a main effect of *L1* on accuracy, either. This indicated that accuracy levels were comparable for the two groups. See Appendix 5.D for the full model parameters for accuracy.

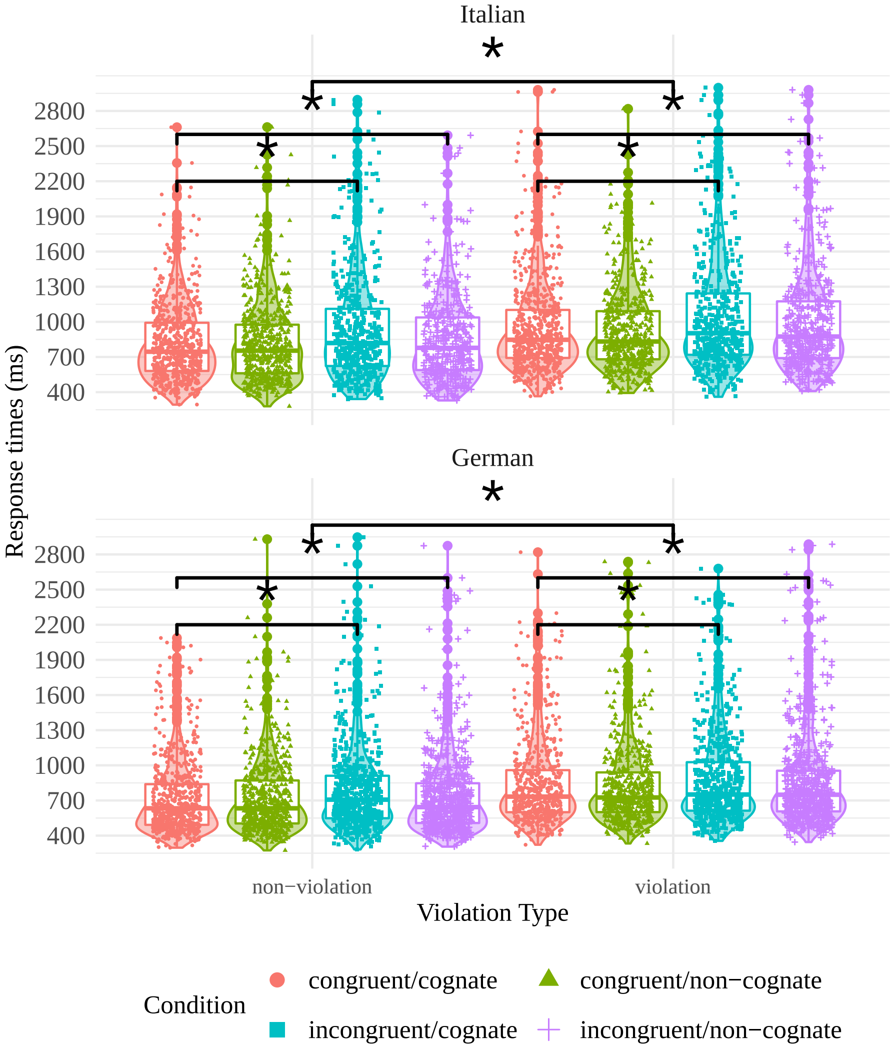
Figure 5.3.1: *Mean accuracy (%) for each group for each condition (n = 57).*



Response times. The maximal model described in section 5.3.2 that included both interaction terms yielded non-convergence. We subsequently simplified the random effects structure and also excluded *LexTALE-Esp score* as a covariate. This simplified model yielded an insignificant interaction effect for *L1* and *violation type* with β = -1.51, t = -0.407, p = 0.684 for Italian and non-violation items compared to German and violation items. We then compared this model to a model which included only the interaction effect between *L1* and *condition*, but not *L1* and *violation type*. This comparison showed no difference in model fit with $\chi^2(1, \text{n} = 57) = 0.001$, p = 0.971. We therefore declared the model containing the interaction effect between *L1* and *condition* and main effects of *L1* and *violation type* as our best-fitting model (Appendix 5.E). Similar to the best-fitting model for accuracy, this model also included *target noun gender* as covariate, by-*subject* random slopes for *violation type* and random intercepts for *participant* and *item* (Appendix 5.E). Subsequently, the best-fitting model was: RTs ~ L1 (Italian vs. German) + violation type (violation vs. non-violation) + L1 * condition (congruent/cognate vs. congruent/non-cognate vs. incongruent/cognate vs. incongruent/non-cognate) + target noun gender (feminine vs. masculine) + (violation type|participant) + (1|item).

Participants were faster for non-violation trials compared to violation trials with β = 128.18, *95% CI*[93.90, 162.45], t = 7.33, p < 0.001. Participants were also significantly faster for congruent/cognate items compared to incongruent/cognate items with β = 105.64, *95% CI*[89.30, 121.98], t = 12.67, p < 0.001, and for incongruent/non-cognates with β = 36.02, *95% CI*[25.35, 46.68], t = 6.62, p < 0.001. Importantly, participants in the German-Spanish group were statistically faster compared to the Italian-Spanish group with β = -82.55, *95% CI*[-100.54, -64.56], t = -8.99, p < 0.001. Moreover, the interaction effect between *L1* and *condition* was significant for Italian and congruent/cognate items compared to German and incongruent/cognate items with β = -63.19, *95% CI*[-102.49, -23.89], t = -3.15, p = 0.002, with Italian participants being significantly slower (Figure 5.3.2). In sum, we found first, that participants were faster for non-violation compared to violation items; second, that participants were faster for congruent/cognate items than for incongruent/cognate and incongruent/non-cognate items; third, that the German-Spanish group was overall faster compared to the Italian-Spanish group; and fourth, that the German-Spanish group was faster for incongruent/cognate items compared to congruent/cognate items than the Italian-Spanish group. This indicated an effect of *L1* on CLI across the two groups for RTs. See Appendix 5.E for the full model parameters for RTs.

Figure 5.3.2: *Mean response times (ms) for each group for each condition (n = 57).*



5.3.4 EEG data exclusion

EEG trials were excluded based on one of the following reasons: first, the participant had indicated that they were unfamiliar with the noun; second, because the participant made an incorrect grammatical judgement; and third, the trial segment contained an artefact. Therefore, we only included familiar, correct and uncontaminated trials in our analysis, provided that the trial rejection threshold did not exceed 60% of trials per participant. Subsequently, we excluded four participants from the Italian-Spanish group, and four participants from the German-Spanish group. Moreover, one participant from the German-Spanish group was lost due to a recording failure. In total, we included 57 datasets, 29 from the Italian-Spanish group, and 28 from the German-Spanish group. We included the same participants in the behavioural analyses (see previous section 5.3.1).

5.3.5 EEG data pre-processing

We pre-processed our EEG data before the statistical analysis using BrainVision Analyzer (Brain Products, GmbH, Munich). For both groups, we re-referenced to the mastoid electrodes TP9 and TP10 and re-used the original reference channel as a data channel. For the German-Spanish group, we additionally implemented linear derivation to obtain an average HEOG signal. Next, we applied a high-pass filter of 0.1 Hz and a low-pass filter of 30 Hz. We then corrected for residual drift using a maximum amplitude of $\pm 200\ \mu$V for the HEOG channel, and $\pm 800\ \mu$V for the VEOG channel. We used ocular independent component analysis to correct for blink activity using both the VEOG and the HEOG channel as a baseline. We performed artefact correction according to the following criteria: we allowed a maximal voltage step of 50 μV/ms for the gradient, a maximal difference in 100 ms - intervals of 200 μV; maximal amplitudes of $\pm$ 200 μV, and the lowest allowable amplitude in 100 ms - intervals of 0.5 μV. Next, we segmented our data from -200 ms prior to the onset of the stimulus to 1,200 ms after the onset of the stimulus for familiar and correct trials. We applied a baseline correction to each segment using the signal in the 200 ms before stimulus onset. In a final step, we exported all available voltage amplitude samples for each time point, segment, data channel (excluding HEOG, VEOG and the reference channels) and participant to perform our statistical analysis. In this, we exported 29 data channels for the Italian-Spanish group (Fp1, Fp2, Fz, F3, F4, F7, F8, FCz, FC1, FC2, FC5, FC6, Cz, C3, C4, T7, T8, CP1, CP2, CP5, CP6, Pz, P3, P7, P4, P8, Oz, O1 and O2) and 31 channels for the German-Spanish group (Fp1, Fp2, AFz, Fz, F3, F4, F7, F8, FCz, FC3, FC4, FT7, FT8, Cz, CPz, CP3, CP4, C3, C4, T7, T8, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1 and O2). Each channel was assigned to one of the following topographic regions: left anterior, mid anterior, right anterior; left central, mid central, right central; and finally, left posterior, mid posterior and right posterior regions.

5.3.6 EEG data analysis

For the statistical analysis, we employed a data-driven approach to model *voltage amplitudes* over time. For this, we first conducted a permutation analysis to determine our region of interest (ROI) in terms of channels. Second, we used generalised additive mixed models (GAMMs) to establish our time window of interest for a potential P600 effect (Meulman et al., 2015) and to model group differences in terms of the P600 effect and CLI effects.

To determine our ROI, we performed a cluster-based permutation analysis using the *permutes* package (Voeten, 2019) in R to highlight potentially significant effects of *violation type* and *condition* on *voltage amplitudes*. We visualised the outcomes of the permutation analysis in Figure 5.3.3 for the Italian-Spanish speakers, and in Figure 5.3.4 for the German-Spanish speakers. Potentially significant effects of *violation type* and *condition* are highlighted in red colours. Note that the figure for the German-Spanish speakers is identical to Von Grebmer Zu Wolfsthurn et al. (2021a). For the Italian-Spanish group, the outcome tentatively suggested channels *C4, CP2, CP6, Pz, P3, P4, P7* and *P8* as ROI, these channels were located in centro-parietal regions with a slight left lateralisation. In contrast, for the German-Spanish group the outcome yielded *CPz, CP3, CP4, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1* and *O2* as a potential ROI. These electrodes were located in left posterior, central posterior and right posterior regions, consistent with the classical topography of the P600 component (Steinhauer et al., 2009).

Figure 5.3.3: *Permutation analysis outcome for the Italian-Spanish group (n = 29). Note that higher F-values are visualised in red colours, and lower F-values in yellow.*

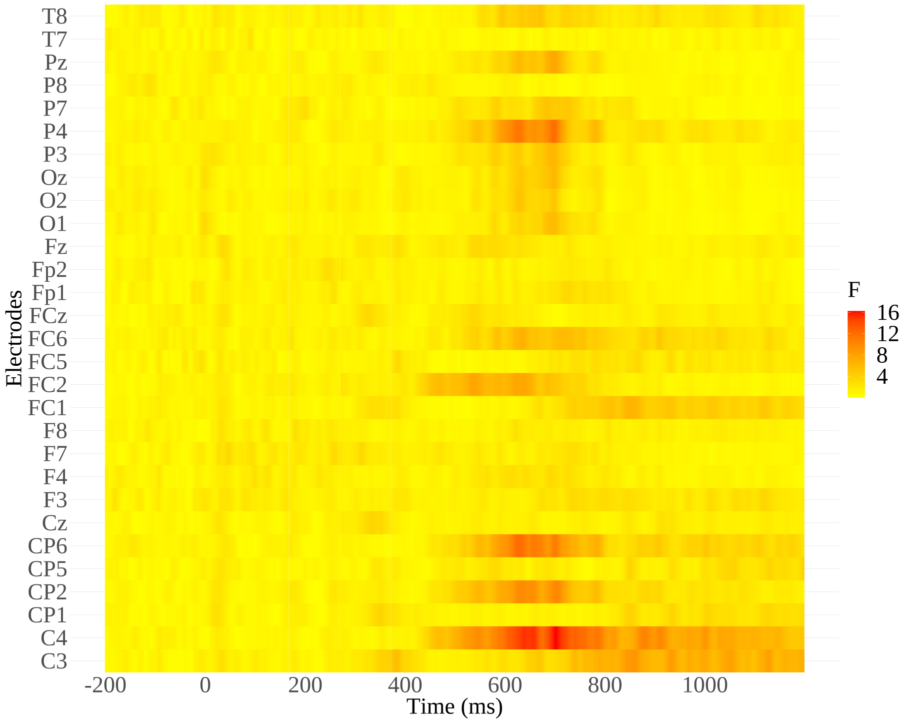
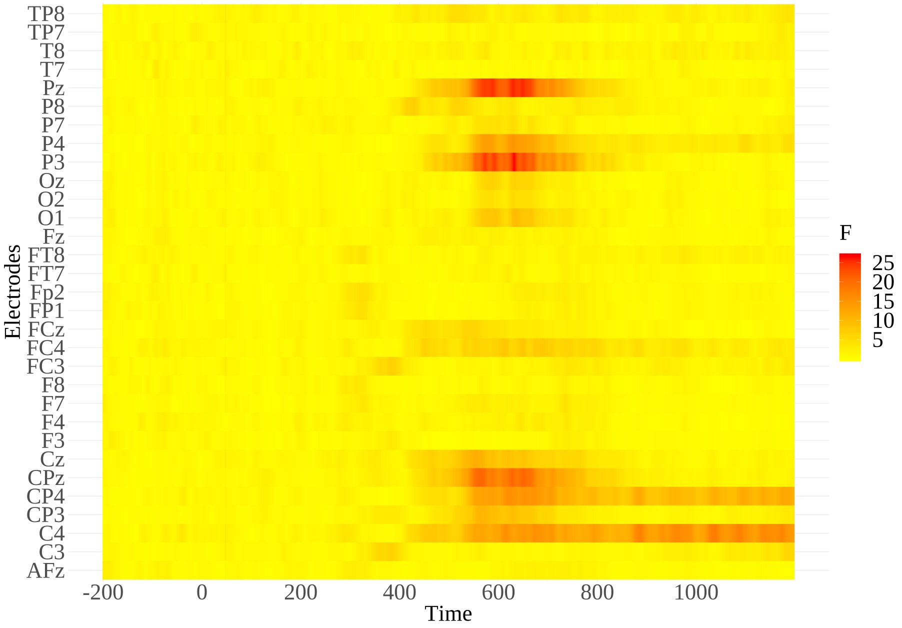


Figure 5.3.4: *Permutation analysis outcome for the German-Spanish group (n = 28). Note that higher F-values are visualised in red colours, and lower F-values in yellow.*



Pooling the ROI channels for both groups, we selected only channels which were present in the montage of each group, namely *Pz, P3, P4, P7* and *P8* as our ROI (Appendix 5.C). In a second step, we modelled *voltage amplitudes* over time in our ROI using a generalised additive mixed model (GAMM) to determine our time window of interest. A detailed discussion of this method and its application in EEG research can be found in Meulman et al. (2015) and in Tremblay and Newman (2015). Briefly, GAMMs not only allow for the inclusion of by-participant and by-item random effects (as do GLMMs), but are also robust against missing data following the missing-at-random mechanism and unbalanced observations per participant. Most importantly, GAMMs allow for the inclusion of non-linear terms to flexibly model the non-linear effects of voltage amplitudes over time, which cannot be captured with linear functions. Here, the non-linear term *time* is modelled flexibly using (penalised) splines, resulting in a smooth fit for the oscillatory trend of voltage amplitudes over time (Meulman et al., 2015). To avoid over-fitting our data, we constructed a simpler, theoretically plausible model which included the interaction effect of *L1* and *violation type*, the interaction effect of *L1* and *condition*, as well as *channel* as a covariate. Next, we added a non-linear term for *time*, and interaction effects between: *time* and *L1*, *time* and *violation type*, *time* and *condition*, and *time* and *channel*. We further created additional variables to test for our critical interaction effects over *time*, namely *L1* and *violation type*, and *L1* and *condition*. Finally, we added random intercepts for *participant* and *item*, random slopes for each participant for the effects of *time*, *violation type*, *condition* and *channel*; and random slopes for each item for the effects of *time* and *channel*. This model was fitted using the *mgcv* package (Wood, 2021) with the fast restricted likelihood estimation (fREML) using a scaled t-distribution to account for heavy tails in the residuals (Meulman et al., 2015). For storage efficiency reasons, we further applied discretisation. We carefully checked the model diagnostics for problematic residual patterns, the appropriate number of basis functions (k-parameter), the goodness of fit and for strong autocorrelation (De Cat et al., 2015). Further, we assumed missing data to be following the missing-at-random (MAR) mechanism (Ibrahim, Chen & Lipsitz, 2001).

To answer our first research question about the presence of a P600 effect in both groups, we used the *itsadug* package (Van Rij, Wieling & Baayen, 2020) in R to plot the predicted differences in voltage amplitudes for non-violation vs. violation trials separately for both groups. This also provided us with a precise time window of interest for the P600 component (Appendix 5.G). For our second research question, we generated conditional plots for the interaction effect of *L1* and *violation type* over time. Similarly, we created conditional plots for the interaction effect of *L1* and *condition* over time to tackle our third research question.

5.3.7 EEG data results

We visualised raw voltage amplitudes for our ROI for each violation type for both groups in Figure 5.3.5, which illustrates the oscillatory trend of voltage amplitudes over time. The first 250 ms post-stimulus onset show the early visual processing response typical for visual stimuli (Eulitz et al., 2000). Critically, the signal yielded a deviation in voltage amplitudes around 450 ms post-stimulus onset across both groups. Descriptively speaking, voltage amplitudes appeared lower for non-violation trials compared to violation trials between 450 ms and 900 ms post-stimulus onset in both groups in Figure 5.3.5, which tentatively suggested a P600 effect for both groups. In contrast, Figure 5.3.6 shows mean voltage amplitudes for each condition for the Italian-Spanish and the German-Spanish group. Importantly, Appendix 5.F visualises the large variance and individual differences in the EEG signal across both groups, which is a critical aspect to keep in mind when dealing with large EEG datasets.

Figure 5.3.5: *Mean voltage amplitudes over time for each violation type for channels Pz, P3, P4, P7 and P8 for both groups.*

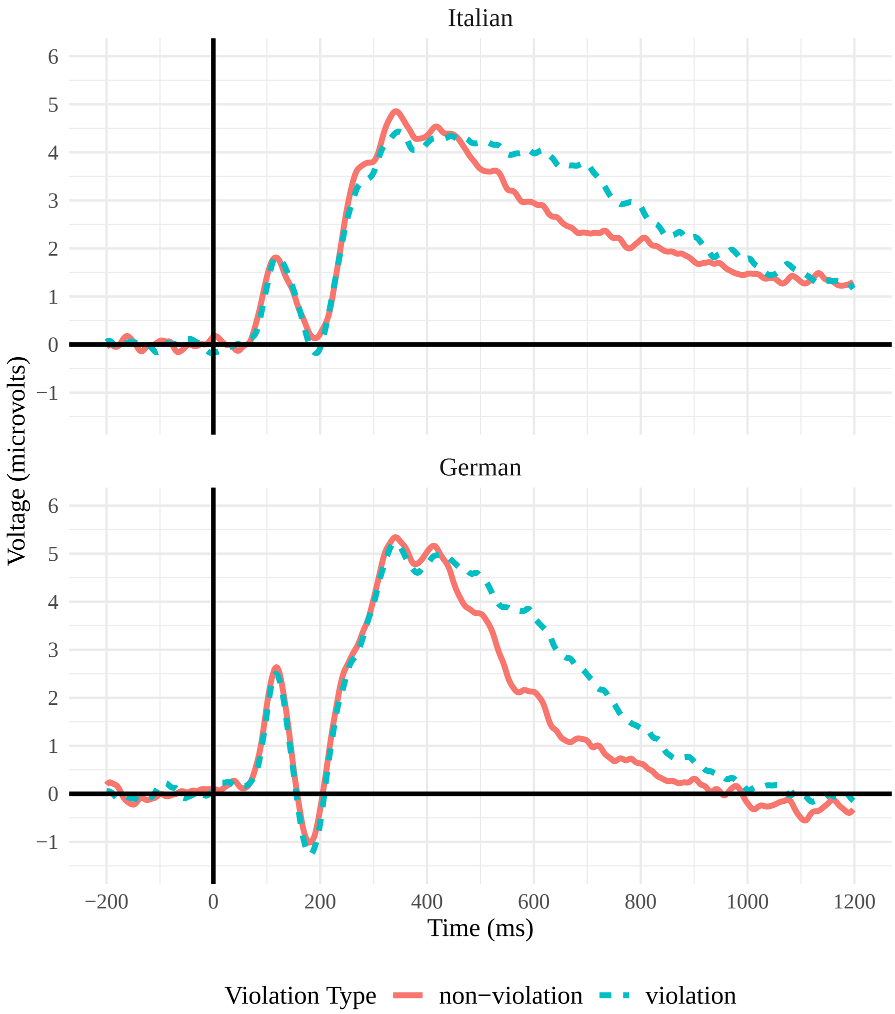
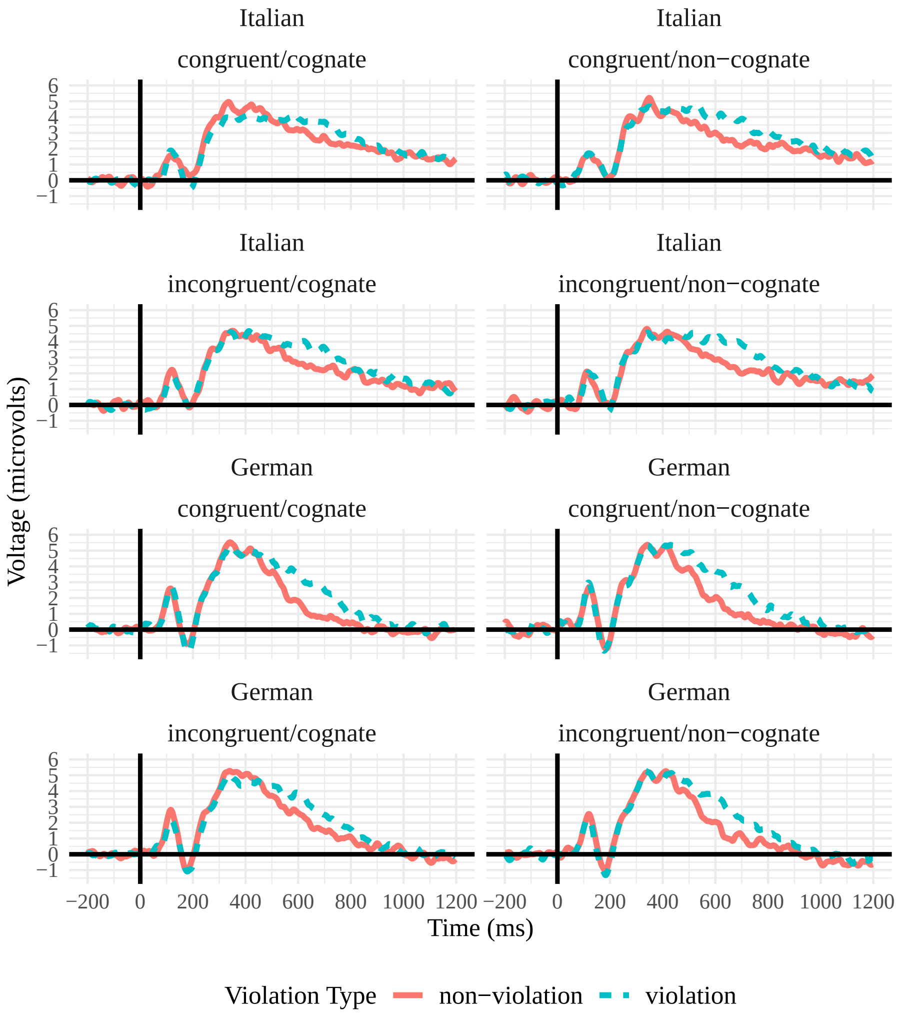


Figure 5.3.6: *Mean voltage amplitudes over time for each condition for channels Pz, P3, P4, P7 and P8 for both groups.*



As described above, our fitted GAMM model was as follows: voltage amplitudes ∼ L1 * violation type + L1 * condition + channel + s(time, k = 20) + s(time, by = L1, k = 20) + s(time, by = violation type, k = 20) + s(time, by = condition, k = 20) + s(time, by = L1 *violation type, k = 20) + s(time, by = L1 * condition, k = 20) + s(time, by = channel, k = 20) + s(participant, time, bs = "re") + s(participant, violation type, bs = "re") + s(participant, condition, bs = "re") + s(participant, channel, bs = "re") + s(participant, bs = "re") + s(item, time, bs = "re") + s(item, bs = "re")[1]. See Appendix 5.G for the exact model parameters. The model captured 9.61% of the variance in the data.

With respect to our first research question, we found a significant difference between non-violation and violation trials over time with $F = 636.46$, $p < 0.001$, which is indicative of a P600 effect. We examined this effect individually for each group and found a significant difference between non-violation and violation trials between 477.82 ms and 1056.79 ms post-stimulus onset for the Italian-Spanish group (Figure 5.3.7) and between 491.94 ms and 1056.79 ms for the German-Spanish group (Figure 5.3.8). In addition, the German-Spanish group showed a small difference at 350 ms post-stimulus onset, which is likely linked to the early visual response. Taken together, we found a P600 effect for both the Italian-Spanish group and the German-Spanish group.

[1] Our model diagnostics revealed autocorrelation and we subsequently generated a model where we corrected for this autocorrelation (De Cat et al., 2015). However, this model did not reach convergence and is therefore not reported here. Importantly, while the correction for autocorrelation may have a small impact on the model parameters, it likely does not affect the overall results.

Figure 5.3.7: *Marginal plot of predicted differences in voltage amplitudes over time for violation vs. non-violations for channels Pz, P3, P4, P7 and P8 for the Italian-Spanish group (n = 29).*

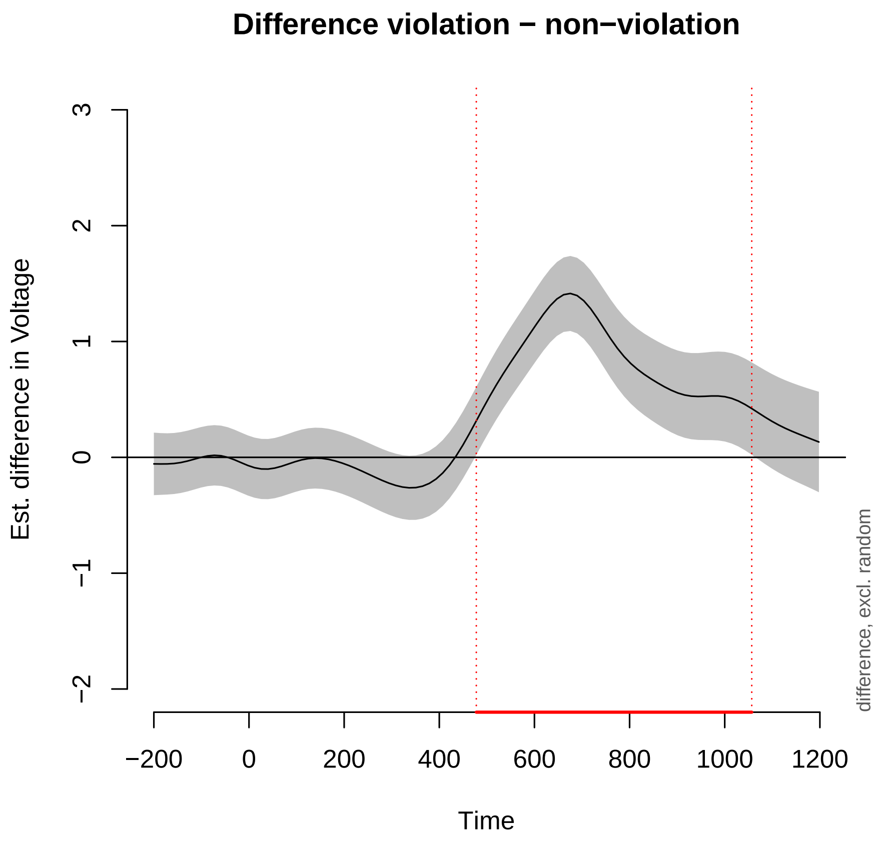
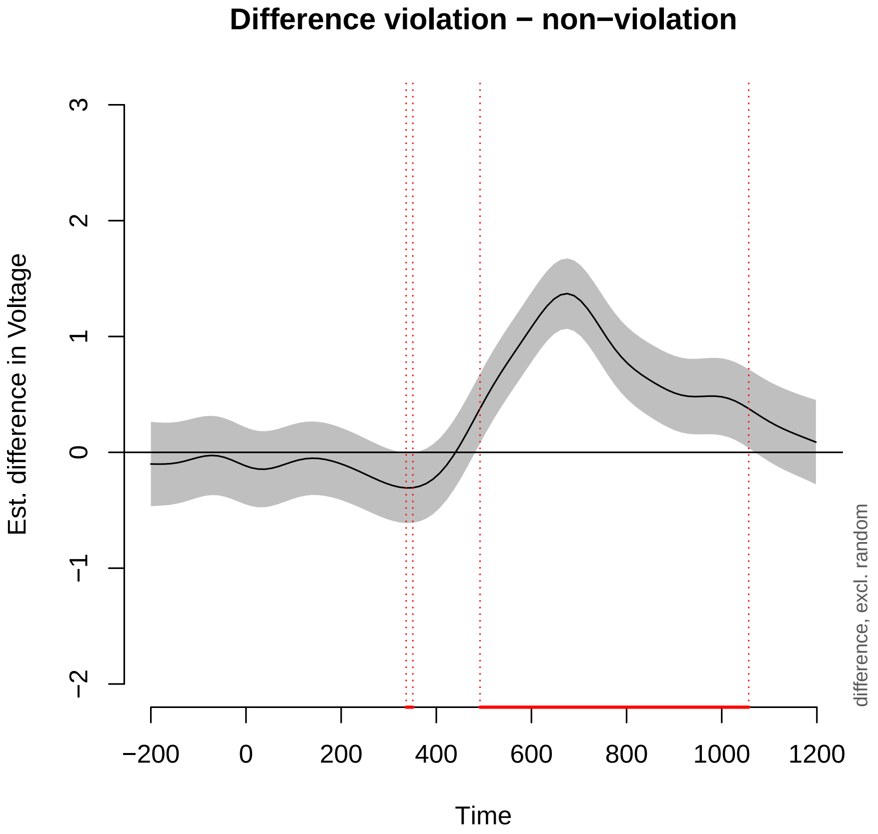
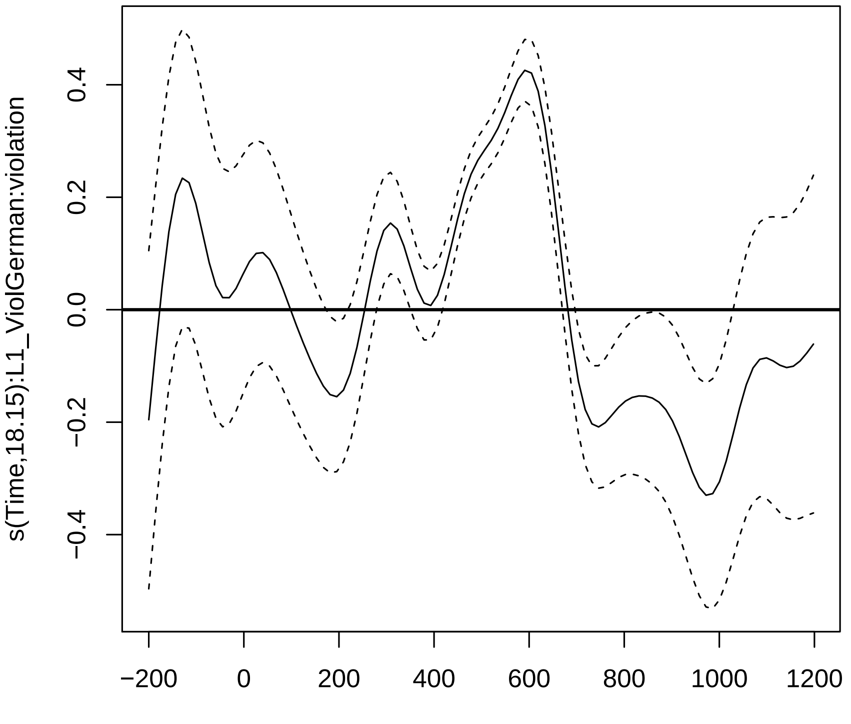


Figure 5.3.8: *Marginal plot of predicted differences in voltage amplitudes over time for violation vs. non-violations for channels Pz, P3, P4, P7 and P8 for the German-Spanish group (n = 28).*



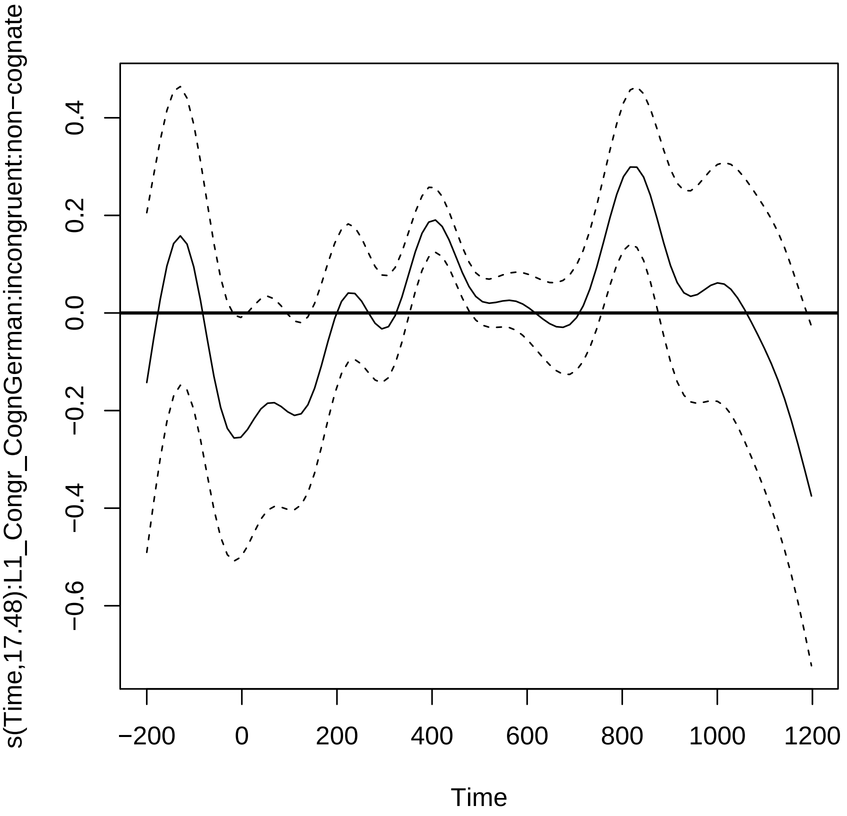
With respect to our second research question, the interaction effect of *L1* and *violation type* was significant over time with $F = 61.46$, $p < 0.001$. The conditional plot suggested a small, but robust difference in the P600 effect between the two groups (Figure 5.3.9). Figure 5.3.9 visualises this difference in voltage amplitudes between non-violation trials vs. violations trials over time for the Italian-Spanish group compared to the German-Spanish group. This figure shows a significant non-zero difference in P600 effects around 600 ms, with a larger P600 effect linked to the Italian-Spanish group compared to the German-Spanish group (Figure 5.3.9). The effect difference was close to zero for the the remaining time points and therefore not significant. Note that Figure 5.3.9 visually suggests a large difference in P600 effect size across the two groups, but was in fact much smaller as predicted by the model (Appendix 5.G). We captured this notion in Appendix 5.H, which shows this small difference in P600 effects in relation to our original scale.

Figure 5.3.9: *Conditional plot of predicted difference in voltage amplitudes over time for violations vs. non-violations for channels Pz, P3, P4, P7 and P8 across both groups (n = 57). The dashed lines represent the standard error.*



With respect to our third and final research question, the interaction effect of *L1* and *condition* was significant over time with $F = 29.30$, $p < 0.001$. This suggested that CLI effects differed over time between the groups. The conditional plot for this particular effect showed a small difference at two separate time points post-stimulus onset (Figure 5.3.10). More specifically, CLI effects were significantly larger around 400 ms and around 800 ms for the Italian-Spanish group compared to the German-Spanish group. For the remaining time points, the difference in CLI effects was close to zero and therefore not significant. Importantly, as Appendix 5.I shows, these differences in CLI effects across the two groups are small, but statistically significant according to the model. See Appendix 5.G for the exact model parameters.

Figure 5.3.10: *Conditional plot of predicted difference in voltage amplitudes over time for the CLI effects for channels Pz, P3, P4, P7 and P8 across both groups (n = 57). The dashed lines represent the standard error.*



In summary, our ERP findings were the following: first, we found evidence for a P600 effect for both groups. This was indicated by higher voltage amplitudes for violation trials compared to non-violation trials. Second, results suggested a statistically larger P600 effect around 600 ms for the Italian-Spanish compared to the German-Spanish group over time. Finally, voltage amplitudes connected to CLI effects were larger around 400 ms and around 800 ms for the Italian-Spanish compared to the German-Spanish group.

5.4 Discussion

The primary aim of the current study was to investigate typological similarity effects on cross-linguistic influence (CLI) and on the neural correlates of syntactic violation processing at the behavioural and neural level. More specifically, we examined typological similarity effects using a syntactic violation paradigm in speakers of typologically similar languages (Italian-Spanish) and of typologically less similar languages (German-Spanish), all of whom were late learners of Spanish. During the syntactic violation paradigm, we measured accuracy, RTs and voltage amplitudes over time, with a particular focus on the P600 component. We probed first, whether there was a P600 effect across both groups; second, whether this potential P600 effect was larger for one group compared to the other; and third, whether there were different CLI effects across the two groups. On the basis of the LDH (Zawiszewski & Laka, 2020) outlined in the introduction, we predicted an overall processing advantage for the Italian-Spanish group compared to the German-Spanish group.

From a behavioural perspective, we first predicted that speakers would be more accurate and faster for non-violation compared to violation trials, and for congruent/cognate items compared to incongruent/non-cognate items. Next, we hypothesised that the Italian-Spanish group would be more accurate and faster for non-violation than for violation trials compared to the German-Spanish group. Finally, we predicted that the Italian-Spanish group would be more accurate and faster at processing congruent/cognate items than for incongruent/non-cognate items compared to the German-Spanish group. This would reflect first, an advantage for speakers of typologically similar languages (Italian-Spanish) in detecting syntactic violations compared to speakers of typologically less similar languages (German-Spanish); and second, more pronounced CLI effects for the Italian-Spanish group.

Behavioural results suggested the following: for accuracy, we found that participants were indeed more accurate for non-violation trials compared to violation trials, in line with our hypothesis. Next, we also found a small effect of condition, indicating a difference in accuracy as a function of CLI. Here, participants were more accurate for congruent/cognate items compared to incongruent/cognate items, thereby suggesting a small effect of gender congruency. However, with respect to our second and third research question, results from accuracy indicated neither an influence of typological similarity on syntactic violation processing, nor on overall CLI effects as both critical interaction effects yielded non-significance.

In contrast, results from RTs provided us with a more extensive picture. Participants were faster for non-violation trials compared to violation items, and for congruent/cognate items compared to incongruent/cognate and incongruent/non-cognate items. This yields a processing advantage for congruent/cognates compared to incongruent/cognates both at the level of accuracy and RTs. One possible interpretation of this particular result could be that incongruent/cognates are potentially particularly difficult to process because of the simultaneous occurrence of similarity at the word form level and the unexpected mismatch at the gender level. Subsequently, the processing effort for incongruent/cognates may be comparatively high in contrast to cases where the similarity manifests itself at the word form level as well as at the gender level. Another critical finding was that Italian-Spanish speakers were overall slower compared to the German-Spanish speakers. This suggests a more general processing advantage for the typologically less similar German-Spanish pair compared to the Italian-Spanish pair in terms of RTs. Critically, with respect to our second research question about the differential processing of violation vs. non-violation trials across groups, we did not find evidence for this notion, contrary to our behavioural predictions. As for our third research question about differential CLI effects across groups, we found a difference in CLI across the two groups, but in the opposite direction to what we had predicted: Italian-Spanish speakers were significantly slower for incongruent/cognates compared to the German-Spanish speakers.

Taken together, the main finding from our behavioural results was the small effect of typological similarity both on overall RTs but also on CLI: the German-Spanish speakers, but not the Italian-Spanish speakers, displayed an overall behavioural processing advantage in this task. This was both in terms of faster RTs when detecting syntactic violations and in terms of overall smaller CLI effects. This notion is contrary to the predictions made by the LDH (Zawiszewski & Laka, 2020).

There are several possible interpretations of these findings: first, that there was less CLI for the German-Spanish speakers to begin with and therefore they were less subject to CLI effects compared to the Italian-Spanish speakers. Therefore, the processing advantage for the German-Spanish group could be a natural consequence of being less subject to CLI. A second interpretation is that CLI was equally pronounced in both groups, but the German-Spanish speakers employed a more efficient strategy to mitigate CLI effects compared to the Italian-Spanish speakers. Finally, the predictions of the LDH (Zawiszewski & Laka, 2020) may be limited to morphosyntactic similarity and may not apply to similarity at the level of gender and word form overlap as tested in this current study, at least in terms of behaviour. Our current design does not allow for the discrimination of these interpretations, but they should be subject to future research. Nevertheless, these results provide evidence for an effect of typological similarity on CLI favouring speakers of typologically less similar languages. To get a clearer interpretation of our findings, in the next section we corroborated these behavioural findings with the ERP findings.

In terms of ERPs, we first expected a P600 effect for both groups. In line with our predictions, we found significantly higher voltage amplitudes for violation trials compared to non-violation trials for both the Italian-Spanish and the German-Spanish group, which reflects the classical P600 effect (Friederici et al., 1999, 2002; Sabourin & Stowe, 2008; Swaab et al., 2011). In turn, this indicated that both groups were highly sensitive to syntactic violations at the level of gender. Notably, both groups displayed a highly similar onset of the P600 effect around 490 ms post-stimulus onset, as well as a comparable P600 effect latency until around 1,000 ms post-stimulus onset. Therefore, answering to our first research question, our data suggest a P600 effect for both the Italian and the German late learners of Spanish (S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005).

For our second research question, we predicted a larger, more native-like P600 effect for the typologically more similar Italian-Spanish group compared to the typologically less similar German-Spanish group, in line with the LDH (Zawiszewski & Laka, 2020). Supporting this prediction, our data provided evidence for a small, but robust statistical difference in P600 effect sizes (around 600 ms post-stimulus onset), with a larger P600 effect for the Italian-Spanish group than for the German-Spanish group. This indicates a processing advantage for typologically more similar languages compared to less similar languages. Further, these findings corroborate the results by Sabourin and Stowe (2008), who reported a larger P600 effect for the typologically more similar language combination of German and Dutch compared to the combination of Romance languages and Dutch when processing syntactic violations in the non-native language Dutch. By extension, the results from our study support the notion of enhanced sensitivity to syntactic violations in speakers of typologically more similar languages compared to less similar languages, i.e., Italian-Spanish vs. German-Spanish, see also Sabourin and Stowe (2008). Therefore, as for our second research question, we provide evidence that typological similarity directly impacts P600 effect sizes. This notion expands on work by Zawiszewski and Laka (2020), who demonstrated a modulation of ERP effects by morphological similarity in highly proficient speakers. Therefore, our study contributes novel findings about the facilitatory role of gender similarity and word form similarity to existing accounts on the role of morphosyntactic similarity on non-native comprehension.

For our third research question, we predicted larger CLI for the Italian-Spanish group compared to German-Spanish group, as reflected in larger, more native-like voltage amplitudes for CLI effects for the typologically more similar group. In line with our predictions, we found that CLI effects were larger for the Italian-Spanish group compared to the German-Spanish group. Subsequently, this represents evidence for a modulation of CLI by typological similarity, as well as a processing advantage for typologically similar languages compared to less similar languages. These results extend the LDH by Zawiszewski and Laka (2020) in that we provide evidence that also similarity at the level of gender and orthographic and phonological form overlap (cognate status) elicited more native-like, larger ERP components. Moreover, these results are also in line with studies suggesting overall larger CLI for typologically similar languages compared to less similar languages (Mosca, 2017; Sabourin & Stowe, 2008; Tolentino & Tokowicz, 2011).

Taking both the behavioural and the ERP data together, results suggested a general typological similarity effect on non-native comprehension. Interestingly, however, they indicated a typological effect in opposite directions: on the one hand, the behavioural data suggested a behavioural processing disadvantage for the Italian-Spanish group in the form of overall slower RTs and slower RTs for processing CLI compared to the German-Spanish group. This contrasts with our predictions on the basis of the LDH (Zawiszewski & Laka, 2020). In turn, it could imply that the model's behavioural predictions were only applicable to morphosyntactic similarity but not to overlap at the level of gender, orthography and phonology. On the other hand, the ERP data suggested a processing advantage for the Italian-Spanish group at the neural level, with larger and more native-like P600 effects and larger CLI effects compared to the German-Spanish group. These results support the predictions of the LDH (Zawiszewski & Laka, 2020). Our interpretation is that the notion of larger, more native-like ERPs for similar languages holds not only for morphological similarity, but also for gender system similarity, and orthographic and phonological word form similarity.

Differential findings across behavioural data and ERP data are not uncommon in the non-native language processing literature

(Acheson et al., 2012; Bosma & Pablos, 2020; Jiao et al., 2020). In this current study, behavioural findings support a processing advantage for typologically less similar languages, whereas neural findings support a processing advantage for typologically similar languages. Critically, we argue that this contrasts highlights the complex association between behavioural and neural cognitive mechanism, which goes far beyond the more traditional interpretation that neural measures index ongoing processes and behavioural measures index the outcomes of those processes (White, Genesee & Steinhauer, 2012). Moreover, our contrasting results could also indicate that typological similarity effects differ not only across behavioural and neural measures, but potentially also in terms of the different linguistic domains, such as phonological similarity, orthographic similarity or lexico-semantic similarity. Our study design and research questions did not allow for a more nuanced investigation of whether the typological similarity effect is in fact an interplay between several similarity effects across different domains. Therefore, more refined research is needed first, to tease apart a potentially differential impact of typological similarity on behaviour and neural correlates; and second, to characterise typological similarity effects not as a unified effect, but as a combination of individual similarity effects.

Another direction for future research is concerned with examining the exact role of proficiency in modulating typological similarity effects more closely. As discussed in the introduction, some studies suggested more pronounced typological similarity effects at lower proficiency levels (Tokowicz & MacWhinney, 2005). However, more direct comparisons are needed between typological similarity effects at different levels of non-native proficiency and AoA. This was beyond the scope of the current study, but will be essential for characterising the role of typological similarity on non-native processing more broadly and to model the potentially dynamic effects of typological similarity over time with evolving proficiency levels.

Returning to our broader question of whether typological similarity impacts non-native processing, the results of this study suggest an affirmative answer. In turn, this notion indicates that the L1 and the non-native language are intrinsically linked with each other in our late language learners at this specific proficiency level. However, since studies on this particular topic are scarce, we argue for the need of more comprehensive studies to tackle this question in a more nuanced manner.

5.4.1 Conclusions

In this study, we investigated typological similarity effects in non-native comprehension in Italian-Spanish speakers (typologically similar group) and German-Spanish speaker (typologically less similar pair). On the basis of the Language Distance Hypothesis, LDH (Zawiszewski & Laka, 2020), we predicted a processing advantage for speakers of the typologically more similar language pair, as reflected in higher accuracy, shorter RTs and larger, more native like P600 amplitudes during a syntactic violation paradigm. We found different typological similarity effects: on the one hand, the Italian-Spanish speakers were overall slower during the task compared to the German-Spanish speaker. On the other hand, ERP evidence showed a larger P600 effect for the Italian-Spanish speakers as well as larger voltage amplitudes for CLI compared to the German-Spanish speakers. This latter finding was in line with the LDH (Zawiszewski & Laka, 2020). Therefore, our results indicate a general typological similarity effect at the level of both behavioural and neural measures. Moreover, our results suggest an intimate functional link between the L1 and the non-native language in the multilingual brain. Questions remain as to whether typological similarity effects are uniform across behavioural and neural measures and whether they are equally pronounced across different linguistic domains and proficiency levels.

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 all members of the Neurolinguistics Laboratory at the University of Konstanz for their support, in particular Carsten Eulitz, Oleksiy Bobrow, Charlotte Englert, Khrystyna Oliynyk, Anna-Maria Waibel, Zeynep Dogan and Fangming Zhang. We also thank Núria Sebastián Gallés and Ege Ekin Özer from the Speech Acquisition and Perception group at the Center for Brain and Cognition at Pompeu Fabra University. We further thank Marjolein Fokkema, Julian Karch and Franz Wurm for their support in the statistical analyses. A special thanks goes to Frank Doornkamp and Willem Hoogerworst, who consulted on the analysis for the EEG data and implemented the initial GAMM analysis. The statistical analyses were performed using the computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. Finally, we also want to thank all of our participants for taking part in this research.

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/e6acy/?view_only=798087b3751b46d88654f76eaf26ec67`

Citation diversity statement

We included this statement to make readers aware of research suggesting that authors identifying as female or as members of a minority group are underrepresented in the reference list of scientific studies (Dworkin et al., 2020; Zurn et al., 2020). This is a particularly important topic to consider during a global pandemic (Viglione, 2020). Our references included 23% woman/ woman authors, 36% man/ man, 24% woman/ man and finally, 8% man/ woman authors. 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 in the reference list of publications in the field of neuroscience (Dworkin et al., 2020).

Appendix

5.A Linguistic profile: Italian-Spanish group

Table 5.A.1: *Overview of the native and non-native languages acquired by the Italian-Spanish group (n = 29).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Italian	n = 29					**29**
Spanish		n = 2	n = 17	n = 8	n = 2	**29**
English		n = 24	n = 4			**28**
French		n = 3	n = 6	n = 3		**12**
German			n = 1	n = 1		**2**
Catalan				n = 1	n = 1	**2**
Portuguese					n = 3	**3**
Total	**29**	**29**	**28**	**13**	**6**	

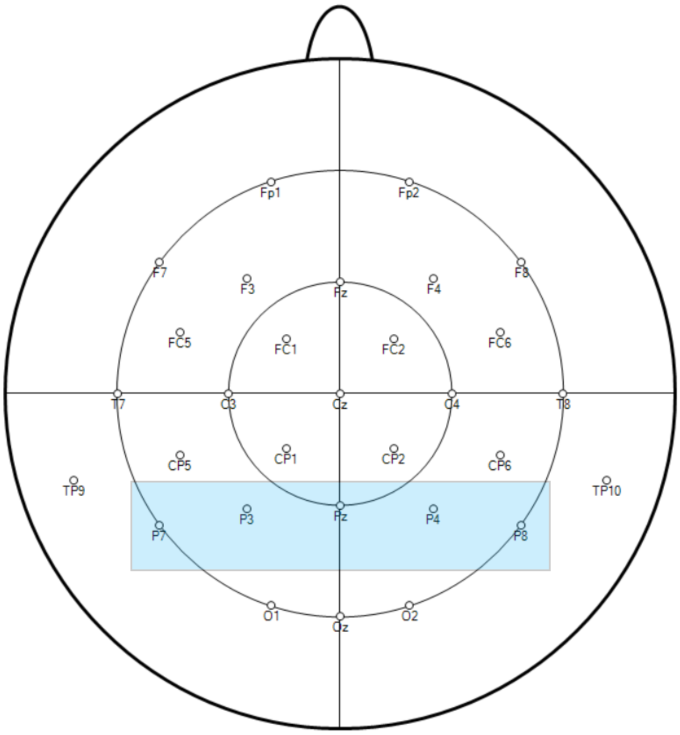
5.B Linguistic profile: German-Spanish group

Table 5.B.1: *Overview of the native and non-native languages acquired by the German-Spanish group (n = 28).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
German	n = 28					**28**
Spanish			n = 15	n = 11	n = 2	**28**
English		n = 26	n = 2			**28**
French		n = 2	n = 8	n = 5		**15**
Latin			n = 2	n = 1	n = 1	**4**
Russian			n = 1		n = 1	**2**
Swedish				n = 1		**1**
Portuguese					n = 1	**1**
Arabic					n = 1	**1**
Catalan					n = 1	**1**
Italian					n = 1	**1**
Mandarin					n = 1	**1**
Total	**28**	**28**	**28**	**18**	**9**	

5.C EEG data: region of interest

Figure 5.C.1: *Region of interest and the corresponding channels Pz, P3, P4, P7 and P8 for the EEG analysis, shown in the shaded area in the montage of the Italian-Spanish group.*



5.D Model parameters: accuracy

Table 5.D.1: *Model parameters for the best-fitting model for accuracy (n = 57).*

Formula: accuracy $\sim$ L1 (Italian vs. German) + violation type (violation vs. non-violation) + L1 * condition (congruent/cognate vs. congruent/non-cognate vs. incongruent/cognate vs. incongruent/non-cognate) + LexTALE-Esp score + target noun gender (feminine vs. masculine) + (violation type|participant) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	65.65 [32.57, 132.35]	11.70	< 0.001
L1 [German]	1.50 [0.673, 3.35]	0.993	0.321
Violation type [violation]	0.412 [0.279, 0.609]	-4.45	< **0.001**
Condition [congruent/non-cognate]	1.63 [0.752, 3.54]	1.24	0.215
Condition [incongruent/ cognate]	0.258 [0.132, 0.504]	-3.96	< **0.001**
Condition [incongruent/ non-cognate]	0.714 [0.342, 1.49]	-0.897	0.370
LexTALE-Esp score	1.02 [1.01, 1.03]	4.15	< 0.001
Target noun gender [m]	0.686 [0.478, 0.986]	-2.04	0.042
L1 [German] * Condition [congruent/ non-cognate]	0.349 [0.121, 1.01]	-1.94	0.052
L1 [German] * Condition [incongruent/ cognate]	1.68 [0.631, 4.46]	1.04	0.300
L1 [German] * Condition [incongruent/ non-cognate]	0.661 [0.238, 1.84]	-0.792	0.428

Random effects	
σ^2	3.29
$\tau_{00 Item}$	1.77
$\tau_{00 Participant}$	0.36
$\tau_{11 Participant[non-violation]}$	0.16
$\rho_{01 Participant}$	-0.35
ICC	0.39
$N_{Participant}$	57
N_{Item}	448
Observations	9,972
Marginal R^2/ Conditional R^2	0.111/0.461

5.E Model parameters: response times

Table 5.E.1: *Model parameters for the best-fitting model for RTs (n = 57).*

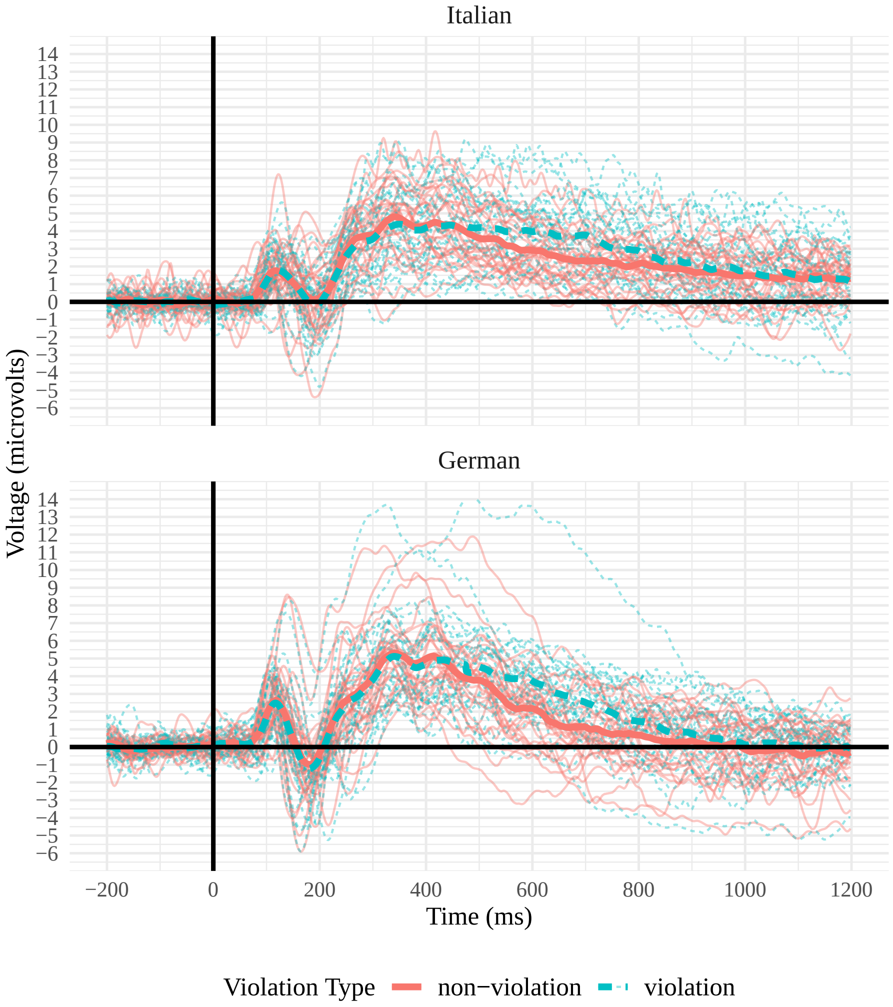
Formula: RTs ~ L1 (Italian vs. German) + violation type (violation vs. non-violation) + L1 * condition (congruent/cognate vs. congruent/non-cognate vs. incongruent/cognate vs. incongruent/non-cognate) + target noun gender (feminine vs. masculine) + (violation type|participant) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	854.22 [791.17, 917.26]	26.56	< 0.001
L1 [German]	-82.55 [-100.54, -64.56]	-8.99	$<$ **0.001**
Violation type [violation]	128.18 [93.91, 162.45]	7.33	$<$ **0.001**
Condition [congruent/ non-cognate]	-2.05 [-47.75, 43.65]	-0.088	0.930
Condition [incongruent/ cognate]	105.64 [89.31, 121.98]	12.67	$<$ **0.001**
Condition [incongruent/ non-cognate]	36.02 [25.36, 46.67]	6.62	$<$ **0.001**
Target noun gender [m]	14.47 [-9.69, 38.63]	1.17	0.241
L1 [German] * Condition [congruent/ non-cognate]	0.297 [-49.29, 49.88]	0.012	0.991
L1 [German] * Condition [incongruent/ cognate]	-63.19 [-102.49, -23.89]	-3.15	**0.002**
L1 [German] * Condition [incongruent/ non-cognate]	-27.28 [-71.27, 16.71]	-1.22	0.224
Random effects			
σ^2	0.14		
$\tau_{00 Item}$	3966.19		

$\tau_{00 Participant}$	8558.85
$\tau_{11 Participant[non-violation]}$	5839.45
$\rho_{01 Participant}$	-0.18
ICC	1.00
$N_{Participant}$	57
N_{Item}	448
Observations	9,393
Marginal R^2/ Conditional R^2	0.359/1.00

5.F EEG data: by-violation type mean voltage amplitudes

Figure 5.F.1: *Mean voltage amplitudes over time for each violation type for each participant for channels Pz, P3, P4, P7 and P8 (n = 57). Mean amplitudes for violation type are shown as thicker lines.*



5.G Model parameters: P600 component

Table 5.G.1: *Model parameters of the GAMM model for the effect of L1 and time on voltage amplitudes for channels Pz, P3, P4, P7 and P8 (n = 57). 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 ∼ L1 * violation type + L1 * condition + channel + s(time, k = 20) + s(time, by = L1, k = 20) + s(time, by = violation type, k = 20) + s(time, by = condition, k = 20) + s(time, by = L1 * violation type, k = 20) + s(time, by = L1 * condition, k = 20) + s(time, by = channel, k = 20) + s(participant, time, bs = "re") + s(participant, violation type, bs = "re") + s(participant, condition, bs = "re") + s(participant, channel, bs = "re") + s(participant, bs = "re") + s(item, time, bs = "re") + s(item, bs = "re")

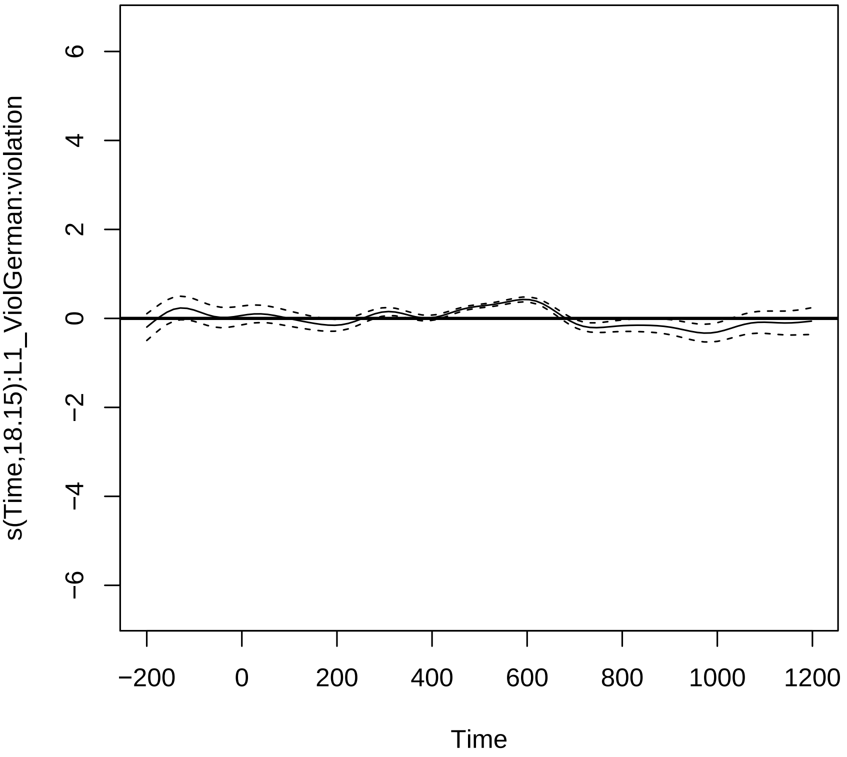
Linear terms	**Estimate**	**SE**	**t-value**	**p-value**
(Intercept)	2.32	0.243	9.58	< 0.001
L1 [German]	-0.541	0.301	-1.80	0.072
Violation type [violation]	0.328	0.151	2.17	0.029
Condition [congruent/ non-cognate]	0.044	0.171	0.255	0.798
Condition [incongruent/ cognate]	0.003	0.171	0.018	0.985
Condition [incongruent/ non-cognate]	0.083	0.181	0.460	0.645
Channel [P4]	0.490	0.176	2.78	0.005
Channel [P7]	-2.18	0.176	-12.39	< 0.001
Channel [P8]	-1.20	0.176	-6.81	< 0.001
Channel [Pz]	0.496	0.176	2.82	0.005
L1 [German] * Violation type [violation]	-0.044	0.214	-0.207	0.863
L1 [German] * Condition [congruent/ non-cognate]	0.039	0.220	0.176	0.860

L1 [German] * Condition [incongruent/ cognate]	-0.016	0.220	-0.071	0.943
L1 [German] * Condition [incongruent/ non-cognate]	-0.152	0.250	-0.608	0.543
Non-linear terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Time)	17.97	18.00	3971.89	< 0.001
s(Time) * L1 [German]	18.89	18.99	653.49	< 0.001
s(Time) * Violation type [violation]	17.86	18.77	636.46	< **0.001**
s(Time) * Condition [congruent/ non-cognate]	17.51	18.66	25.98	< 0.001
s(Time) * Condition [incongruent/ cognate]	17.73	18.74	27.93	< 0.001
s(Time) * Condition [incongruent/ cognate]	17.98	18.76	25.37	< 0.001
s(Time) * [German/ violation]	18.15	18.85	61.46	< **0.001**
s(Time) * [German/ incongruent/ non-cognate]	17.48	18.62	29.30	< **0.001**
s(Time) * Channel [P3]	18.88	19.00	249.47	< 0.001
s(Time) * Channel [P4]	1.00	1.00	43.60	< 0.001
s(Time) * Channel [P7]	18.98	19.00	2619.93	< 0.001
s(Time) * Channel [P8]	18.98	19.00	1385.98	< 0.001
s(Time) * Channel [Pz]	18.98	19.00	474.41	< 0.001

s(Time, Participant	54.98	55.00	16133637.33	< 0.001
s(Violation type, Participant)	71.20	114.00	1476223.16	1.00
s(Condition, Participant)	174.60	226.00	275408.80	1.00
s(Channel, Participant)	252.74	284.00	1563283.21	1.00
s(Participant)	0.003	57.00	0.166	1.00
s(Time, Item)	437.67	441.00	77707.90	0.017
s(Item)	432.69	444.00	75336.34	0.217

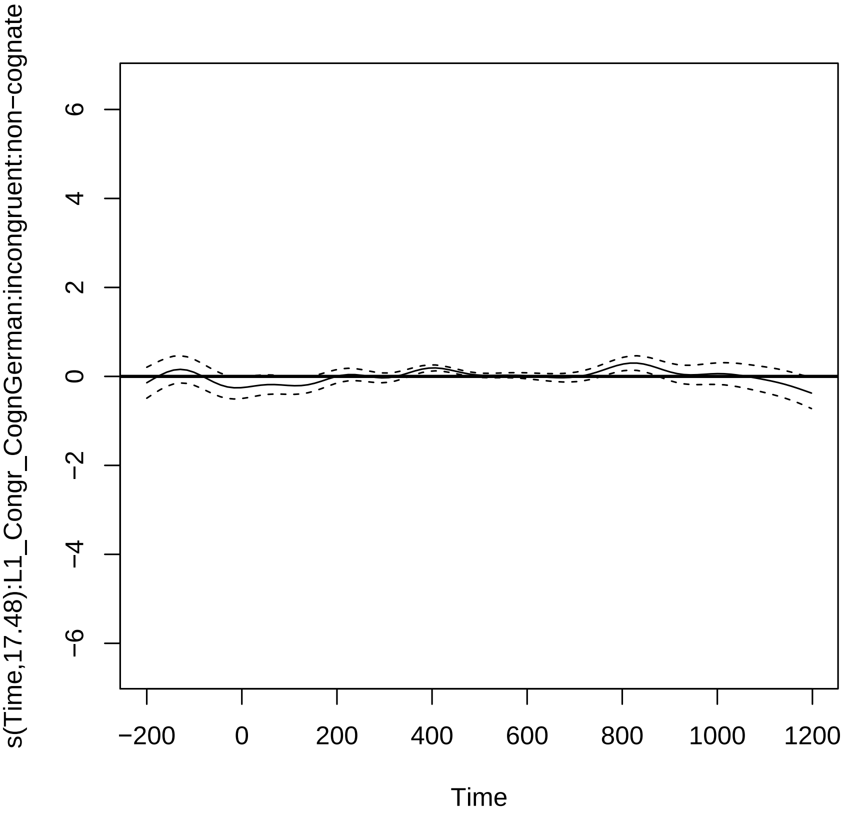
5.H P600 effect sizes: unscaled predicted differences

Figure 5.H.1: *Conditional plot of predicted difference in voltage amplitudes over time for violation vs. non-violations for channels Pz, P3, P4, P7 and P8 across both groups (n = 57) on the original scale. The dashed lines represent the standard error.*



5.I CLI effect sizes: unscaled predicted differences

Figure 5.I.1: *Conditional plot of predicted difference in voltage amplitudes over time for the CLI effects for channels Pz, P3, P4, P7 and P8 across both groups (n = 57) on the original scale. The dashed lines represent the standard error.*



CHAPTER 6

Does your native language matter? Neural correlates of typological similarity in non-native production

This article is published as: Von Grebmer Zu Wolfsthurn, S., Pablos, L., & Schiller, N. O. (2022). Does your native language matter? Neural correlates of typological similarity in non-native production. Lingue e Linguaggio, 21(1), 143-169.

Abstract: Cross-linguistic influence (CLI) and typological similarity are key features in multilingual language processing. Here, we study whether CLI effects in language production are more pronounced in typologically similar vs. dissimilar languages in late language learners. In a picture-naming task, we manipulated gender congruency and cognate status as indices for CLI in a group of Italian learners of Spanish and a group of German learners of Spanish. Further, we explored modulations of P300 amplitudes indexing inhibitory control. Behaviourally, we observed effects of CLI, but not of typological similarity. At the neural level, P300 amplitudes were modulated by CLI effects. However, we did not find evidence for a typological similarity effect on P300 amplitudes. Therefore, our results suggest a limited role of typological similarity. This study has crucial implications for non-native language production mechanisms in light of the similarity between the native and the non-native language.

Keywords: *typological similarity, non-native production, cross-linguistic influence, P300 ERP effect, late language learners*

6.1 Introduction

Anecdotally, multilingual language learners sometimes describe learning a particular language as “easy” because their native language (L1) is “similar” to the non-native language they are acquiring. Here, the term “multilinguals” describes individuals from diverse linguistic backgrounds who have acquired two or more languages at varying proficiency levels (Cenoz, 2013). In turn, proficiency refers to the extent to which language abilities match the age-based standard in comparison to native speaker (Bedore et al., 2012). Beyond anecdotal accounts, the notion of similarity between the L1 and the non-native language refers to *typological similarity*, i.e., the structural cross-linguistic similarities with respect to lexico-semantics and morphosyntax (Foote, 2009; Rothman & Cabrelli Amaro, 2010). For example, Italian and Spanish may be considered as more typologically similar than German and Spanish due to more overlap in terms of morphosyntax and lexicon (Schepens et al., 2012). The question of how much typological similarity affects language processing is crucial because it directly addresses the debate of the functional organisation of the L1 and a non-native language in the multilingual brain (Costa, Heij & Navarrete, 2006; Tolentino & Tokowicz, 2011; Zawiszewski & Laka, 2020).

At the root of so-called *typological similarity effects* are the shared cognitive representations and neurocognitive resources of the L1 and the non-native language (MacWhinney, 2005). Several neuroimaging and electroencephalographic (EEG) studies on highly proficient speakers suggest that more neurocognitive resources are shared across the L1 and the non-native language for typologically similar linguistic features compared to typologically less similar features (De Diego Balaguer, Sebastián-Gallés, Díaz & Rodríguez-Fornells, 2005; Jeong et al., 2007). In turn, these shared representations are linked to language co-activation and *cross-linguistic influence*, hereafter CLI (Lago et al., 2021; Nozari & Pinet, 2020). CLI refers to the bi-directional influence of the L1 and the non-native language on the underlying processing mechanisms. Moreover, CLI effects are found for different ages of acquisition (AoA) and may be larger and more pronounced at lower non-native proficiency levels (Heidlmayr et al., 2021; MacWhinney, 2005; Ringbom, 1987). Connected to typological similarity, research also suggests a link between high typological similarity and increased CLI (Cenoz, 2001; Costa, Heij & Navarrete, 2006; Yamasaki et al., 2018). Relevant for the purpose of our study, CLI was found to significantly impact non-native production (Lemhöfer et al., 2008; Paolieri et al., 2019). Multilinguals are said to recruit a language control network to manage language co-activation and CLI, in order to select the appropriate target language and obtain successful communication (D. W. Green, 1998; Stocco et al., 2014). A prominent theoretical framework for language control is the Inhibitory Control (IC) model (D. W. Green, 1998), which postulates the suppression of the non-target language prior to any linguistic output. Further, some evidence also suggests that language control forms part of domain-general cognitive control (Declerck et al., 2021) paramount to everyday functioning.

6.2 Background

A crucial question at this point is the following: do typologically similar languages bear a processing advantage over typologically less similar languages, or does typological similarity between languages result in a processing disadvantage instead? The relevant literature remains inconclusive with respect to this particular question, as will be discussed below. The *Conditional Routing Model* (CRM) by Stocco et al. (2014) proposes that increased CLI for typologically similar languages effectively trains and strengthens the control network throughout development (Yamasaki et al., 2018). This implies that speakers of typologically similar languages develop overall superior cognitive control skills to mitigate CLI effects and therefore have an advantage over speakers of typologically less similar languages (Declerck, Koch, Duñabeitia, Grainger & Stephan, 2019; Yamasaki et al., 2018; Zawiszewski & Laka, 2020). In this study, we used this theoretical framework to generate testable predictions about the role of typological similarity in non-native production. Participants were late learners and speakers of languages with differing degrees of typological similarity. We defined late learners as speakers who have acquired Spanish at a later age (AoA $>$ 14 years), and who have not yet reached high proficiency levels. We tested Italian-Spanish speakers for the typologically similar language pair, and German-Spanish speakers for the typologically less similar language pair. To directly probe the influence of typological similarity on non-native production, we examined its effects on two features shown to be prone to CLI from a behavioural and neural perspective: grammatical gender (hereafter gender) and cognates. Both of these CLI effects offer a unique window into the mechanisms of non-native production in light of differing degrees of typological similarity. According to the CRM, we should find that Italian-Spanish speakers show overall smaller CLI effects compared to the German-Spanish speakers as a function of increased training in mitigating CLI effects over time. This would imply that speakers of typologically similar languages indeed bear a processing advantage over speakers of typologically less similar languages.

The first CLI effect we explored in this study was the *gender congruency effect*. It was previously proposed to reflect CLI of the gender systems (Costa et al., 2003; Lemhöfer et al., 2008). Relevant to this study, Italian and Spanish both feature a feminine and masculine gender value, marked by *la* and *il* in Italian, and *la* and *el* in Spanish, respectively. In contrast, German has a three-way gender system characterised by masculine, feminine and neuter marked by *der*, *die* and *das*, respectively (Schiller & Caramazza, 2003). The core feature of the gender congruency effect is that *congruent* items, i.e., items belonging to the same nominal gender category across languages (e.g., *die*$_F$ *Kerze*$_F$– *la*$_F$ *vela*$_F$ [the candle] in German and Spanish), are processed more accurately and faster compared to *incongruent* items, i.e., items mismatching in gender categories across languages (e.g., *der*$_M$ *Schlüssel*$_M$ – *la*$_F$ *llave*$_F$ [the key] in German and Spanish)[1]. To this date, few studies have investigated the effect of typological similarity on gender processing in language production. Further, those who have studied it have in many cases looked at highly proficient speakers, and have not always obtained consistent results. For example, a study by Paolieri et al. (2019) explored the gender congruency effect in a translation task in highly proficient Italian-Spanish (typologically similar pair) and Russian-Spanish (typologically dissimilar pair) speakers. Results revealed a gender congruency effect in both groups, but the effect was more consistently found in the Italian-Spanish group than in the Russian-Spanish group. This finding suggested that typological similarity facilitated gender processing in Italian-Spanish speakers compared to Russian-Spanish speakers, in line with the CRM account (Stocco et al., 2014). By contrast, Costa et al. (2003) conducted a picture-naming study with highly proficient Spanish-Catalan, Catalan-Spanish, Italian-French and Croatian-Italian speakers to study the gender congruency effect. The authors found no evidence for a gender congruency effect, and therefore argued that typological similarity may not play a role in non-native production (see also Costa et al., 2006). This evidence therefore suggests a limited effect of typological similarity on non-native production, in contrast to the CRM account.

The second CLI effect examined in the context of typological similarity was the *cognate facilitation effect.* It reflects CLI of orthographic and phonological systems (Lemhöfer et al., 2008; Peeters et al., 2013). *Cognates*, i.e., items with a large semantic, phonological and orthographic form overlap across languages (e.g., *Tomate*

[1] Note that we are not assuming that gender values are conceptually identical across languages. However, as a result from explicit instructions in language courses, learners may perceive some gender values as sufficiently similar to each other.

– *tomate* [tomato] in German and Spanish), were found to be processed faster and more accurately compared to *non-cognates* (e.g., *Erdbeere* – *fresa* [strawberry]). Italian and Spanish share a larger amount of cognates compared to German-Spanish (Schepens et al., 2012). In turn, this suggests a strong structural overlap for the typologically similar Italian-Spanish pair. Therefore, the cognate facilitation effect per se provisionally suggests a processing advantage for orthographically similar and form-related structures compared to more less similar structures. To the authors' knowledge, no study has previously directly investigated the role of typological similarity in cognate production.

In sum, current evidence suggests first, a processing advantage for typologically similar structures (e.g., gender congruent items and cognates) compared to typologically dissimilar structures (e.g., gender incongruent items and non-cognates). Second, studies on the gender congruency effect suggest a tentative trend towards a production advantage for typologically similar languages compared to typologically less similar languages in highly proficient speakers, supporting the CRM (Stocco et al., 2014). However, in light of the conflicting results with respect to the gender congruency effect and typological similarity and the lack of direct evidence on the role of typological similarity on cognate production, we systematically explored typological similarity effects on both gender and cognate processing in this EEG study.

EEG and event-related potentials (ERPs) are critical tools in examining the temporal unfolding of the cognitive mechanisms underlying multilingual language processing (D. W. Green & Kroll, 2019). In this study, we probed the neural correlates of the typological similarity effect by focusing on the P300 component. This ERP component was previously linked to more general cognitive mechanisms such as inhibitory control, working memory and allocation of attentional resources in control network paradigms, such as language switching or the Flanker task (Declerck et al., 2021; González Alonso et al., 2020; Polich, 2007). In light of its functional involvement in control network processes, and because non-native production heavily relies both on the successful mitigation of CLI effects and on language control to inhibit the non-target language prior to articulation, the P300 component is the most relevant ERP component for our study. We predicted the P300 component to be an index of CLI effect mitigation where any differences in P300 amplitudes across groups could reflect a typological similarity effect. To the authors' knowledge, only the study by Von Grebmer Zu Wolfsthurn et al. (2021b) reported a P300 effect in an overt non-native picture-naming task. The present study is the first ERP study exploring whether typological similarity effects and CLI effects are detectable at the neural level in the form of distinct P300 correlates in non-native production.

The aim of this study was twofold: first, we systematically studied the gender congruency effect and the cognate facilitation effect in non-native language production in late learners with intermediate proficiency levels. Secondly, and more importantly, we investigated the effect of typological similarity on CLI in two language pairs with differing degrees of typological similarity: Italian-Spanish and German-Spanish. The research question investigated in this study was the following: does typological similarity have an impact on the potential effects of gender congruency and cognate status in non-native language production? In this, we employed an overt picture-naming task where participants named pictures in the non-native language Spanish using determiner + noun constructions, e.g., *la llave* [the key]. This was done to ensure the processing of grammatical gender since the correct determiner had to be produced alongside the noun (Schiller & Caramazza, 2003). During this task, we measured naming accuracy, naming latencies and P300 amplitudes.

6.2.1 Hypotheses

Hypotheses for behavioural data

We predicted higher accuracy and faster naming latencies for congruent and cognate items compared to incongruent and non-cognate items as an index of CLI. Next, and in line with the CRM

(Stocco et al., 2014), we predicted a non-native production advantage for the Italian-Spanish group compared to the German-Spanish group, reflected in overall higher accuracy and naming latencies. Finally, we predicted CLI effects to vary as a function of typological similarity, i.e., smaller CLI effects for the Italian-Spanish group compared to the German-Spanish group.

Hypotheses for ERP data

We first hypothesised P300 amplitudes to be modulated by gender congruency and cognate status. Specifically, we predicted less positive P300 amplitudes for congruent and cognate items compared to incongruent and non-cognate items. Second, as a reflection of a typological similarity effect and in line with the CRM (Stocco et al., 2014), we expected a production advantage for the typologically similar over the typologically less similar languages. Accordingly, we predicted different neural signatures of CLI between groups in the form of overall smaller P300 amplitudes for the Italian-Spanish group compared to the German-Spanish group.

6.3 Methods

6.3.1 Participants

The Italian-Spanish group included 33 participants (24 females) with M = 27.12 years of age (SD = 4.08), recruited from the Universitat Pompeu Fabra (Barcelona, Spain). The German-Spanish group consisted of 33 participants (27 females) with M = 23.06 years of age (SD = 2.47) recruited from the University of Konstanz (Germany). To avoid confounding effects by proficiency, we restricted our participant selection to late learners with intermediate proficiency levels in the B1/B2 range according to the Common European Framework of Reference for Languages, *CEFR* (Council of Europe, 2001). Note that the majority of participants were recruited from Spanish language courses specifically aimed at B1/B2 proficiency levels. Further eligibility criteria for this study were the following: right-handedness, no additional language learnt before the age of five, AoA of Spanish from fourteen years onwards, no language, reading, vision or hearing impairments and no psychological or neurological issues at the time of testing. Finally, the age limit was between 18 and 35 years. Given that the Italian-Spanish speakers were tested in Spain and the German-Spanish speakers were tested in their native language environment, we balanced any potential differences in immersion by imposing additional inclusion criteria for the Italian-Spanish group. These extra criteria were to only accept those individuals who had started learning Spanish shortly before or upon their arrival to Barcelona and those who had been living in a Spanish-speaking country for less than one year that conformed to B1/B2 proficiency levels.

Linguistic profile

Participants' linguistic profile including self-reported proficiency and experience with Spanish from the LEAP-Q is summarised in Table 6.3.1. The Italian-Spanish group spent on average 0.46 years (SD = 0.343) in a Spanish-speaking country. This compares to an average of 0.96 years (SD = 0.690) for the German-Spanish group. In the Italian-Spanish group, thirteen participants reported Spanish as their strongest language after Italian, fourteen participants as their second, five participants as their third, and one participant as their fourth strongest. Of the German-Spanish group, four participants stated that Spanish was the strongest language after German, twenty-six participants as their second, and three as their third strongest. The proficiency measures were rated on a ten-point scale, ten corresponding to being maximally proficient. The linguistic profiles for both groups were therefore highly comparable. See Appendix 6.A for an overview of the languages acquired by the Italian-Spanish speakers, and Appendix 6.B for the German-Spanish speakers.

Table 6.3.1: *Linguistic profile of participants' Spanish proficiency and experience from the LEAP-Q for the Italian-Spanish group (N = 33) and the German-Spanish group (N = 33). The self-reported proficiency measures are highlighted in bold.*

Native language	**Italian**	**German**
Mean AoA (years)	23.93 (*SD* = 5.07)	16.29 (*SD* = 2.39)
Mean fluency age (years)	24.88 (*SD* = 4.48)	18.53 (*SD* = 2.29)
Mean reading onset age (years)	24.36 (*SD* = 4.91)	17.27 (*SD* = 3.03)
Mean reading fluency age (years)	24.24 (*SD* = 4.82)	18.42 (*SD* = 2.62)
Exposure (%)	40 (*SD* = 18.37)	10 (*SD* = 9.48)
Speaking proficiency	6.09 (*SD* = 1.76)	6.76 (*SD* = 1.00)
Comprehension proficiency	7.26 (*SD* = 1.67)	7.34 (*SD* = 0.92)
Reading proficiency	7.36 (*SD* = 1.48)	7.18 (*SD* = 1.07)

6.3.2 Materials and design

Prior to the experimental session, participants completed the LEAP-Q background questionnaire which provides information regarding language proficiency and language experience from multilinguals (Marian et al., 2007). During the experimental session, participants were first presented with the LexTALE-Esp[2] (Izura et al., 2014), a lexical decision task to measure vocabulary size in Spanish, followed by the picture-naming task. Both these tasks were programmed in E-prime2 (Schneider et al., 2002).

Stimuli

Picture-naming task. We followed an identical stimulus selection procedure for both groups. We selected the picture stimuli from the MultiPic database (Duñabeitia et al., 2018). We chose

[2] For our study, we transformed the LexTALE-Esp version by Izura et al. (2014) into an E-prime equivalent using the same instructions and stimuli words.

those stimuli with the largest proportion of valid and correct responses during the validation phase from the database. We selected 96 stimuli pictures for each group. There were 24 stimuli pictures for each one of the four conditions: congruent cognates, congruent non-cognates, incongruent cognates, and incongruent non-cognates, see Table 6.3.2 and Table 6.3.3 for examples for each group. Identical cognates (e.g., *das Taxi - el taxi* for German and *il taxi - el taxi* for Italian [the taxi]) were not included in our sample; neither were items with biological gender (e.g., *il judice - el juez* [the judge], *der Sänger - el cantante* [the singer]), English loanwords (e.g., *el boomerang* [the boomerang]), or gender-ambiguous nouns (*la mar/el mar* [the sea]). The classification of cognates as such was based on semantic, orthographic and phonological overlap. To increase the validity of our study, we modelled the distribution of terminal phonemes of the Spanish stimuli after work by Clegg (2011), for example, approximately 30% of the stimuli nouns ended in [*a*] and 10% of nouns in [*e*]. Nevertheless, terminal phoneme was included as a covariate in the statistical analyses (see section 6.4.2).

Table 6.3.2: *Example noun phrase stimuli for each condition for the Italian-Spanish group.*

Condition	**Noun phrase**	**Italian translation**	**English translation**
congruent/ cognate	la$_F$ llave$_F$	la$_F$ chiave$_F$	*the key*
congruent/ non-cognate	la$_F$ fresa$_F$	la$_F$ fragola$_F$	*the strawberry*
incongruent/ cognate	el$_M$ bolso$_M$	la$_F$ borsa$_F$	*the handbag*
incongruent/ non-cognate	el$_M$ caracol$_M$	la$_F$ lumaca$_F$	*the slug*

Table 6.3.3: *Example noun phrase stimuli for each condition for the German-Spanish group.*

Condition	**Noun phrase**	**German translation**	**English translation**
congruent/ cognate	la$_F$ jirafa$_F$	die$_F$ Giraffe$_F$	*the giraffe*
congruent/ non-cognate	la$_F$ pera$_F$	die$_F$ Birne$_F$	*the pear*
incongruent/ cognate	el$_M$ melón$_M$	die$_F$ Melone$_F$	*the melon*
incongruent/ non-cognate	el$_M$ tenedor$_M$	die$_F$ Gabel$_F$	*the fork*

EEG recordings

Italian-Spanish group. EEG data were collected via 32 active electrodes using the BrainVision Recorder software (Version 1.10) by BrainProducts using a standard 10/20 montage with a 500 Hz sampling rate. We recorded the vertical electrooculogram (VEOG) from an additional facial electrode placed underneath the participant's left eye (FT9), and the horizontal electrooculogram (HEOG) from one electrode at the outer canthus of the left eye (FT10). The original reference electrode was FCz. The ground electrode was placed on the right cheek of the participant. Electrodes were configured via the BrainVision Recorder software to ensure optimal conductivity. Impedances were kept below 10 kΩ.

German-Spanish group. EEG data were collected from 32 passive electrode locations via the BrainVision Recorder software (Version 1.23.0001) at a sampling rate of 500 Hz. We used the standard 10/20 montage with an EasyCap electrode cap. The HEOG was measured from two electrodes at the outer canthus of the left and right eye. The VEOG was recorded using an electrode underneath the left eye. The ground electrode was placed on the right cheek of the participant. Electrodes were initially referenced to the Cz electrode. Impedances of the electrodes were checked and configured using actiCAP ControlSoftware (Version 1.2.5.3). We kept impedances below 5 kΩ. for the reference and ground electrode. For the remaining channels, impedances were below 10 kΩ.

6.3.3 Procedure

Complying with the ethics code for linguistic research in the Faculty of Humanities at Leiden University, participants signed an informed consent form prior and after participation. Further, they were given an information sheet prior to the experiment and a written and oral debrief upon termination of the experiment in their respective L1. Participants received a monetary reimbursement for their participation. The procedure for both the Italian-Spanish and the German-Spanish group was identical, with the difference that for the former group oral instructions were provided in Italian, and for the latter in German by a native speaker.

LexTALE-Esp

For the LexTALE-Esp, instructions were provided in black font on a white screen. Next, a fixation cross was displayed for 1,000 ms. Then, a letter string corresponding to either a Spanish word or a pseudoword was displayed in the centre of the screen. Participants were asked to make a lexical decision via a button press about whether or not the string was a Spanish word. The letter string remained on the screen until the participant's response. The original stimuli from Izura et al. (2014) consisted of 60 words and 30 pseudowords. For both the Italian-Spanish and the German-Spanish group, three stimuli were eliminated from the stimuli list before the experiment due to overlap with the picture stimuli. Therefore, the total number of trials was 87 for both groups. Each letter string was only shown once, and trial order was randomised for each participant. Offline, we computed LexTALE-Esp vocabulary size scores by subtracting the percentage of incorrectly identified pseudowords from the correctly identified words (Izura et al., 2014). LexTALE-Esp scores were subsequently included as a covariate in our statistical analyses (see section 6.4.2).

Picture-naming task

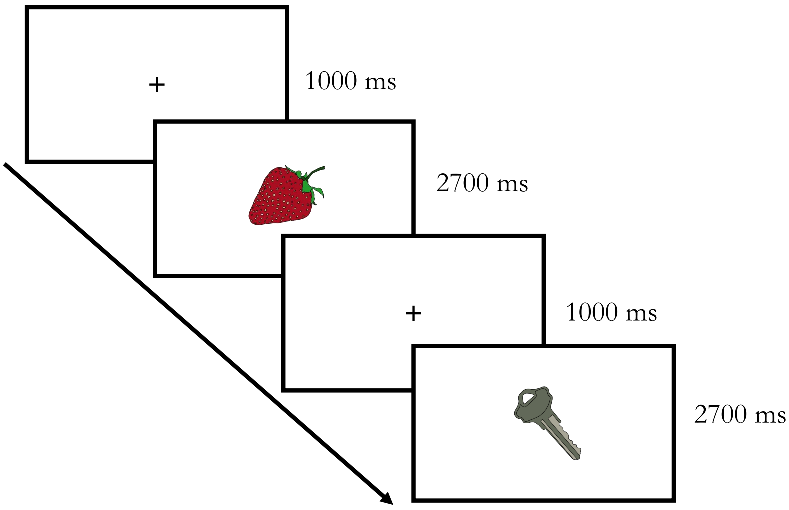
For the picture-naming task, we manipulated gender congruency and cognate status in a 2 x 2 fully factorial within-subjects design. Half of the stimuli were congruent items, whereas the other half were incongruent items. Half of the congruent and incongruent items were cognates, respectively, whereas the other half were noncognates. We divided the task into a familiarisation phase and an experimental phase, during which we recorded the EEG signal. The familiarisation phase consisted of three rounds. In each round, participants were shown each picture and had to overtly produce the corresponding noun together with the correct definite determiner (e.g., *la llave* [the key]). If either the noun or the determiner, or both, were incorrect, the experimenter provided oral feedback to the participant to ensure the participants' familiarity with each picture. The total number of trials in the familiarisation phase was 288. In the experimental phase, a trial was initiated by a black fixation cross on a white screen, which was displayed for 1,000 ms. Next, a picture appeared on the screen for 2,700 ms (Figure 6.3.1). Participants were instructed to overtly name each picture on the screen as fast and accurately as possible. They were explicitly encouraged to minimise all movements during the experiment. Each picture was shown once in a unique trial order during the experimental phase, resulting in a total of 96 trials. There were two self-paced breaks to restore participants' engagement with the task.

6.4 Results

6.4.1 Behavioural data exclusion

Data for one participant of the German-Spanish group were lost due to a recording failure. For naming latencies, we included only correct trials in the analysis. Moreover, the behavioural analyses consisted of the identical data sets as the EEG analysis (see section 4.4 for details on EEG data exclusion). Therefore, we included a total of 28 participants from the Italian-Spanish group and 30

Figure 6.3.1: *Trial sequence for the picture-naming task.*



participants from the German-Spanish group in the behavioural analyses (n = 58).

6.4.2 Behavioural data analysis

We calculated naming accuracy and naming latencies in Praat (Broersma & Weenink, 2019). We pooled the behavioural data for both groups to probe modulatory effects of CLI and typological similarity on *naming accuracy* and *naming latencies* across the Italian-Spanish speakers and the German-Spanish speakers. We employed a single-trial linear mixed effects modelling (LMM) approach using the *lme4* package (Bates et al., 2020) in RStudio (Team, 2020). To model *naming accuracy*, we used a generalised linear mixed effect model (GLMM) with a binomial distribution using the *glmer()* function. For our positively skewed *naming latencies*, we used a GLMM with a gamma distribution and the identify link function with the *glmer()* function (Lo & Andrews, 2015). See Appendix 6.C for details about the included interaction effects and the model fitting procedure, which is also described in Von Grebmer Zu Wolfsthurn et al. (2021b). Our fixed effects structure consisted of the following: *typological similarity* (typologically similar vs. typologically dissimilar), *gender congruency* (congruent vs. incongruent) and *cognate status* (cognate vs. non-cognate). Moreover, we controlled for the covariates *LexTALE-Esp score, familiarisation phase performance, target noun gender, word length, order of acquisition of Spanish* and *terminal phoneme* of the target word in both analyses[3]. Finally, our random effects structure included *subject* and *item* (i.e., the individual picture) as well as random slopes for our main manipulations.

6.4.3 Behavioural data results

We first computed descriptive statistics for the remaining participants of each group. See Table 6.4.1 for descriptives for *naming accuracy*, reported as percentages, and for *naming latencies* reported in ms. Mean LexTALE-Esp scores were $M = 25.92$ ($SD = 13.69$, range between -7.37 and 49.30) for the Italian-Spanish group, and $M = 18.45$ ($SD = 20.52$, range between -23.16 and 60.18) for the German-Spanish group. Scores above 60 were previously linked to C1/C2 levels, thereby confirming the B1/B2 range of our participants (Lemhöfer & Broersma, 2012). A two-sample t-test yielded no statistical difference in LexTALE-Esp scores between the two groups with $t(50.83) = 1.64$, *95% CI*[1.68, 16.60], $p = 0.101$. Nevertheless, we included LexTALE-Esp scores as a covariate in our analysis. Further, we plotted naming accuracy for both groups in Figure 6.4.1 and naming latencies in Figure 6.4.2.

Naming accuracy. The model of best fit included both *gender congruency* and *cognate status* as main effects. Participants were more accurate for congruent items compared to incongruent items ($\beta = 0.728$, *95% CI*[0.542, 0.978], $z = -2.11$, $p = 0.035$). Further, participants were more accurate for cognates compared to non-cognates ($\beta = 0.696$, *95% CI*[0.519, 0.934], $z = -2.41$, $p =$

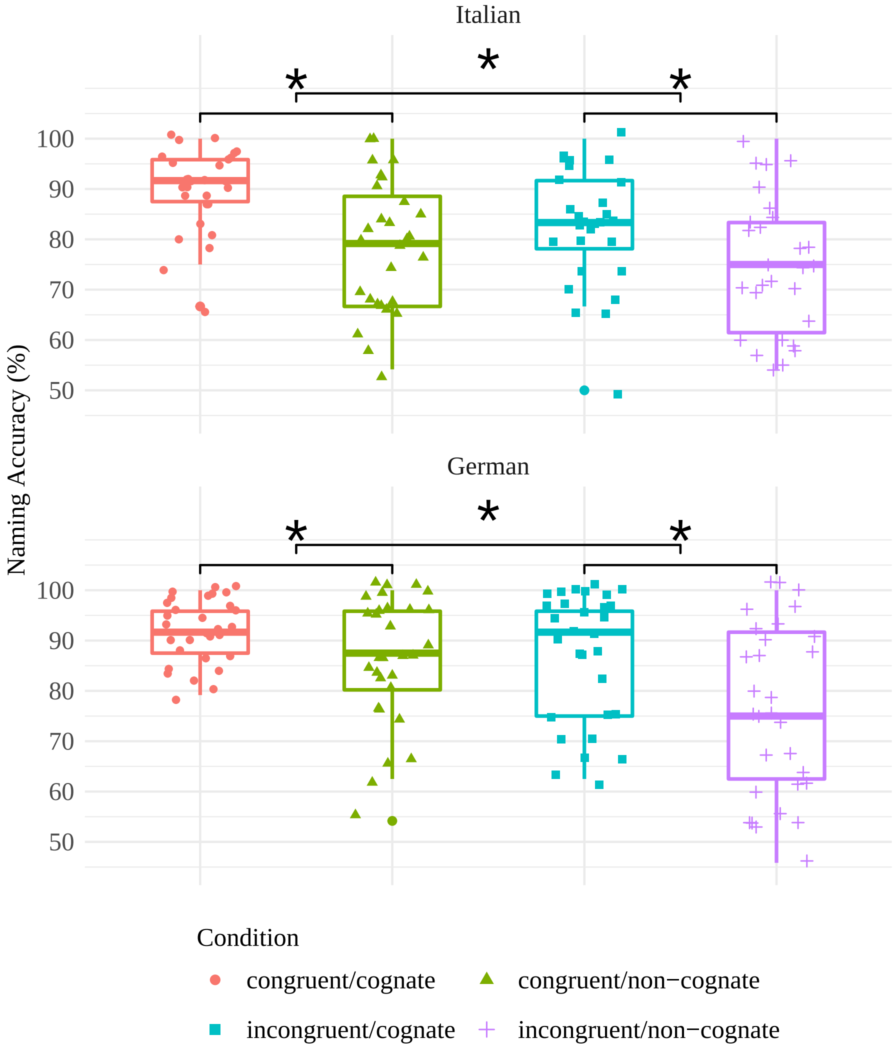
[3] Note that in Von Grebmer Zu Wolfsthurn et al. (2021b) acquisition of French was also considered as a potential covariate; however, no significant effect on our outcome variables was found. We therefore did not include acquisition of French in the current analyses.

Table 6.4.1: *Mean naming accuracy (%) and mean naming latencies (ms) for each condition for the Italian-Spanish group (n = 28) and the German-Spanish group (n = 30).*

Native language	**Condition**	**Mean (%)**	**SD**	**Mean (ms)**	**SD**
Italian	congruent/ cognate	89.88	30.18	911.12	253.91
	congruent/ non-cognate	78.57	41.06	1011.90	297.67
	incongruent/ cognate	82.59	37.95	1027.61	305.52
	incongruent/ non-cognate	75.00	43.33	1068.59	281.19
German	congruent/ cognate	92.08	27.02	891.46	236.85
	congruent/ non-cognate	86.39	34.31	932.52	293.93
	incongruent/ cognate	87.22	33.41	971.27	312.85
	incongruent/ non-cognate	75.83	42.84	977.70	303.22

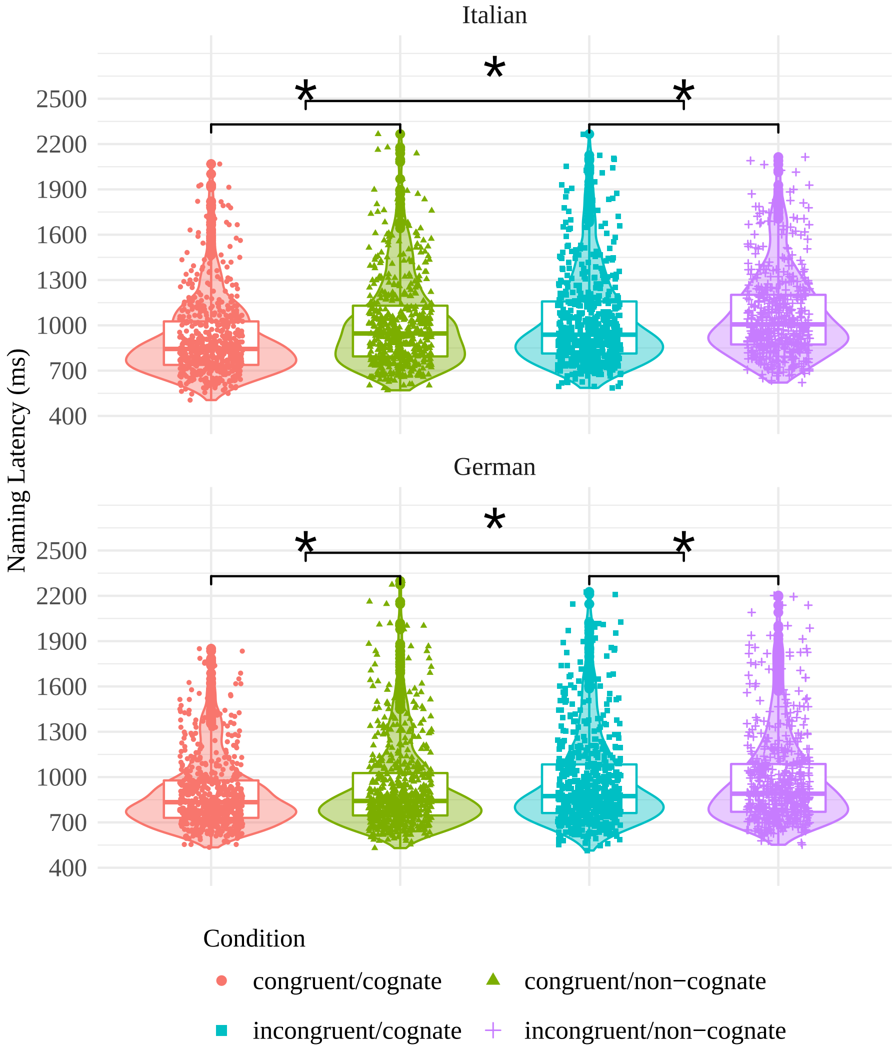
0.016). The model also included *familiarisation phase performance* as a covariate, and *subject* and *item* as random effects. *Typological similarity* did not significantly improve the model fit with χ^2(1, n = 58) = 0.235, p = 0.125 and was therefore not included in the final model. The best-fitting model was the following: naming accuracy ~ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (1|subject) + (1|item). See Appendix 6.D and Figure 6.4.1.

Figure 6.4.1: *By-condition naming accuracy (%) for each group. The significance brackets reflect the statistical difference in accuracy between congruent and incongruent items and between cognate and non-cognate items, with higher accuracy rates for congruent and cognate items, as well as for the German-Spanish group (n = 30) compared to the Italian-Spanish group (n = 28).*



Naming latencies. For naming latencies, the model of best fit included main effects for *gender congruency* and *cognate status*. Participants were faster at naming congruent compared to incongruent items, and cognate compared to non-cognate items with β = 0.064, *95% CI*[0.019, 0.108], t = 2.82, p = 0.005 and β = 0.046, *95% CI*[0.005, 0.087], t = 2.18, p = 0.029, respectively. We included *familiarisation phase performance* as a covariate, however, we found no evidence that *typological similarity* contributed to a better model fit. Similarly, the other covariates did not improve the model fit or led to non-convergence and were therefore excluded from the model. Furthermore, we included a by-*subject* random slope for the correlated effects of *gender congruency* and *cognate status* as well as *item* as random effect. The final model was therefore the following: naming latency ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|subject) + (1|item), see Appendix 6.E and Figure 6.4.2 for details. Taken together, we found no behavioural evidence that *typological similarity* modulated *naming accuracy* or *naming latencies* across the two groups.

Figure 6.4.2: *By-condition naming latencies (ms) for each the Italian-Spanish group (n = 28) and the German-Spanish group (n = 30). The significance brackets reflect the statistical difference in accuracy between congruent and incongruent items and between cognate and non-cognate items, with faster naming latencies for congruent and cognate items.*



6.4.4 EEG data exclusion

Upon completion of pre-processing of our EEG data, we determined a set of identical inclusion criteria for both groups: first, we only modelled the EEG signal for correct trials, i.e., trials where participants provided the correct noun phrase in the experimental phase (Christoffels et al., 2007). Secondly, we only included trials which were not contaminated by artefacts. Finally, we did not include participants with heavily contaminated datasets, i.e., > 40% of trials lost due to artefacts. The resulting threshold for inclusion was a remainder of at least 60% of trials at the end of pre-processing and the application of the inclusion criteria. Following these criteria, we excluded five EEG datasets from the Italian-Spanish group and two EEG datasets from the German-Spanish group. The corresponding behavioural data were also excluded from the behavioural analysis. This amounted to 28 Italian-Spanish datasets and 30 German-Spanish datasets for further statistical analyses (n = 58).

6.4.5 EEG data pre-processing

Prior to any statistical analyses, we performed a vigorous pre-processing procedure to separate the task-related EEG signal from articulatory artefacts frequently found in production paradigms (Ganushchak, Christoffels & Schiller, 2011). We followed a largely identical procedure for both groups. Note that as stated above, the recording parameters differed between the two groups. The EEG data were analysed in BrainVision Analyser V2.2. First, we re-referenced all of our data channels to the average of the mastoid channels, TP9 and TP10. For the Italian-Spanish group, the original reference channel FCz was reused as a data channel, adding to a new total of 29 data channels. FT9 and FT10 were used as VEOG and HEOG, respectively. For the German-Spanish group, we reused the reference Cz channel as a data channel, amounting to a total of 31 data channels. For this group, we performed linear derivation on the two HEOG channels to form a single HEOG channel. For both groups, we subsequently applied a high-pass filter of 0.1 Hz, and a low-pass filter of 30 Hz. We interpolated channels where appropriate. We then performed residual drift detection on our HEOG and VEOG channels in order to improve the precision of the subsequent blink correction using ocular ICA. Next, we performed artefact rejection on all data channels. After increasing the signal-to-noise ratio by means of pre-processing, we segmented our data at 200 ms prior and 1,200 ms after picture onset. Segments which included artefacts were excluded in this process. After segmentation, we applied a baseline correction using the 200 ms interval prior to picture onset. Finally, we exported all voltage amplitudes for all segments, channels and participants in order to perform single-trial LMM in RStudio (R Core Team, 2020). This method was recently introduced as powerful alternative to the more traditional grand-averaging of EEG data because it captures both by-subject and by-item individual variance and correlations between individual data points while preserving statistical power (Frömer et al., 2018).

6.4.6 EEG data analysis

Next, we tentatively explored our EEG data via a permutation analysis to visualise potential effects of condition on voltage amplitudes. First, we divided our 29 data electrodes for the Italian-Spanish group (Fp1, Fp2, Fz, F3, F4, F7, F8, FCz, FC1, FC2, FC5, FC6, Cz, C3, C4, T7, T8, CP1, CP2, CP5, CP6, Pz, P3, P7, P4, P8, Oz, O1 and O2) and the 31 electrodes for our German-Spanish group (Fp1, Fp2, AFz, Fz, F3, F4, F7, F8, FCz, FC3, FC4, FT7, FT8, Cz, CPz, CP3, CP4, C3, C4, T7, T8, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1 and O2) into nine topographic areas: anterior, central and posterior, in turn divided into left, midline and right sections. Next, we used the *permu.test()* function from the *permutes* package (Voeten, 2019) for a permutation analysis on each group. See Figure 6.4.3 for the outcome of the permutation test for the Italian-Spanish group, and Figure 6.4.4 for the German-Spanish group. In both figures, darker colours correspond to higher F-values indicating potential differences in voltage between conditions, whereas lighter colours correspond to lower F-values following an F-distribution under the null hypothesis that there are no differences by condition (Maris & Oostenveld, 2007; Voeten, 2019). For the Italian-Spanish group, the output from the permutation analysis showed that a potential region of interest was clustered in centro-parietal regions around the following eight electrodes: *Pz, P3, P4, P7, P8, Oz, O1* and *O2*. Further, the permutation analysis suggested a time window of interest between 350 ms and 600 ms post-stimulus onset. On the other hand, the permutation test for the German-Spanish group showed potentially significant effects in centro-parietal regions at the following thirteen electrodes: *CPz, CP3, CP4, TP7, TP8, Pz, P3, P4, P7, P8, Oz, O1* and *O2* in a similar time window, around 350 ms to 600 ms post-stimulus onset. Due to increasing articulatory artefacts in proximity to the articulatory onset, we deliberately set the upper time window threshold to 600 ms post-stimulus onset to avoid signal contamination by motor articulation.

Figure 6.4.3: *Permutation test outcome for the Italian-Spanish group (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.*

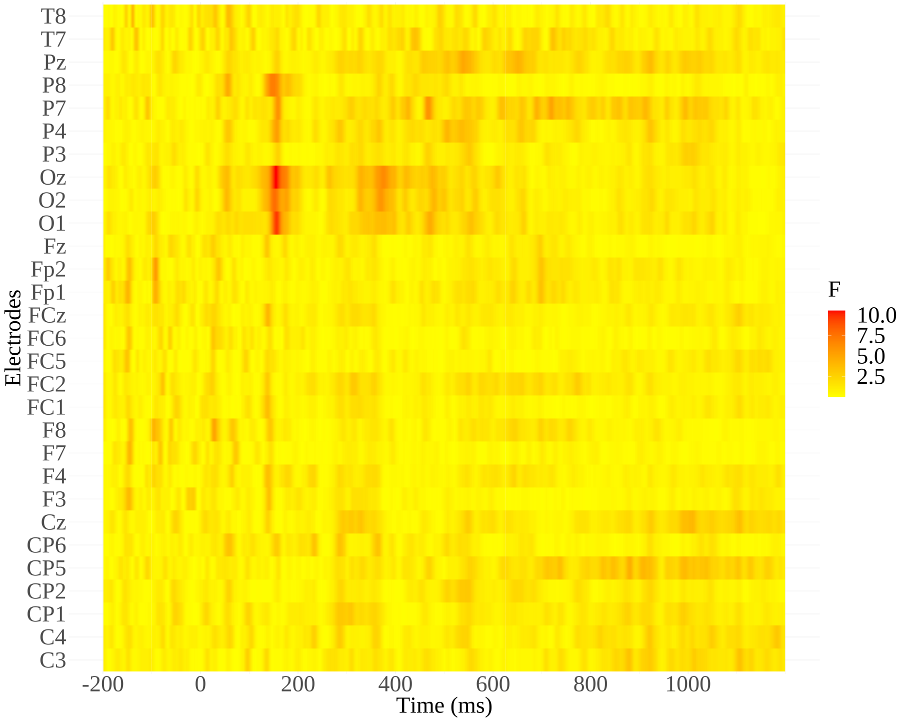
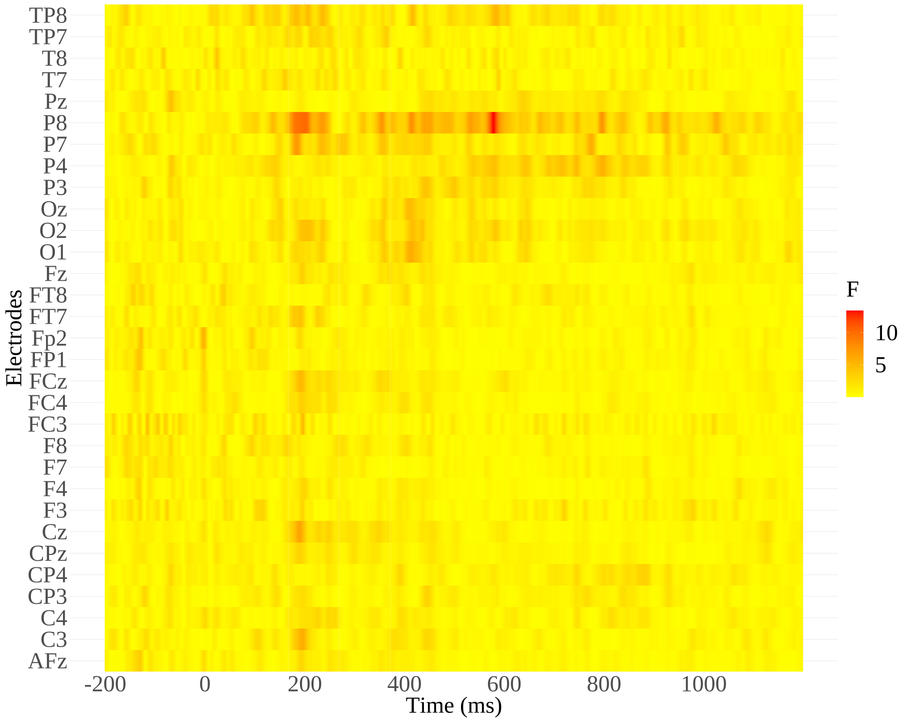


Figure 6.4.4: *Permutation test outcome for the German-Spanish group (n = 30). Larger F-values are shown in darker colours and denote an increased likelihood for a statistically relevant effect of our manipulations on voltage amplitudes.*



As a next step, we pooled our EEG data and proceeded to the statistical analysis. On the basis of the previous literature and the outcome of the permutation analysis, we selected centro-parietal regions as our region of interest, and 350 ms to 600 ms as our time window of interest. We only selected the eight electrodes which were present in the montage of both groups: *Pz, P3, P4, P7, P8, Oz, O1* and *O2*. See Appendix 6.F for a visualisation of the region of interest for the group analysis. Mirroring the behavioural data analysis, we modelled *voltage amplitudes* as a function of *gender congruency*, *cognate status* and *typological similarity* in our fixed effects structure using the *lme4* package (Bates et al., 2020). Moreover, we carefully controlled for potential confounding factors such as *hemisphere, LexTALE-Esp score, familiarisation phase per-*

formance, target noun gender, word length, order of acquisition and *terminal phoneme*. *Subject* and *item* were defined as random effects for the random effects structure and random slopes for our main manipulations were included. Lastly, the final model was re-fitted using the restricted maximum likelihood criteria (REML) for unbiased estimates (Mardia et al., 1999).

6.4.7 EEG data results

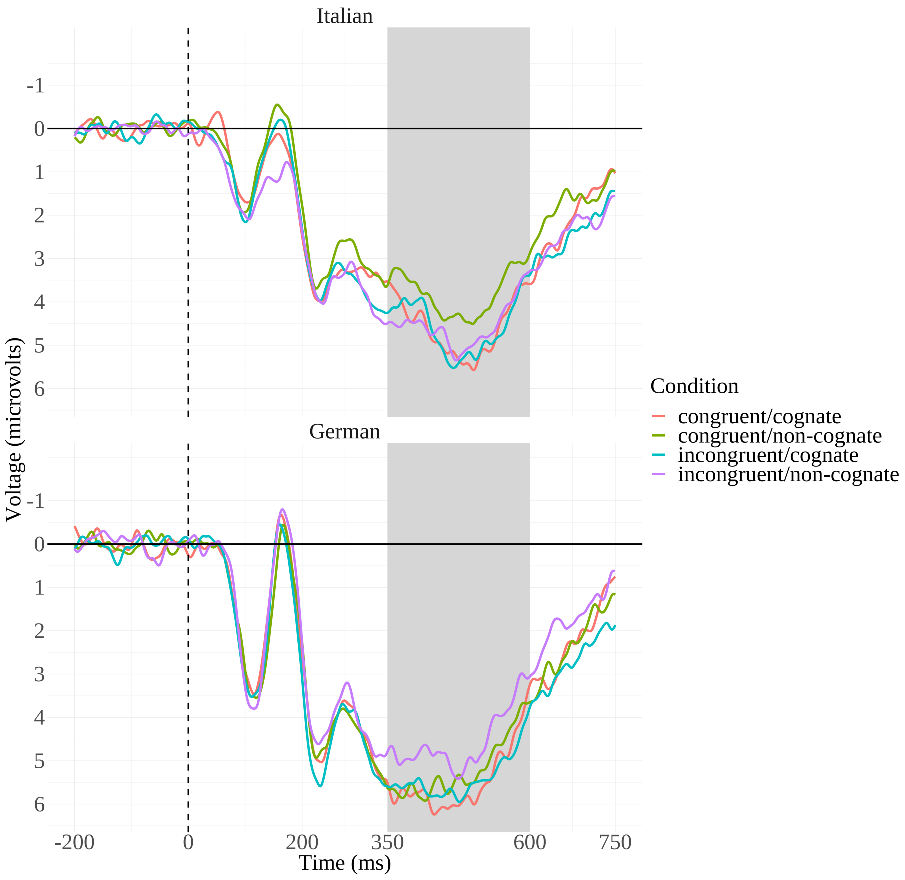
Mean voltage amplitudes by condition for each group are summarised in Table 6.4.2. Descriptively speaking, the Italian-Spanish group yielded overall lower voltage amplitudes compared to the German-Spanish group in our time window and region of interest. Visual inspection of voltage amplitudes in Figure 6.4.5 showed a positive oscillation followed immediately by a negative oscillation shortly after stimulus onset, which reflected the classical P1/N2 complex linked to early visual processing (P. Chen et al., 2017). In line with the outcomes of the permutation test, both groups then showed a positive-going waveform across centro-parietal regions between 350 ms and 600 ms after stimulus onset (indicated by grey shading in Figure 6.4.5). This particular waveform in our time window of interest, together with the topographic distribution of the channels, indicated a P300 component in both groups. Voltage amplitudes peaked around 500 ms post-stimulus onset. Finally, they visibly dropped back to baseline and became increasingly noisier closer to the articulatory onset around 700 ms.

Table 6.4.2: *Mean voltage amplitudes (μV) by condition for centro-parietal regions for the time window of interest (350 ms to 600 ms).*

Native language	**Condition**	**Mean (μV)**	**SD**
Italian	congruent/cognate	4.58	8.85
	congruent/non-cognate	3.80	9.07
	incongruent/cognate	4.57	8.23
	incongruent/non-cognate	4.54	8.54
German	congruent/cognate	5.48	9.59
	congruent/non-cognate	5.19	9.13
	incongruent/cognate	5.39	9.53
	incongruent/non-cognate	4.55	9.36

For voltage amplitudes, the model of best fit included *gender congruency* and *cognate status* as main effects, as well as *hemisphere* and *familiarisation phase performance* as covariates. *Subject* and *item* were included as random effects, as well as correlated by-*subject* random slopes for *gender congruency* and *cognate status.* Therefore, the final model was the following: voltage amplitudes $\sim$ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|subject) + (1|item). *Voltage amplitudes* were significantly higher for cognates compared to non-cognates with β = -0.477, *95% CI*[-0.934, -0.021], t = -2.05, p = 0.040. In contrast, there was no significant effect of *gender congruency* on voltage amplitudes with β = 0.025, *95% CI*[-0.479, 0.529], t = 0.097, p = 0.922 for congruent compared to incongruent items. Despite a descriptive trend, *typological similarity* did not significantly improve the model fit, with χ^2(1, n = 58) = 1.31, p = 0.252 when comparing the model with and without typological similarity in the fixed effects structure, see Appendix 6.G and Figure 6.4.5 for further details. Therefore, mirroring the behavioural results, *typological similarity* did not appear to influence electrophysiological measures of non-native production.

Figure 6.4.5: *By-condition voltage amplitudes (μV) for centro-parietal regions for each group. The time window of interest (350 ms to 600 ms) is highlighted in light grey. Note that positive voltage amplitudes are plotted downwards, and negative voltage amplitudes are plotted upwards.*



6.5 Discussion

In this study, we investigated the effect of typological similarity on CLI in non-native production in Italian learners of Spanish (typologically similar group) and German learners of Spanish (typologically dissimilar group) with intermediate proficiency levels (B1/B2). More specifically, we examined a processing advantage for the typologically similar group within the framework of the CRM (Stocco et al., 2014) and explored how *gender congruency*, *cognate status* and *typological similarity* modulated overt picture-naming in intermediate learners of Spanish, e.g., when producing *la llave* [the key]. We modelled *naming accuracy* and *naming latencies*, as well as P300 *voltage amplitudes*. Outcomes of this study have crucial implications for non-native language processing, as well as the functional organisation of languages with respect to typological similarity.

Behaviourally, we expected more accurate and faster processing of congruent and cognate items compared to incongruent and non-cognate items. In addition, we predicted the typologically similar Italian-Spanish group to show an overall production advantage compared to the typologically dissimilar German-Spanish group in terms of naming accuracy and naming latencies. Finally, we predicted CLI effects to vary between our two groups. Our results partially supported our hypotheses. Participants were faster and more accurate at naming congruent compared to incongruent items and at naming cognates compared to non-cognates, displaying both the classical gender congruency effect (Paolieri et al., 2019) and the cognate facilitation effect (Lemhöfer et al., 2008). However, we found no main effect of typological similarity nor an interaction effect, indicating that the mitigation of CLI effects was equally successful in both groups. This is in line with previous studies (Costa, Heij & Navarrete, 2006; Costa et al., 2003).

From a theoretical point of view, our behavioural results showing more accurate and faster processing of congruent and cognate items indicated first that we found evidence for a sensitivity of late learners to both gender congruency and cognate status. Secondly, these results suggest that both groups experience comparable levels of CLI, as reflected by similar behavioural performances. One possible interpretation is that at intermediate proficiency levels, the facilitatory effects (when processing congruent items and cognates) and the hampering effects (when processing incongruent items and non-cognates) are balanced across the two groups. At this stage of learning, the overall similarity between the respective native languages and Spanish showed little influence on multilingual language production. Therefore, our behavioural results do not support the predictions of the CRM by Stocco et al. (2014). Based on these results, we postulate that gender congruency and cognate status are the main modulating factors of non-native production in this study and that typological similarity plays a limited role at intermediate proficiency levels. However, to obtain a more comprehensive interpretation of these data, we integrated the behavioural findings with the EEG findings.

At the neural level, we originally predicted a modulation of the P300 component as a function of *gender congruency* and *cognate status*. Moreover, we expected smaller P300 amplitudes for the Italian-Spanish group compared to the German-Spanish group to reflect an effect of *typological similarity*. The first critical finding was that the EEG signal in our time window and region of interest was consistent with a P300 component in both the Italian-Spanish and the German-Spanish group, although this P300 showed a delayed voltage peak compared to previous research (Polich, 2007). These results mirror Von Grebmer Zu Wolfsthurn et al. (2021b). As previously discussed, the P300 has been typically linked to conflict monitoring, cognitive control and interference, and more recently to attentional resources and working memory (González Alonso et al., 2020; Polich, 2007). On the basis of our findings, we argue that the P300 may be a critical index for cognitive control in the non-native production process because to succeed at this task, speakers need to mitigate language co-activation and CLI effects. This implies a strong element of cognitive control in non-native production for speakers at intermediate proficiency levels.

The second critical finding in line with our hypotheses was the modulation of the P300 voltage amplitudes by CLI of cognates: they elicited larger P300 voltage amplitudes compared to non-cognates. The notion of larger voltage amplitudes for cognates was found in previous work (Christoffels et al., 2007; Strijkers et al., 2010), but studies also reported the opposite pattern (Peeters et al., 2013).

Here, we argue that the difference in voltage amplitudes for cognates compared to non-cognates is reflective of the mitigatory processes to manage CLI, and that the P300 modulation may be a critical marker of the processes underlying CLI. The third important finding was that we found no evidence that gender congruency significantly modulated P300 voltage amplitudes. Therefore, our data suggest that cognate status is the more salient modulating feature during non-native production at the neural level.

Finally, we found no evidence for distinct neural signatures across the Italian-Spanish and the German-Spanish group: we did not find an interaction effect between typological similarity and the two CLI effects we tested that could indicate different CLI effects across groups. Statistically speaking, voltage amplitudes were comparable for both groups. Therefore, our ERP results suggest a limited effect of typological similarity on CLI effects and non-native production as a whole (Costa, Heij & Navarrete, 2006; Costa et al., 2003). This finding speaks directly to the theoretical framework regarding the directionality of the typological similarity effect in that it does not support an advantage of typologically similar languages (Stocco et al., 2014). These ERP findings are contrary to what we hypothesised, but are in line with the behavioural results we obtained. The question here is whether there could be another modulating factor at play that is potentially more powerful than typological similarity effects.

As discussed in the introduction, Costa et al. (2003) did not find evidence for an effect of typological similarity in groups of highly proficient Spanish-Catalan, Catalan-Spanish, Italian-French and Croatian-Italian speakers. This is in line with our findings. Similar results were found in a later study by Costa, Heij and Navarrete (2006) with highly proficient speakers, which also did not show evidence for typological similarity effect in a bilingual picture-naming task in Spanish-Catalan (typologically similar) and Spanish-Basque (typologically dissimilar) speakers. One possible interpretation of our findings is embedded within the context of proficiency, which is a key feature in mitigating CLI and language control (D. W. Green, 1998). More specifically, the IC model (D. W. Green, 1998) proposes similar language control mechanisms for the L1 and the non-native language when proficiency levels are highly similar. The speakers in our study were intermediate speakers of Spanish, implying that typological similarity effects on non-native production were limited when the difference in proficiency between the L1 and the non-native language was modest. Nevertheless, this does not exclude the possibility that typological similarity effects could increase with decreased non-native proficiency.

Taken together, our results suggest that CLI is a driving factor of the mechanisms underlying non-native production at intermediate proficiency levels in our study, and not typological similarity. However, the exact contribution of typological similarity may change dynamically as a function of other key factors in multilingual language processing, such as non-native proficiency. Given the scarcity of research, more empirical studies are needed to delve deeper into this issue, see section 6.5.1.

6.5.1 Conclusions

In this study, we asked the question whether or not typologically similar languages bear a processing advantage in non-native production. Our data showed CLI effects at the behavioural level in the form of a processing advantage of congruent and cognate items compared to incongruent and non-cognate items. Unexpectedly, we found no evidence for an effect of typological similarity on behavioural measures: the behavioural performance and CLI effects were comparable across the Italian-Spanish and German-Spanish group. For the EEG data, we found a P300 component across both groups. Further, P300 amplitudes were modulated by cognate status. This highlights the crucial involvement of the P300 component in mitigatory CLI processes. In contrast, there was no traceable effect of typological similarity on voltage amplitudes, reflecting highly similar neural signatures and CLI across the two groups. Taken together,these results suggest a limited role of typological similarity on non-native production at intermediate proficiency levels. We argue that this may be linked to the relatively small difference in proficiency between the L1 and the non-native language.

6.5.2 Future directions

Our study does not allow for a direct assessment of a typological similarity effect at lower non-native proficiency levels in driving CLI. Future research could incorporate varying groups of late learners at different acquisition stages to support our claims. Further, while we carefully recruited our participants to fit the B1/B2 proficiency range and controlled for a number of linguistic variables, future studies should incorporate a more direct measure of overall proficiency to include in the statistical models. In turn, this could be used for more nuanced analysis for participants in the lower vs. higher B1/B2 range. Finally, there is an urgent need for an objective measure of typological similarity to quantify any effects on multilingual language processing and cognition. While the classification of our groups was unambiguous in the current study, this remains a frequent debate in the literature (Van der Slik, 2010).

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 are grateful to the members of the neurolinguistics lab at the University of Konstanz, Carsten Eulitz, Oleksiy Bobrow, and the research assistants Charlotte Englert, Khrystyna Oliynyk, Zeynep Dogan, Fangming Zhang and Anna-Maria Waibel. We also thank Núria Sebastián-Gallés and the CBC team, in particular Ege Ekin Özer, at the Universitat Pompeu Fabra for providing us with the testing facilities. A special thanks goes to Theo Marinis and Maria del Carmen Parafita Couto for their feedback, to Julian Karch for his support in the statistical analyses and to the POLAR lab for their feedback. 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 time.

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/xraz2/?view_only=a708afaf8a684441a95d928829ed6733`

Citation diversity statement

We made a particular effort to include this Citation Diversity Statement in our manuscript. We aim to raise awareness about the systematic underrepresentation of female authors and authors from minority populations in academic publishing (Dworkin et al., 2020). Here, we report the proportion of gender representation in our reference list, wherever this information was available. Our reference list contained 19% woman/woman authors, 41% man/man, 15% woman/man and finally, 17% man/woman authors. In comparison, the equivalents for neuroscience are 6.7% for woman/woman, 58.4% for man/man, 25.5% woman/man, and lastly, 9.4% for man/woman authored references, see Dworkin et al. (2020). In the future, we are convinced that this binary gender classification system will be further improved and that data will become available for additional research areas.

Appendix

6.A Linguistic profile: Italian-Spanish group

Table 6.A.1: *Overview of the native and non-native languages acquired by the Italian-Spanish speakers included in the analysis (n = 28).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Italian	n = 28					**28**
Spanish		n = 2	n = 15	n = 8	n = 3	**28**
English		n = 23	n = 4			**27**
French		n = 3	n = 7	n = 3		**13**
German			n = 1	n = 2		**3**
Portuguese					n = 2	**2**
Catalan					n = 1	**1**
Total	**28**	**28**	**27**	**13**	**6**	

6.B Linguistic profile: German-Spanish group

Table 6.B.1: *Overview of the native and non-native languages acquired by the Italian-Spanish speakers included in the analysis (n = 30).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
German	n = 30					**30**
Spanish			n = 16	n = 12	n = 2	**30**
English		n = 28	n = 2			**30**
French		n = 2	n = 8	n = 5		**15**
Latin			n = 3	n = 1	n = 1	**5**
Russian			n = 1		n = 1	**2**
Swedish				n = 1		**1**
Portuguese					n = 1	**1**
Catalan					n = 1	**1**
Italian					n = 1	**1**
Mandarin					n = 1	**1**
Total	**30**	**30**	**30**	**19**	**8**	

6.C Model fitting procedure

The model-fitting procedure was as follows: we first constructed a theoretically plausible maximal model. This model included an elaborate fixed effects structure which consisted of our fixed effects and any pre-hypothesised interactions, as well as the covariates. For both our behavioural data and the EEG data, the maximal model included an interaction effect of *typological similarity* with *gender congruency* and *cognate status* in order to test the hypotheses of finding potential differential effects across groups. Our random effects structure was specified as maximally as possible with random slopes as well as random intercepts if supported by the data (Barr, 2013). In the case of non-convergence or singular fit, we first simplified our random effects structure. Next, we proceeded to simplify the fixed effects structure and systematically tested for statistical significance of the fixed effects and the covariates in a top-down fashion. By default, GLMMs were fitted using the maximum likelihood (ML) method with the Laplace approximation (Bates et al., 2020). Absolute test-statistic values larger than ±1.96 at $\alpha = 0.05$ were defined as statistically significant (Alday et al., 2017). Model comparisons were performed to assess the contribution of each fixed effect using the *anova()* function. This function is based on the Information Criteria AIC and BIC and the loglikelihood ratio. Fixed effects which did not significantly improve the model fit were excluded from the model selection procedure. Model fit was assessed after each model by examining the model residuals using the *DHARMa* package (Hartig, 2020). Treatment coding was used as our default contrast.

6.D Model parameters: naming accuracy

Table 6.D.1: *Model outcome parameters for naming accuracy (n = 58).*

Formula: naming accuracy ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (1|subject) + (1|item)

Term	**Odds Ratio [95% CI]**	**z-value**	**p-value**
(Intercept)	0.641 [0.416, 0.987]	-2.02	0.043
Gender congruency [incongruent]	0.728 [0.542, 0.978]	-2.11	**0.035**
Cognate status [non-cognate]	0.696 [0.519, 0.934]	-2.41	**0.016**
Familiarisation phase performance [one correct]	6.342 [4.67, 8.61]	11.82	< 0.001
Familiarisation phase performance [two correct]	26.75 [19.50, 36.69]	20.39	< 0.001
Familiarisation phase performance [three correct]	58.02 [41.63, 80.87]	23.97	< 0.001
Random effects			
σ^2	3.29		
τ_{00Item}	0.55		
$\tau_{00Subject}$	0.59		
ICC	0.26		
$N_{Subject}$	58		
N_{Item}	192		
Observations	5,568		
Marginal R^2/ Conditional R^2	0.286/0.469		

6.E Model parameters: naming latencies

Table 6.E.1: *Model outcome parameters for naming latencies (n = 58).*

Formula: naming latency $\sim$ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|subject) + (1|item)

Term	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	1.39 [1.31, 1.46]	34.80	< 0.001
Gender congruency [incongruent]	0.064 [0.019, 0.108]	2.82	**0.005**
Cognate status [non-cognate]	0.046 [0.005, 0.087]	2.18	**0.029**
Familiarisation phase performance [one correct]	-0.182 [-0.239, -0.125]	-6.25	< 0.001
Familiarisation phase performance [two correct]	-0.371 [-0.424, -0.318]	-13.71	< 0.001
Familiarisation phase performance [three correct]	-0.445 [-0.499, -0.392]	-16.39	< 0.001

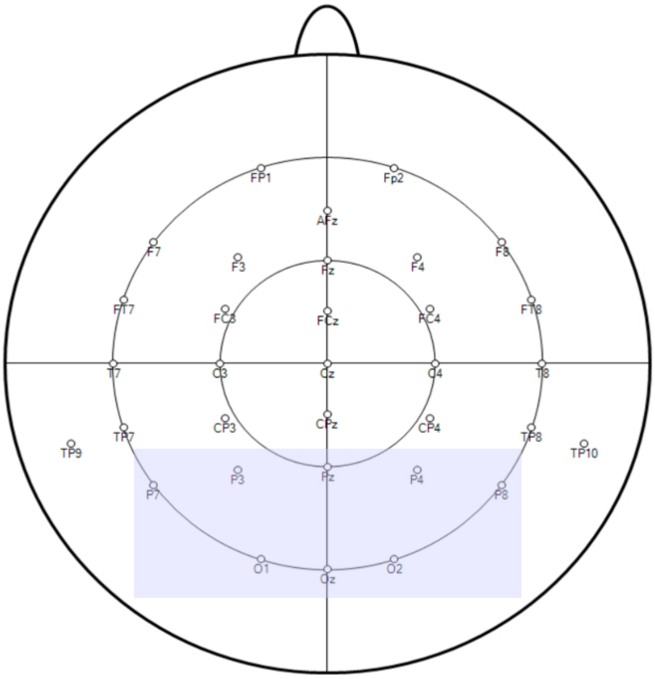
Random effects

σ^2	0.05
$\tau_{00 Item}$	0.00
$\tau_{00 Subject}$	0.00
$\tau_{11 Subject[incongruent]}$	0.00
$\tau_{11 Subject[non-cognate]}$	0.00
$\rho_{01 Subject[incongruent]}$	-0.21
$\rho_{01 Subject[non-cognate]}$	-0.25
ICC	0.16
$N_{Subject}$	58
N_{Item}	192

Observations	4,650
Marginal R^2/ Conditional R^2	0.174/0.310

6.F EEG data: region of interest

Figure 6.F.1: *Region of interest and the corresponding data channels for the group comparison, illustrated in the montage of the German-Spanish group.*



6.G Model parameters: P300 component

Table 6.G.1: *Model outcome parameters for voltage amplitudes (n = 58).*

Formula: voltage amplitudes ∼ gender congruency (congruent vs. incongruent) + cognate status (cognate vs. non-cognate) + hemisphere (left vs. midline vs. right) + familiarisation phase performance (zero correct vs. one correct vs. two correct vs. three correct) + (gender congruency + cognate status|subject) + (1|item)

Term	Estimate [95% CI]	t-value	p-value
(Intercept)	4.96 [4.17, 5.74]	12.40	< 0.001
Gender congruency [incongruent]	0.025 [-0.479, 0.529	0.097	0.922
Cognate status [non-cognate]	-0.477 [-0.934, -0.021]	-2.05	**0.040**
Hemisphere [midline]	0.619 [0.598, 0.639]	59.14	< 0.001
Hemisphere [right]	-0.426 [-0.444, -0.408]	-45.63	< 0.001
Familiarisation phase performance [one correct]	0.009 [-0.052, 0.070]	0.292	0.770
Familiarisation phase performance [two correct]	0.113 [0.056, 0.170]	3.87	< 0.001
Familiarisation phase performance [three correct]	0.172 [0.114, 0.230]	5.81	< 0.001
Random effects			
σ^2	73.66		
τ_{00Item}	1.78		
$\tau_{00Subject}$	7.60		
$\tau_{11Subject[incongruent]}$	1.68		
$\tau_{11Subject[non-cognate]}$	0.99		
$\rho_{01Subject[incongruent]}$	-0.42		

$\rho_{01 Subject[non-cognate]}$	-0.001
ICC	0.11
$N_{Subject}$	58
N_{Item}	192
Observations	4,500,216
Marginal R^2/ Conditional R^2	0.003/0.112

CHAPTER 7

When left is right: The role of typological similarity in multilinguals' inhibitory control performance

This article is published as: Von Grebmer Zu Wolfsthurn, S., Gupta, A., Pablos-Robles, L., & Schiller, N. O. (2022). When left is right: The role of typological similarity in multilinguals' inhibitory control performance. Bilingualism: Language and Cognition, 25(5), 1-14.

Abstract: Both inhibitory control and typological similarity between two languages feature frequently in current research on multilingual cognitive processing mechanisms. Yet, the modulatory effect of speaking two typologically highly similar languages on inhibitory control performance remains largely unexplored. However, this is a critical issue because it speaks directly to the organisation of the multilingual's cognitive architecture. In this study, we examined the influence of typological similarity on inhibitory control performance via a spatial Stroop paradigm in native Italian and native Dutch late learners of Spanish. Contrary to our hypothesis, we did not find evidence for a differential Stroop effect for the typologically similar group (Italian-Spanish) compared to the typologically dissimilar group (Dutch-Spanish). Our results therefore suggest a limited influence of typological similarity on inhibitory control performance. The study has critical implications for characterising inhibitory control processes in multilinguals.

Keywords: *inhibitory control performance, typological similarity, spatial Stroop task, late language learners*

7.1 Introduction

A remarkable feature of multilingual speakers is the ability to engage with several acquired languages, seemingly without effort. In this paper, we will broadly refer to multilinguals as those language users who have acquired one or more non-native language(s) in addition to their native language, L1 (Cenoz, 2013; De Groot, 2017). Over the past decades, numerous studies have attempted to capture the complexity of the multilingual experience. In particular, they focused on the cognitive, structural and functional consequences of managing several languages in the brain (Abutalebi & Green, 2007; Bialystok et al., 2012; D. W. Green, 1998; Kroll et al., 2015; Mosca & De Bot, 2017; Pliatsikas, 2020; Schwieter, 2016) (Sebastián-Gallés & Kroll, 2003).

A well-established aspect of the cognitive architecture of multilingualism is the parallel activation of languages across a range of proficiency levels, language combinations and linguistic domains (Blumenfeld & Marian, 2013; Colomé, 2001; Costa et al., 2000; Dijkstra, Van Heuven & Grainger, 1998; Guo & Peng, 2006; Hoshino & Thierry, 2011). In order to successfully mitigate parallel activation and to ultimately select the appropriate target language, multilinguals must employ a language control mechanism on the non-target language (Abutalebi & Green, 2007; Christoffels et al., 2007; Costa & Santesteban, 2004; Declerck et al., 2019; D. W. Green, 1998). Here, language control is conceptualised as a collection of control mechanisms applied to multilingual speech production and comprehension (Abutalebi, 2008; D. W. Green & Abutalebi, 2013). From a theoretical point of view, this notion is featured in the Inhibitory Control (IC) model of language control by D. W. Green (1998), which postulates that the non-target language needs to be suppressed prior to the linguistic output.

The exact nature of the mechanisms underlying language control is yet to be established. There is a substantial amount of evidence suggesting that language control is strongly associated with domain-general inhibitory control, also termed cognitive control or executive control (Bialystok et al., 2012; Declerck et al., 2021; Festman, Rodriguez-Fornells & Münte, 2010). Inhibitory control is an executive function used to regulate and inhibit irrelevant information with respect to thoughts or behaviour, as well as switching attention (Diamond, 2013; Miyake et al., 2000). Some studies indicate that language control impacts executive functions, for example inhibitory control (Bialystok & Martin, 2004; Bialystok, 2010; Kroll & Bialystok, 2013; D. W. Green & Abutalebi, 2013; Miyake et al., 2000; Wiseheart, Viswanathan & Bialystok, 2016). Critically, evidence further suggests that language control may share some underlying processing mechanisms with inhibitory control (Declerck et al., 2021; D. W. Green, 1998; Linck, Hoshino & Kroll, 2005; Weissberger, Gollan, Bondi, Clark & Wierenga, 2015), although this notion is still debated (Branzi, Della Rosa, Canini, Costa & Abutalebi, 2016; Calabria, Hernandez, Branzi & Costa, 2012).

In the current study, we investigated the impact of multilingualism on inhibitory control performance. More specifically, we examined whether the typological similarity between languages of a multilingual plays a role in modulating inhibitory control performance. Typological similarity, also termed typological distance or language similarity in the literature, refers to linguistic and structural (dis)similarities across different languages spoken by multilinguals (Foote, 2009; Putnam et al., 2018; Westergaard, Mitrofanova, Mykhaylyk & Rodina, 2017). For example, Italian and Spanish may be considered as more typologically similar languages compared to language pairs such as Dutch and Spanish because of the larger degree of overlap in morphosyntax, gender systems and cognates

(Paolieri et al., 2019; Schepens et al., 2012; Serratrice et al., 2012).

Several studies have focused on the modulating effects of typological similarity on language control, for example, within the context of a classical Stroop paradigm (Brauer, 1998; Coderre, Van Heuven & Conklin, 2013; Van Heuven, Conklin, Coderre, Guo & Dijkstra, 2011) However, studies directly investigating the effect of typological similarity on domain-general inhibitory control performance are scarce, but see Bialystok et al. (2005), Linck et al. (2005) and Yamasaki et al. (2018). Typical experimental paradigms to explore domain-general inhibitory control are the Simon task (Bialystok, Craik, Klein & Viswanathan, 2004; Bialystok et al., 2005; Simon & Small, 1969), and the spatial Stroop task (Hilbert et al., 2014; Lu & Proctor, 1995; Luo & Proctor, 2013). The core feature of the Simon task is a conflict between the physical location of a stimulus and the response, e.g., a stimulus appearing on the right side of a screen while the corresponding response button is located on the left side. The Simon effect quantifies the difference in response times (RTs) between trials in which stimulus and response location match and trials in which stimulus and response location mismatch. Typically, longer RTs are linked to the mismatch trials. Accordingly, a smaller Simon effect reflects better inhibitory control performance, whereas a larger Simon effect reflects lower inhibitory control performance (Bialystok et al., 2004).

In this study, we used the spatial Stroop task (Hilbert et al., 2014; Lu & Proctor, 1995), which is a combination of the Simon task and the classical colour-word Stroop task (MacLeod, 1992; Stroop, 1935). While the classical Stroop task involves the naming of a colour-word written in either the matching ink colour (congruent trial), e.g., the word *RED* written in red ink, or the mismatching ink colour (incongruent trial), e.g., the word *RED* written in blue ink, the spatial Stroop task focuses on spatial stimulus-stimulus conflicts. The basic feature of the spatial Stroop task is that a target word (“left”, “right”, “up”, “down”) either matches its location on the screen, e.g., *LEFT* shown on the left side of the screen (*congruent* trial), or it does not match its location on the screen, e.g., *LEFT* shown on the right side of the screen (*incongruent* trial). The key to success in this task is to inhibit the irrelevant spatial stimulus property (e.g., the location of the word) and to instead focus on the relevant target stimulus property (the target word itself). In this task, inhibitory control performance is reflected in the spatial *Stroop effect*, which describes the quantitative difference in RTs between *congruent* and *incongruent* trials (Hilbert et al., 2014; La Heij, Van der Heijden & Plooij, 2001; Marian et al., 2013; Roelofs, 2021; Van Heuven et al., 2011). Drawing parallels between the Simon task, a smaller Stroop effect is reported to indicate better inhibitory control performance (Costa, Hernández & Sebastián-Gallés, 2008; Heidlmayr et al., 2014; Pardo, Pardo, Janer & Raichle, 1990).

In the current study, the critical question we sought to answer was the following: does typological similarity between the two languages significantly modulate inhibitory control performance in multilinguals? A relevant theoretical framework for this particular question is the Conditional Routing Model (CRM) by Stocco et al. (2014). The model is based upon the notion that the multilingual experience dynamically impacts domain-general executive functions, including inhibitory control, as a result of the parallel activation of the languages (Bialystok & Martin, 2004; Festman et al., 2010). Here, the model postulates that executive functions are effectively trained over time (Kroll & Bialystok, 2013; Yamasaki et al., 2018), which results in a strengthening of the neural circuits underlying these executive functions. When the languages within a multilingual system are highly typologically similar, one may predict a higher degree of cross-language interference (Cenoz, 2001; J. Chen, Zhao, Zhaxi & Liu, 2020; De Bot, 2004). In turn, this implies that speakers of these languages develop better inhibitory control skills compared to speakers of typologically less similar languages (Yamasaki et al., 2018). Therefore, the CRM provides us with a testable prediction for the effect of typological similarity on inhibitory control performance: speakers of typologically similar languages should exhibit a better inhibitory control performance compared to speakers of typologically less similar languages. Applied to the context of a spatial Stroop task used in this study, speakers of typologically similar languages (e.g., Italian-Spanish) should therefore show a smaller Stroop effect compared to speakers of typologically less similar languages (e.g., Dutch-Spanish speakers).

7.1.1 The current study

We explored the modulatory role of typological similarity on inhibitory control performance in a spatial Stroop task (hereafter simply *Stroop task*) in two groups of speakers with differing degrees of typological similarity. Participants were native Italian learners of Spanish, and native Dutch learners of Spanish. On the basis of typological work by Schepens et al. (2012) and Van der Slik (2010), we defined our Italian-Spanish group as our typologically similar group, and our Dutch-Spanish group as our typologically dissimilar group. All participants had a Spanish proficiency level in the B1/B2 range within the CEFR framework (Council of Europe, 2001). We followed a spatial Stroop paradigm inspired by Hilbert et al. (2014), who used the location words "left", "right", "up" and "down" to study the Stroop effect in native speakers of German: see also Lu and Proctor (1995) and Shor (1970). In our paradigm, we exploited the conflict between the target word and the location of the target word on the screen, for example, the Spanish location word [*izquierda*] "left" displayed on the right side of the screen, or the Spanish word [*derecha*] "right" displayed on the left side of the screen. The translation equivalents for [*izquierda*] "left" and [*derecha*] "right" are "sinistra" and "destra" in Italian, and "links" and "rechts" in Dutch, respectively. In the *congruent* condition, the target word and the target word location matched. In contrast, in the *incongruent* condition, the target word and the target word location did not match. We measured accuracy and RTs during this task. Post-experiment, we calculated the Stroop effect by subtracting the RTs for congruent trials from RTs for incongruent trials. Importantly, we employed an equiprobable Stroop task design, whereby the probability of each condition occurring in the subsequent trial is identical. Within the framework of the Dual Mechanisms of Control

(DMC) model (Braver, 2012), an equiprobable Stroop task design is linked to a *proactive* control strategy. At the core of this particular strategy is the maintenance of goal-relevant information over time to succeed at the task (Braver, 2012; Gonthier, Braver & Bugg, 2016). Therefore, our Stroop task taps not only into inhibitory control performance per se, but also into the cognitive mechanisms of monitoring the task.

Research questions

Our research questions were the following: first, is there a difference in terms of RTs as a function of typological similarity (typologically similar vs. typologically dissimilar)? Secondly, connected to this first question, is the Stroop effect larger for one group compared to the other, thereby reflecting an effect of typological similarity on inhibitory control performance?

Hypotheses

Based on the literature outlined above, we first predicted a Stroop effect for both the Italian-Spanish group and the Dutch-Spanish group. Behaviourally speaking, this would be reflected in higher accuracy and shorter RTs for congruent trials compared to incongruent trials. Next, in line with the CRM (Stocco et al., 2014), we hypothesised overall shorter RTs for the Italian-Spanish group compared to the Dutch-Spanish group. Finally, we expected a difference in inhibitory control performance as a function of typological similarity: we expected an interaction effect of condition (congruent vs. incongruent) and typological similarity (typologically similar vs. typologically dissimilar) on the size of the Stroop effect. A smaller Stroop effect for the Italian-Spanish group would imply that the overall inhibitory control performance is better for the typologically similar languages compared to the less typologically similar Dutch-Spanish group. In turn, this would support the CRM (Stocco et al., 2014).

7.2 Methods

In addition to the spatial Stroop task, we asked participants to complete the Language Experience and Proficiency Questionnaire, LEAP-Q (Marian et al., 2007). The LEAP-Q is a questionnaire designed to obtain a measure for the linguistic profile of our participants in terms of their proficiency levels and experiences with the languages within their multilingual system (Marian et al., 2007). Finally, participants also completed the LexTALE-Esp (Izura et al., 2014), a lexical decision task to establish vocabulary size in Spanish, for descriptive purposes.

7.2.1 Participants

For the Italian-Spanish group, we recruited 33 healthy, right-handed native speakers of Italian (24 females) with a B1/B2 level of Spanish at Pompeu Fabra University (Barcelona, Spain). Mean age of the Italian-Spanish group was 27.12 years (*SD* = 4.08). Our recruitment criteria for this group were the following: no additional language learnt before the age of three, age of acquisition of Spanish from fourteen years onwards, a maximum time spent in a Spanish-speaking country of no longer than one year, no psychological, neurological, visual, auditory, or language-related impairments; and finally, an age range between 18 and 35 years. For the Dutch-Spanish group, we recruited and tested 25 healthy, right-handed native speakers of Dutch (16 females) with a B1/B2 level of Spanish at Leiden University (Leiden, The Netherlands). Mean age of the Dutch-Spanish group was 22.84 years (*SD* = 3.05). Our recruitment criteria were identical to the Italian-Spanish group, with the cap on maximum time spent in a Spanish-speaking country less stringent due to the testing location. Data from the LEAP-Q was analysed to establish a detailed linguistic profile of each participant. See Appendix 7.A and Appendix 7.B for an overview of the profiles for the Italian-Spanish group and the Dutch-Spanish group, respectively.

LEAP-Q: Linguistic profile of participants

Italian-Spanish group. With respect to their linguistic profile in Spanish, two participants acquired Spanish as first foreign language, whereas eighteen participants acquired Spanish as second foreign language. Spanish was the third foreign language for ten participants, and three participants acquired Spanish as fourth foreign language (Appendix 7.A). The mean age of acquisition (AoA) of Spanish was 23.93 years (*SD* = 5.07). On average, participants reported to be fluent in Spanish at the age of 24.88 years (*SD* = 4.48), to have started reading in Spanish at the age of 24.36 years (*SD* = 4.91) and to be fluent readers by the age of 24.24 (*SD* = 4.82). On average, participants spent 0.46 years (*SD* = 0.343) in a Spanish-speaking country and had learnt Spanish for 0.93 years (*SD* = 1.17) either at school as a foreign language, or as a language course in Spain. Twenty-five participants were completing or had completed a formal Spanish language course that was not part of the school curriculum shortly before or upon their arrival in Spain (mean length of course: 0.53 years, *SD* = 0.889 years). Finally, participants quantified their current daily exposure to Spanish as 40% (*SD* = 18.37%) of the time with respect to the other languages spoken. In terms of dominance, thirteen participants classified Spanish as their most dominant language after Italian, fourteen participants as their second most dominant language after Italian, five participants as their third most dominant language after Italian, and one participant as their fourth most dominant language after Italian. On a ten-point scale, ten being maximally proficient, participants rated their speaking proficiency at 6.09 (*SD* = 1.76), comprehension proficiency at 7.26 (*SD* = 1.67) and their reading proficiency at 7.36 (*SD* = 1.48).

Dutch-Spanish group. In this group, nine participants stated that they acquired Spanish as their second foreign language, nine participants as their third foreign language and seven participants as their fourth foreign language (Appendix 7.B). Mean AoA of Spanish was 17.84 years (*SD* = 3.16). People stated to be fluent in Spanish on average at the age of 19.60 years (*SD* = 2.52), that they started reading in Spanish at the age of 18.44 years (*SD* = 3.24), and that they were on average fluent in reading by 19.76 years (*SD* = 3.41). Eighteen out of the twenty-five participants spent on average 0.57 years (*SD* = 0.66) in a Spanish-speaking country (e.g., Spain, Argentina, Colombia, Mexico). Compared to the other languages, participants quantified their daily exposure to Spanish with 12.96% (*SD* = 10.07). Critically, two participants reported Spanish as their second most dominant language, nineteen as their third most dominant, three as their fourth most dominant and one participant as their fifth most dominant language following Dutch. On a ten-point scale (ten being maximally proficient), participants reported an average speaking proficiency in Spanish of 6.40 (*SD* = 1.47), a comprehension proficiency of 7.08 (*SD* = 1.32) and a reading proficiency of 7.08 (*SD* = 1.22). These ratings are highly comparable with the Italian-Spanish group.

7.2.2 Materials and design

Prior to the experiment, participants completed the LEAP-Q (Marian et al., 2007) at home to reduce self-report biases frequently induced in laboratory settings (Rosenman et al., 2011). During the experiment, we first asked participants to complete the LexTALE-Esp (Izura et al., 2014), followed by the Stroop task.

Tasks and stimuli

LexTALE-Esp. We administered the LexTALE-Esp to establish vocabulary size in Spanish. The task was programmed in E-prime2 (Schneider et al., 2002), using the exact same stimuli as in the original version by Izura et al. (2014).

Stroop task. We administered the Stroop task to measure inhibitory control performance in our Italian-Spanish speakers and Dutch-Spanish speakers. We again generated an E-prime2 script (Schneider et al., 2002) for this task. The target words were the written Spanish words [*izquierda*] “left” and [*derecha*] “right”.

7.2.3 Procedure

Prior to initiating the experiment, participants were provided with an information sheet and the opportunity to ask clarification questions. Then, participants signed the consent form in compliance with the ethics code for linguistic research at the Faculty of Humanities at Leiden University. Before each task, we provided participants with written task instructions in Spanish. Upon termination of all tasks, participants were provided with a debrief sheet in their respective L1, they signed the final consent form and received a monetary compensation for their participation.

LexTALE-Esp

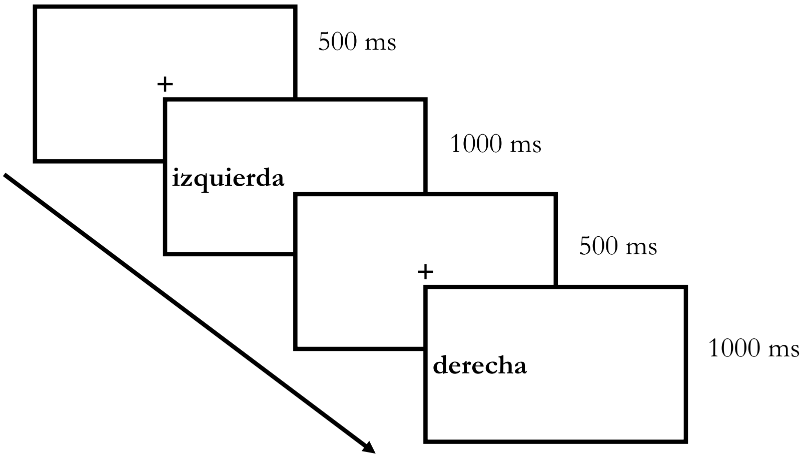
The LexTALE-Esp procedure was identical for both groups. We asked participants to indicate via a button press whether the string corresponded to a Spanish *word* (e.g., [*secuestro*] "kidnapping") or a *pseudoword* (e.g., *plaudir*). Participants were instructed that incorrectly assigning a word status to a pseudoword and vice versa would lead to a deduction in the score. The trial procedure was as follows: first, a black fixation cross was displayed for 1,000 ms on a white screen. Then, a letter string corresponding to either a word or a pseudoword was displayed in the centre of the screen. The letter string remained on the screen until the participants' response. After the participants' response, the next trial was initiated. Sixty trials were Spanish word trials, whereas thirty were pseudoword trials. Trial order was randomised so that each participant was presented with a unique trial order.

Stroop task

The procedure for the Stroop task was the same for both groups. Participants were asked to focus on the target word while ignoring the location of the target word on the screen and to respond to the target word via button presses. The trial procedure was as follows: first, participants saw a black fixation cross for 500 ms in the centre of a white screen. Next, they saw a target word appear on either the left or right side of the screen along the horizontal midline in

Spanish. This target word was either [*izquierda*] “left” or [*derecha*] “right”. The target word was visible on the screen until participants responded or for a maximum display time of 1,000 ms (Figure 7.2.1).

Figure 7.2.1: *Example trial procedure for a congruent trial followed by an incongruent trial.*



The next trial was initiated after participants’ response, or if the response time limit was reached. Half of the trials were *congruent* trials, where the target word matched the location on the screen. The other half of the trials were *incongruent* trials, where the target word and the location on the screen did not match. There were 24 trials for each target word (*izquierda/ derecha*) and target location on the screen (left side/right side), amounting to 48 trials for the congruent condition and 48 trials for the incongruent condition. Prior to the start of the main experimental round, there was a short practise round to familiarise participants with the task procedure. Trial order was randomized in the practice round and in the main experimental round.

7.3 Results

7.3.1 Data exclusion

For the Italian-Spanish group, data from one participant were lost due to a technical failure. Therefore, we included 32 datasets in the analysis. In contrast, for the Dutch-Spanish group we included all 25 datasets in the analysis, adding to a total of 57 datasets.

7.3.2 Data analysis

We analysed our behavioural data using R, Version 4.0.3 in RStudio, Version 1.4.1106 (R Core Team, 2020). We employed a single trial generalised linear mixed effects modelling approach using the *lme4* package (Bates et al., 2020). We first modelled the outcome variables *accuracy* and *RTs* separately for each individual group. Next, we pooled our RT data from both groups for a group comparison analysis to study potential effects of typological similarity on *Stroop effect sizes.* For both the individual group analyses and the group comparison analysis, we applied the following model fitting procedure: first, we constructed a theoretically plausible model with a maximal random effects structure as supported by our data (Barr, 2013; Matuschek et al., 2017). In our case, the maximal model was a random-intercept and random-slope model for both accuracy and RTs. In the case of non-convergence or singular fit, we simplified our random effects structure. Next, we generated the model of best fit in a top-down procedure, whereby we simplified the fixed effects structure in a stepwise fashion. After fitting each model, we performed model diagnostics to establish the goodness of fit using the *DHARMa* package (Hartig, 2020). This involved the plotting of the model residuals against the predicted values, and closely investigating the distribution of the residuals and the presence of influential data points to identify issues in terms of the model fit. Then, we compared models with different fixed effects structures to establish the model of best fit using the *anova()* function, which is based on the Akaike's Information Criterion, AIC (Akaike, 1974), the Bayesian Information Criterion, BIC (Neath & Cavanaugh, 2012) and the log-likelihood ratio. To test for the significance of the terms in the fixed effects structure, absolute test-statistics greater than 1.96 were interpreted as statistically significant at $\alpha = 0.05$ (Alday et al., 2017; Matuschek et al., 2017). Finally, the models of best fit for RTs were re-fitted using the REML criterion (Bates, Mächler, Bolker & Walker, 2014; Verbyla, 2019). All best-fitting models and model parameters are reported in Appendix 7.C, Appendix 7.D and Appendix 7.E. Note that the model parameters for accuracy are reported as odds ratios.

To model accuracy, we used the *glmer()* function with a binomial distribution. This particular function from the *lme4* package uses maximum likelihood estimation via the Laplace approximation (Bates et al., 2020). In contrast, we used the *lmer()* function with a normal distribution to model RTs for correct trials. For the individual group analysis, our fixed effect of interest was *condition* (congruent vs. incongruent), whereas *subject* and *item* were included as random effects. For the group comparison analysis, we used the *lmer()* function to model the interaction effect of *condition* (congruent vs. incongruent) and *typological similarity* (typologically similar vs. typologically dissimilar) as well as their main effects on RTs. *Subject* and *item* were again included as random effects. To control for potential covariates, we included *LexTALE-Esp score* and *order of acquisition of Spanis*h as fixed effects in all analyses.

7.3.3 LexTALE-Esp

Post-experiment, we calculated the LexTALE-Esp vocabulary size score for each participant by subtracting the percentage of incorrectly identified pseudowords from the percentage of correctly identified words (Izura et al., 2014). For the Italian-Spanish group, the mean LexTALE-Esp score was 26.30 (*SD* = 14.04). Large individual differences were evident from the range of scores, which was between -7.37 to 49.30. In contrast, the mean LexTALE-Esp score for the Dutch-Spanish group was 22.69 (*SD* = 17.19). The range was from -11.92 to 54.73, which yielded similar large individual differences between participants. A two-sample t-test yielded no significant statistical difference in LexTALE-Esp scores between the two groups with $t(45.90) = 0.851$, $p = 0.399$. According to calculations provided by Lemhöfer and Broersma (2012), all of our speakers were at or below the B2 level for Spanish according to CEFR standards (Council of Europe, 2001), in line with our recruitment criteria.

7.3.4 Stroop task

We first computed descriptive statistics for accuracy and RTs for both groups. See Table 7.3.1 for descriptive mean accuracy, mean RTs and Stroop effects for the Italian-Spanish group and the Dutch-Spanish group. Descriptively speaking, results yielded overall longer RTs for the Italian-Spanish group compared to the Dutch-Spanish group. Moreover, the Stroop effect was descriptively larger for the typologically similar languages compared to the typologically dissimilar languages. We first discuss the individual analysis for the Italian-Spanish group and the Dutch-Spanish group, respectively. Then, we discuss the group comparison for the Stroop effect.

Table 7.3.1: *Mean accuracy and RTs for the Stroop task for the Italian-Spanish group (n = 32) and the Dutch-Spanish group (n = 25).*

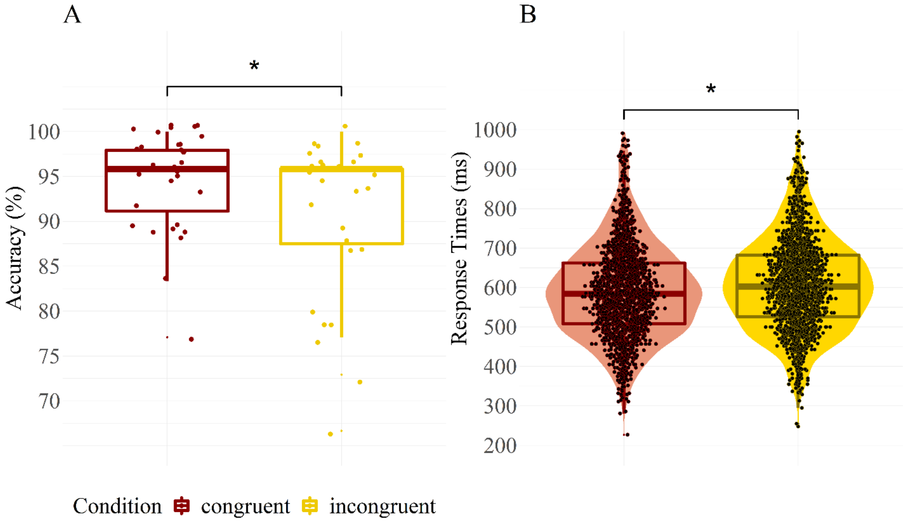
	Italian		**Dutch**	
	Accuracy (%)	**RTs (ms)**	**Accuracy (%)**	**RTs (ms)**
Condition	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
congruent	94.73 (22.35)	579 (113)	95.67 (20.37)	560 (123)
incongruent	91.21 (28.32)	607 (112)	90.67 (29.10)	576 (112)
Stroop effect		**28**		**16**

Italian-Spanish group

Accuracy. For the Italian-Spanish group, the model of best fit included *condition* as fixed effect, as well as *subject* and *item* as random effects. The by-*subject* random slopes for *condition* led to singular fit and were therefore dropped from the model fitting procedure. The covariates *LexTALE-Esp score* and *order of acquisition of Spanish* did not significantly improve the model fit. Therefore, the model of best fit was the following: accuracy $\sim$ condition (congruent vs. incongruent) + (1|subject) + (1|item). Participants were significantly more accurate for congruent trials compared to incongruent trials with β = 0.560, *95% CI*[0.400, 0.784], z = -3.38, p = 0.001 (see Appendix 7.C for the full model parameters). See Figure 7.3.1 for mean accuracy for the Italian-Spanish group.

Response times. For the Italian-Spanish group, the model of best fit yielded an effect of *condition*, a random effect for *subject* and *item* and a by-*subject* random slope for *condition* (Figure 7.3.1). Neither *LexTALE-Esp score* nor *order of acquisition of Spanish* significantly modulated the outcome variable or improved the model fit. These two covariates were therefore excluded from the model fitting procedure. This resulted in the following best-fitting model: RTs $\sim$ condition (congruent vs. incongruent) + (condition|subject) + (1|item). Participants were statistically faster in responding in the congruent condition compared to the incongruent condition with β = 29.46, *95% CI*[18.10, 40.82], t = 5.09, $p < 0.001$ (see Appendix 7.C).

Figure 7.3.1: *Mean accuracy (A) for each participant and response times (B) for each condition for the Italian-Spanish group (n = 32).*



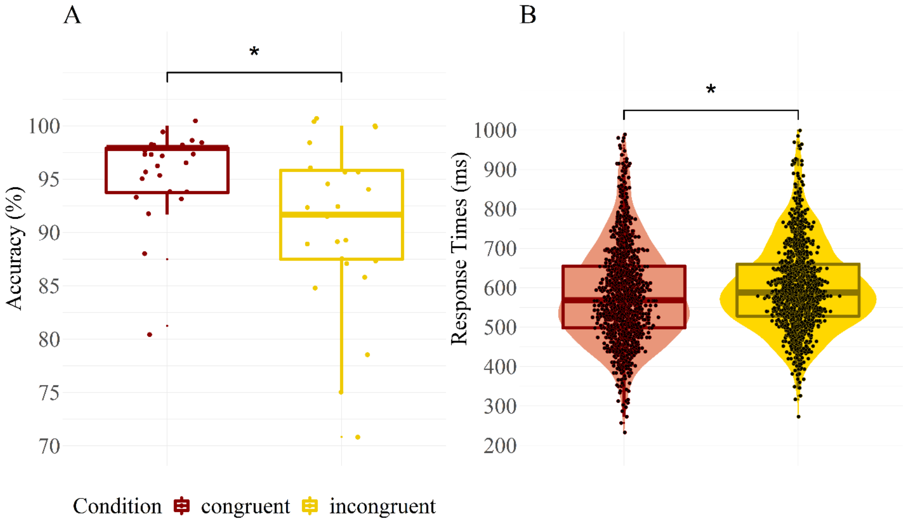
Dutch-Spanish group

Accuracy. For the Dutch-Spanish group, the model of best fit included a fixed effect of *condition*, as well as by-*subject* random slopes for *condition* and *subject* as random effect. *Item* led to singular fit and was excluded from the model fitting procedure. Further, *Lextale-Esp score* and *order of acquisition of Spanish* did not significantly improve the model fit. The model of best fit was as follows: accuracy ∼ condition (congruent vs. incongruent) + (condition|subject). Participants were significantly more accurate in the congruent compared to the incongruent condition with β = 0.477, *95% CI* [0.295 – 0.771], z = -3.02, p = 0.003 (see Appendix 7.D for the full model parameters). See Figure 7.3.2 for mean accuracy for the Dutch-Spanish group.

Response times. For the Dutch-Spanish group, we found that the model of best fit included *condition* as fixed effect, *subject* as random effect and by-*subject* random slopes for *condition* (Figure 7.3.2). The random effect for *item* was not supported by our data and was therefore excluded from the random effects structure.

Neither *LexTALE-Esp score* nor *order of acquisition of Spanish* significantly improved the model fit and were subsequently dropped from the model selection procedure. The resulting model of best fit was the following: RTs $\sim$ condition (congruent vs. incongruent) + (condition|subject). Participants were significantly faster in responding in the congruent condition compared to the incongruent condition, with $\beta = 15.31$, *95% CI*[4.40, 26.21], $t = 2.75$, $p = 0.006$ (see Appendix 7.D).

Figure 7.3.2: *Mean accuracy (A) for each participant and response times (B) for each condition for the Dutch-Spanish group (n = 25).*

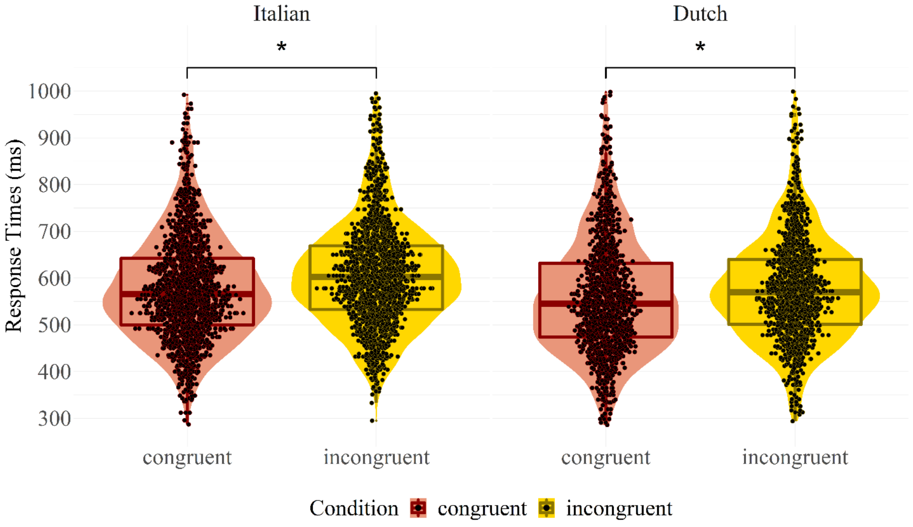


Taken together, data from both the Italian-Spanish group and the Dutch-Spanish groups suggest that participants were significantly more accurate and faster in the *congruent* condition compared to the *incongruent* condition. Therefore, both groups displayed the Stroop effect.

Stroop effect: group comparison

Finally, we compared the Stroop effect (RT incongruent trials minus RTs congruent trials) between the Italian-Spanish and Dutch-Spanish group to explore the possible impact of typological similarity. Here, we explored the interaction effect between *condition* and *typological similarity* on the size of the Stroop effect. Descriptively speaking, the Stroop effect was larger for the Italian-Spanish group compared to the Dutch-Spanish group. However, the model of best fit yielded a main effect of *condition* with participants being faster for *congruent* trials compared to *incongruent* trials with $\beta = 23.20$, *95% CI*[14.56, 31.83] = 4.40, $t = 5.27$, $p < 0.001$. The model also included a main effect of *typological similarity*, with participants from the typologically similar group (Italian-Spanish) being significantly slower compared to the typologically dissimilar group (Dutch-Spanish) with $\beta = 30.70$, *95% CI*[7.72, 53.68] = 11.72, $t = 2.62$, $p = 0.009$. Therfore, the best-fitting model was: RTs $\sim$ typological similarity (high vs. low) + condition (congruent vs. incongruent) + (condition|subject) + (1|item), see Appendix 7.E. There was no evidence for an interaction effect between *condition* and *typological similarity*. See Appendix 7.E for full model specification details, as well as a comparison between the model that included the interaction term and the best-fitting model that did not include the interaction term. Further, the model of best fit also included a by-*subject* random slope for *condition* as well as *item* as random effect. The covariates *LexTALE-Esp score* and *order of acquisition of Spanish* did not significantly contribute to improving the model fit and were therefore not included in the final model. See Figure 7.3.3 for the comparison of the Stroop effect across the Italian-Spanish and Dutch-Spanish group.

Figure 7.3.3: *Mean response times for the Italian-Spanish group (left) and the Dutch-Spanish group (right) for each condition for Spanish Stroop targets (n = 57).*



7.4 Discussion

In this study, we explored the effect of typological similarity on inhibitory control performance in a group of Italian-Spanish speakers and a group of Dutch-Spanish speakers via a spatial Stroop task. The goal of this study was twofold: first, we examined whether or not the typologically similar (Italian-Spanish) group showed a general processing advantage over the typologically dissimilar (Dutch-Spanish) group in terms of RTs. Secondly, we studied whether typological similarity yielded a difference between the two groups in terms of the Stroop effect (difference in RTs between incongruent and congruent trials). Here, a smaller Stroop effect would be indicative of better inhibitory control performance. On the basis of the CRM (Stocco et al., 2014), we expected shorter RTs and a smaller Stroop effect for the Italian-Spanish group compared to the Dutch-Spanish group.

Stroop data from the Italian-Spanish as well as the Dutch-

Spanish group showed that participants were sensitive to the inherent task conflict. More specifically, results demonstrated higher accuracy and shorter RTs for congruent compared to incongruent trials. This yields the typical *Stroop effect*, which is a measure of inhibitory control performance in this task. To succeed at this task, participants had to ignore the irrelevant information (i.e., the location of the target) and instead focus on the target word itself to provide a correct response. Further, as discussed in the introduction, participants had to employ a proactive control strategy (Braver, 2012; Gonthier et al., 2016) and monitor the goal-relevant information during the task, as described in the DMC model (Braver, 2012). Therefore, the presence of a Stroop effect in both groups reflects not only a measure for inhibitory control performance, but also a monitoring strategy to solve this task.

With respect to the first research question, the group comparison analysis showed that the typologically dissimilar (Dutch-Spanish) group was comparatively faster than the typologically similar (Italian-Spanish) group in this task. This finding contrasts with our predictions. The original prediction on the basis of the CRM (Stocco et al., 2014) was a processing advantage for the typologically similar Italian-Spanish group compared to the Dutch-Spanish group due to continuous training of executive functions and inhibitory control skills over time. In contrast, our findings suggest that typologically dissimilar Dutch-Spanish group had a processing advantage in terms of RTs over the Italian-Spanish group. In the literature, similar findings were reported by Bialystok et al. (2005), who investigated the role of typological similarity on the performance during a Simon task in highly proficient Cantonese-English speakers (typologically dissimilar group) and highly proficient French-English speakers (typologically more similar group). Results showed a processing advantage for Cantonese-English speakers compared to the French-English speakers in the form of faster RTs on the Simon task for Cantonese-English speakers, see also Linck et al. (2005). Our results are comparable to Bialystok et al. (2005), and suggest that in this particular task, typological dissimilarity was advantageous over typological similarity. Moreover, these results suggest a qualitative difference between the Italian-Spanish and the Dutch-Spanish group, namely a more efficient inhibitory control strategy for the speakers of the less typologically similar languages. Within the framework of the DMC model (Braver, 2012) and the application of proactive control strategies during this task (Braver, 2012; Gonthier et al., 2016), this implies that Dutch-Spanish speakers were more effective at employing a proactive control strategy, as reflected in overall shorter RTs. In other words, speakers of typologically more dissimilar languages were better at monitoring and actively maintaining goal-related information compared to speakers of typologically similar languages. This has critical implications for the conceptualisation of the underlying cognitive mechanisms for typologically similar vs. dissimilar language combinations.

With respect to our second research question, there was a descriptive trend of a smaller Stroop effect for the Dutch-Spanish group compared to the Italian-Spanish group. However, the overall processing advantage of the Dutch-Spanish group over the Italian-Spanish group was not reflected in the size of the Stroop effect. More concretely, we did not find a statistically significant difference between the Stroop effect size for the Italian-Spanish group compared to the Dutch-Spanish group. This finding was somewhat surprising and contrasts with our original predictions. Our result suggested, first, that the Stroop effect was unaffected by typological similarity, and second, that speakers of both groups demonstrated a highly comparable inhibitory control performance in this task. Importantly, the CRM framework proposed by Stocco et al. (2014) does not fully account for these specific findings. Instead, our findings strongly suggest a limited modulatory role of typological similarity on inhibitory control performance in this study. One arising question here is the following: why were the Dutch-Spanish speaker faster, but not better, compared to the Italian-Spanish speakers at performing the Stroop task?

One interpretation of our findings could be that factors other than typological similarity influence inhibitory control performance in this task. These other potentially modulating factors exert their influence such that one group had an advantage in terms of processing speed, but not in terms of overall performance. A well-established modulatory factor in language control, but less in inhibitory control, is language proficiency, as postulated in the IC model (D. W. Green, 1998). Previous studies have shown that multilingual children with a low non-native proficiency display unilateral cross-language interactions from the L1 into the L2 compared to multilingual children with high non-native proficiency (Brenders, Van Hell & Dijkstra, 2011; Poarch & Van Hell, 2012a). As outlined in Poarch and Van Hell (2012b), this could indicate that less language control effort is needed to manage the native and the non-native languages. In turn, this implies less training of more general executive control functions such as inhibitory control if the difference in proficiency levels between the native and non-native language is considerable. More specific to our intermediate late learners of Spanish, one could argue that our participants have not yet sufficiently trained their inhibitory control skills given their intermediate level of non-native proficiency, in turn accounting for a limited effect of typological similarity in this study. Therefore, one possibility is that there is an interaction effect between typological similarity and non-native proficiency, and only a particular degree of typological similarity paired with a specific proficiency level leads to training of the inhibitory control skills. This tentative hypothesis is partially in line with language control research by Brauer (1998). This study explored the effect of typological similarity on language control via the *within-language Stroop effect* and the *between-language Stroop effect* in speakers of typologically similar languages (German-English) and typologically dissimilar languages (English-Greek and English-Chinese) in the classical Stroop paradigm. The within-language Stroop effect refers to the differences in RTs between the congruent and incongruent condition when the stimulus and response languages are identical. On the other hand, the between-language Stroop effect quantifies the differences in RTs between the congruent and incongruent condition when the stimulus and response languages are different (Brauer, 1998; Marian et al., 2013; Van Heuven et al., 2011). Critically, Brauer (1998) included low- and high-proficient speakers to also explore the effect of proficiency on inhibitory control performance. All three groups showed a within-language and a between-language Stroop effect. On the one hand, low proficiency in the non-native language was linked to larger differences between the within-language and the between-language Stroop effect across the native and non-native language, irrespective of typological similarity. On the other hand, highly proficient speakers in the typologically dissimilar group were linked to larger within-language compared to between-language Stroop effects in both the native and non-native language. Importantly, highly proficient speakers in the typologically similar group showed no difference between the within-language and the between-language Stroop effect. Therefore, these results suggest that when the difference in proficiency levels is considerable (i.e., low proficiency in the non-native language), the effect of typological similarity on language control performance may be limited, potentially because the amount of "training" of the inhibitory skills has not yet been sufficient to elicit any typological similarity effects.

Given the strong link between language control and domain-general inhibitory control (Bialystok et al., 2012; Declerck et al., 2021; Festman et al., 2010), this argument could be applied to our study: our Italian-Spanish and Dutch-Spanish speakers were late language learners of Spanish who had a B1/B2 proficiency level in Spanish. We therefore postulate that the difference in proficiency between the native language (i.e., Italian or Dutch) and the non-native language Spanish was too substantial to elicit a typological similarity effect on inhibitory control performance, even at intermediate B1/B2 proficiency levels. However, we anticipate that with increasing non-native proficiency levels, a typological similarity effect on inhibitory control may be more pronounced. In view of this, it may not be surprising that inhibitory control performance (i.e., the size of the Stroop effect) was statistically equal given that our groups had highly comparable proficiency levels in their non-native language Spanish. Thus, while our findings are not fully compatible with the CRM framework proposed in the introduction (Stocco et al., 2014), they suggest that at intermediate non-native proficiency levels, the modulating role of typological similarity is not yet traceable at the behavioural level.

A second interpretation of our results could be that management of cross-language interference between two typologically similar languages does not directly transfer to strengthening the networks underlying inhibitory control. While we know that speaking multiple languages has a direct impact on language control (Coderre et al., 2013; Coderre & Van Heuven, 2014; D. W. Green, 1998; D. W. Green & Abutalebi, 2013; Mosca & De Bot, 2017), this may not generalise to broader executive functions such as inhibitory control. Contrary to the predictions by the CRM (Stocco et al., 2014), it may be the case that speaking typologically similar languages does not result in a quantitative difference in the amount of training of executive functions over time compared to typologically dissimilar languages. Therefore, the link between speaking typologically similar languages, language control and inhibitory control needs to be more closely inspected in future studies, specifically, the association between language control and inhibitory control.

Considering our compelling findings, the current study takes an important step towards understanding the relative contribution of typological similarity to inhibitory control performance. Taken together, our results suggest that typological similarity only plays a limited role in modulating inhibitory control performance, already at the stage when there is a moderate difference in proficiency levels between the native and the non-native language. However, typological similarity may start to play a role only when non-native proficiency becomes more native-like. Second, our findings further suggest a more complex link between managing multiple languages and more general inhibitory control skills. This could imply that multilingualism primarily influences language control, but that it has only limited effect on domain-general inhibitory control mechanisms. Therefore, our results have important implications for the conceptualisation of the underlying processes of inhibitory control and add novel evidence to the debate around the role of typological similarity in inhibitory control performance.

7.4.1 Conclusions

In this study, we used a spatial Stroop task to examine whether and how inhibitory control performance measured via the Stroop effect was modulated by typological similarity. We found that the typologically dissimilar (Dutch-Spanish) group was faster in performing the task compared to the typologically similar (Italian-Spanish) group. This implied that the Dutch-Spanish group was better at monitoring goal-related information throughout the task compared to the Italian-Spanish group. Critically, this did not impact the overall Stroop task performance. Instead, the size of the Stroop effect, and in turn inhibitory control performance, were similar across both groups, irrespective of typological similarity. Therefore, our results suggest that typological similarity plays a limited role in modulating inhibitory control performance, particularly in intermediate proficient multilinguals with considerable differences in proficiency between their L1 and non-native language(s).

7.4.2 Future directions

Our findings open new avenues to expand on current theoretical frameworks describing the impact of typological similarity on inhibitory control. An emerging line of research could focus on quantifying the degree of interference between typologically similar vs. dissimilar languages and the consequences for language control and/or inhibitory control. For this, future studies should investigate first, language pairs with varying degrees of typological similarity, second, include separate measures for both language control and inhibitory control performance and, finally, recruit speakers of different proficiency levels to tease apart the potentially critical effects of proficiency in modulating inhibitory control performance. Recent years have also seen an increase in research on the neuro-cognition of inhibitory control which combines behavioural measures with electrophysiological and neuroimaging methods (Abutalebi et al., 2012; Christoffels et al., 2007; Constantinidis & Luna, 2019; Grundy, Anderson & Bialystok, 2017). Future studies in this area of research should also incorporate both offline and online measures such as electroencephalography or fMRI measures to model the cognitive and neural mechanisms underlying inhibitory control performance in multilingual language processing.

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. **Anna Gupta**: Conceptualisation, Investigation, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review & Editing. **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 Núria Sebastián-Gallés and the CBC laboratory team from Pompeu Fabra University for providing the facilities for the data collection. A special thanks goes to Philipp Flieger and Ifeoluwa Olusayo Oloruntuyi for their support during the data collection at Leiden University and their feedback on the first draft. We also thank Michal Korenar for his helpful comments. Finally, we would like to thank our anonymous reviewers and all our participants for their time.

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/e5ba9/?view_only=e8c7dbdf07984cbebb0fce50a132a605`

Citation diversity statement

Within academia, research is witnessing a systematic underrepresentation of female researchers and members of minorities in published articles (Dworkin et al., 2020; Rust & Mehrpour, 2020; Torres et al., 2020; Zurn et al., 2020). With this Citation Diversity Statement, we aim to raise awareness about this issue. We classified the first and last author based on their preferred gender for each reference in our reference list (wherever this information was available). Our reference list contained 25% woman/woman authors, 42% man/man, 13% woman/man and finally, 16% man/woman authors. For lack of direct comparison in the psycholinguistic field, we compared this 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 (Dworkin et al., 2020). Note that the limitations of this classification are twofold: first, we need to develop adequate comparison metrics for every research field; and second, we need to improve this rudimentary binary gender classification system. However, we are confident that future work will address both issues.

Appendix

7.A Linguistic profile: Italian-Spanish group

Table 7.A.1: *Linguistic profile of the Italian-Spanish group (N = 33) according to the LEAP-Q (Marian et al., 2007).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Italian	n = 33					**33**
Spanish		n = 2	n = 18	n = 10	n = 3	**33**
English		n = 27	n = 5			**32**
French		n = 4	n = 8	n = 3		**15**
German			n = 1	n = 2		**3**
Catalan				n = 1	n = 1	**2**
Portuguese					n = 3	**3**
Total	**33**	**33**	**32**	**16**	**7**	

7.B Linguistic profile: Dutch-Spanish group

Table 7.B.1: *Linguistic profile of the Dutch-Spanish group (N = 25) according to the LEAP-Q (Marian et al., 2007).*

	L1	**L2**	**L3**	**L4**	**L5**	**Total**
Dutch	n = 25					**25**
Spanish			n = 9	n = 9	n = 7	**25**
English		n = 23	n = 2			**25**
German			n = 7	n = 7		**14**
French		n = 1	n = 7			**8**
Portuguese				n = 2	n = 2	**4**
Frisian		n = 1				**1**
Japanese				n = 1	n = 1	**2**
Italian					n = 1	**1**
Mandarin					n = 1	**1**
Total	**25**	**25**	**25**	**19**	**12**	

7.C Model parameters: Italian-Spanish group

Table 7.C.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for the Italian-Spanish group (n = 32).*

	Formula: accuracy ∼ condition (congruent vs. incongruent) + (1|subject) + (1|item)			**Formula**: RTs ∼ condition (congruent vs. incongruent) + (condition|subject) + (1|item)		
Term	**Odds ratio [95% CI]**	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	24.86 [16.56 - 37.32]	15.50	< 0.001	579.21 561.94 - 596.48]	65.77	< 0.001
Condition [incongruent]	0.580 [0.400 - 0.784]	-3.38	**0.001**	29.46 [18.10 - 40.82]	5.09	<**0.001**
Random effects						
σ^2	3.29			10783.88		
$\tau_{00 Subject}$	0.72			2102.95		
$\tau_{00 Item}$	0.01			8.82		
$\tau_{11 Subject[incongr.]}$				307.02		
$\rho_{01 Subject[incongr.]}$				-0.28		
ICC	0.18			0.16		
$N_{Subject}$	32			32		
N_{Item}	4			4		
Observations	3072			2856		
Marginal R^2/ Conditional R^2	0.020/0.198			0.017/0.173		

7.D Model parameters: Dutch-Spanish group

Table 7.D.1: *Models of best fit for accuracy and RTs, including odd ratios/estimates, confidence intervals, test statistics and p-values for the Dutch-Spanish group (n = 25).*

	Formula: accuracy ~ condition (congruent vs. incongruent) + (condition|subject)			**Formula**: RTs ~ condition (congruent vs. incongruent) + (condition|subject)		
Term	**Odds ratio [95% CI]**	**z-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	25.65 [17.30, 38.02]	16.16	< 0.001	559.80 [539.98, 579.63]	55.37	< 0.001
Condition [incongruent]	0.477 [0.295, 0.771]	-3.02	**0.003**	15.31 [4.40, 26.21]	2.75	**0.006**
Random effects						
σ^2	3.29			12180.85		
$\tau_{00 Subject}$	0.33			2289.48		
$\tau_{11 Subject[incongr.]}$	0.43			226.95		
$\rho_{01 Subject[incongr.]}$	-0.20			-0.78		
ICC	0.12			0.13		
$N_{Subject}$	25			25		
Observations	2,400			2,236		
Marginal R^2/ Conditional R^2	0.035/0.155			0.004/0.136		

7.E Model parameters: Stroop effect

Table 7.E.1: *Comparison between the model with the interaction effect of condition and typological similarity (left) and the best-fitting model (right) with main effects for condition and typological similarity (n = 57).*

	Formula: RTs $\sim$ typological similarity TS (high vs. low) * condition (congruent vs. incongruent) + (condition|subject) + (1|item)			**Formula**: RTs $\sim$ typological similarity TS (high vs. low) + condition (congruent vs. incongruent) + (condition|subject) + (1|item)		
Term	**Estimate [95% CI]**	**t-value**	**p-value**	**Estimate [95% CI]**	**t-value**	**p-value**
(Intercept)	559.79 [540.73, 578.85]	57.57	< 0.001	553.48 [534.99, 571.96]	58.68	< 0.001
TS [high]	19.44 [-5.93, 44.82]	1.50	0.133	30.70 [7.72, 53.68]	2.62	**0.009**
Condition [incongruent]	15.33 [4.36, 26.29]	2.74	**0.006**	23.20 [14.56, 31.83]	5.27	<**0.001**
TS [high] * Condition [incongruent]	13.98 [-0.396, 28.36]	1.91	0.057			
Random effects						
σ^2	11399.52			11397.84		
$\tau_{00Subject}$	2101.67			2210.02		
τ_{00Item}	1.10			5.03		
$\tau_{11Subject[incongr.]}$	243.08			307.02		
$\rho_{01Subject[incongr.]}$	-0.50			-0.50		
ICC	0.14			0.15		
$N_{Subject}$	57			57		
N_{Item}	4			4		
Observations	5,092			5,092		
Marginal R^2/ Cond. R^2	0.023/0.161			0.027/0.170		

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* $\underline{\textit{*den}_m\ \textit{Land}_n}$ *auf einem kräftigen Kamel.*

[She travels $\underline{\text{*the}_m\ \text{land}_n}$ on a strong camel]

[2] *Sie bereist* $\underline{\textit{das}_n\ \textit{Land}_n}$ *auf einem kräftigen Kamel.*

[She travels $\underline{\text{the}_n\ \text{land}_n}$ on a strong camel]

[3] *Le olive farcite con* $\underline{\textit{*la}_f\ \textit{peperone}_m}$ *sono ottime.*

[The olives stuffed with $\underline{\text{*the}_f}$ $\underline{\text{bell pepper}_m}$ are excellent]

[4] *Le olive farcite con* $\underline{\textit{il}_m\ \textit{peperone}_m}$ *sono ottime.*

[The olives stuffed with $\underline{\text{the}_m}$ $\underline{\text{bell pepper}_m}$ are excellent]

[5] $\underline{\textit{*La}_f\ \textit{piano}_m}$ *estaba viejo y desafinado.*

[$\underline{\text{*The}_f\ \text{piano}_m}$ was old and off-key]

[6] $\underline{\textit{El}_m\ \textit{piano}_m}$ *estaba viejo y desafinado.*

[$\underline{\text{The}_m \text{ piano}_m}$ 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 *dog”_, 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

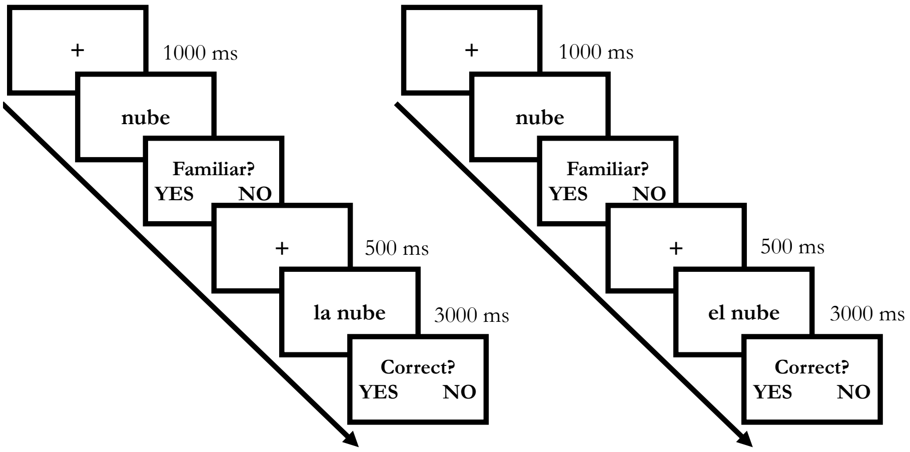
Participants were asked to make a lexical decision whether the letter string displayed on the screen corresponded to a Spanish word or a pseudoword, identical to what was described in Von Grebmer Zu Wolfsthurn et al. (2021a). Following the display of a fixation cross for 500 ms, the letter string remained 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.

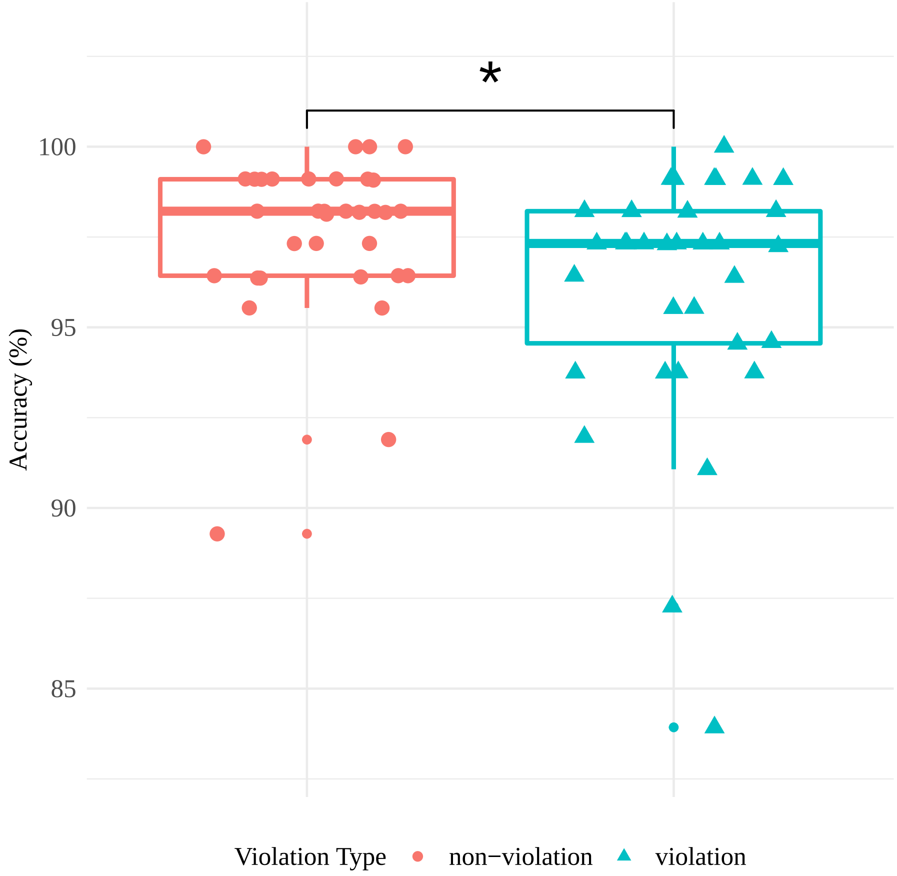
Table 8.3.1: *Mean accuracy and RTs by violation type (n = 34).*

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 β = 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.

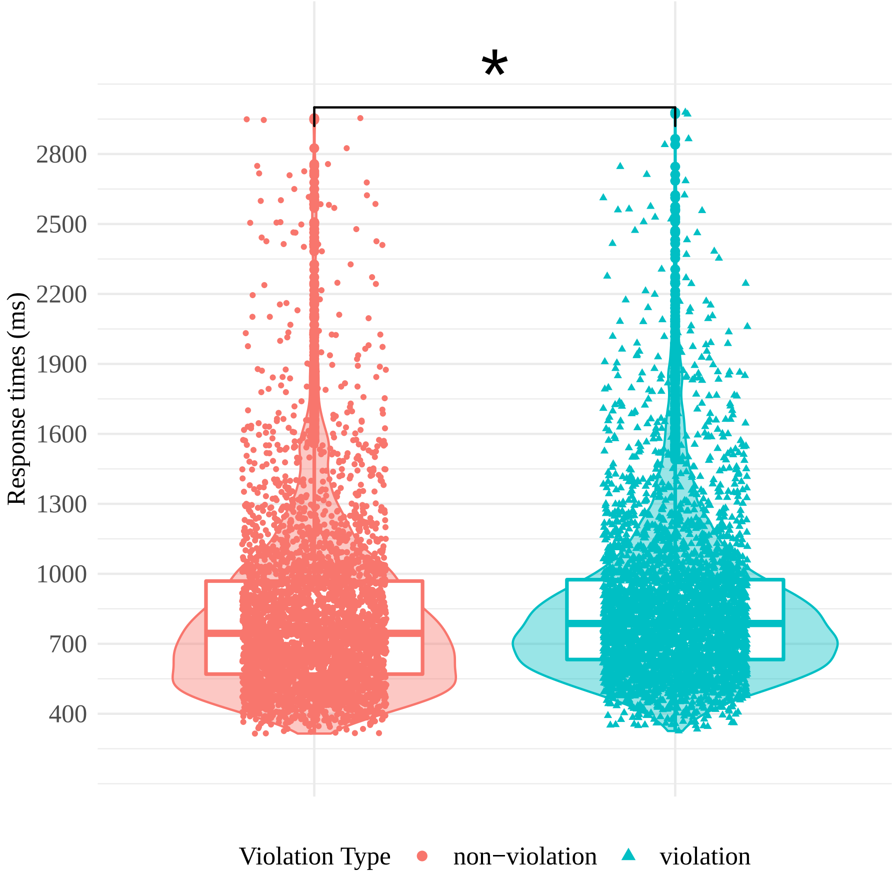
Figure 8.3.1: *Mean accuracy for each violation type (n = 34).*



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

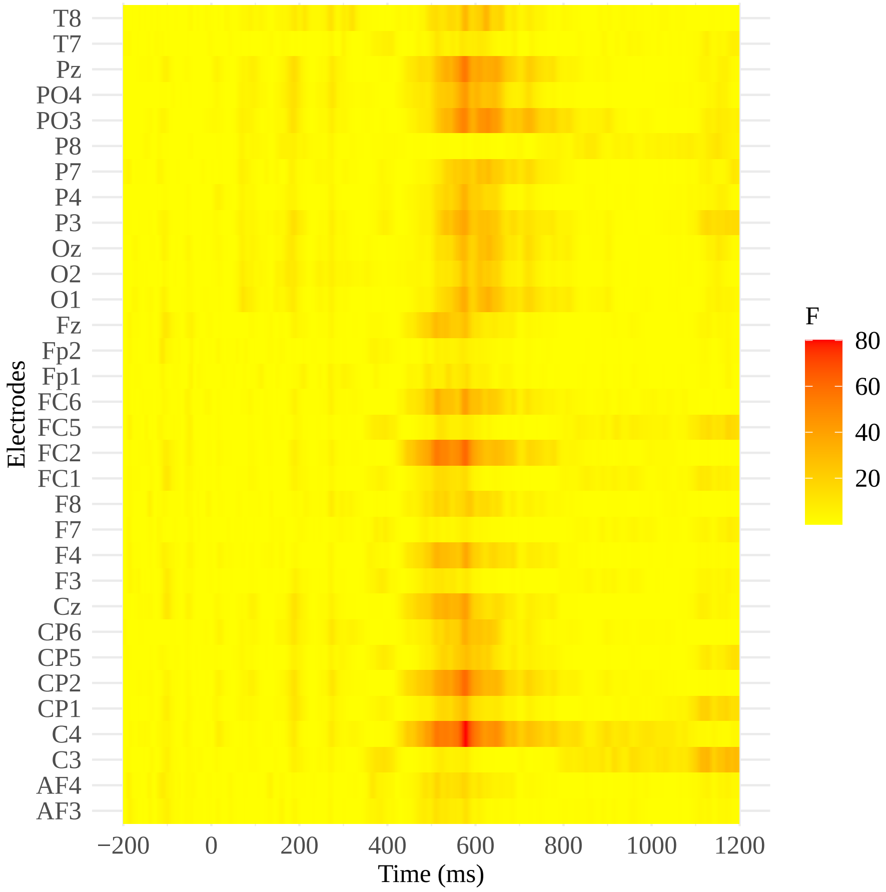
Prior to the statistical analyses of our EEG data, we first performed EEG data pre-processing to increase the signal-to-noise ratio using BrainVision Analyzer (Brain Products, GmbH, Munich). First, we re-referenced the signal from the implicit reference channel to the average of the two mastoid channels. 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.*

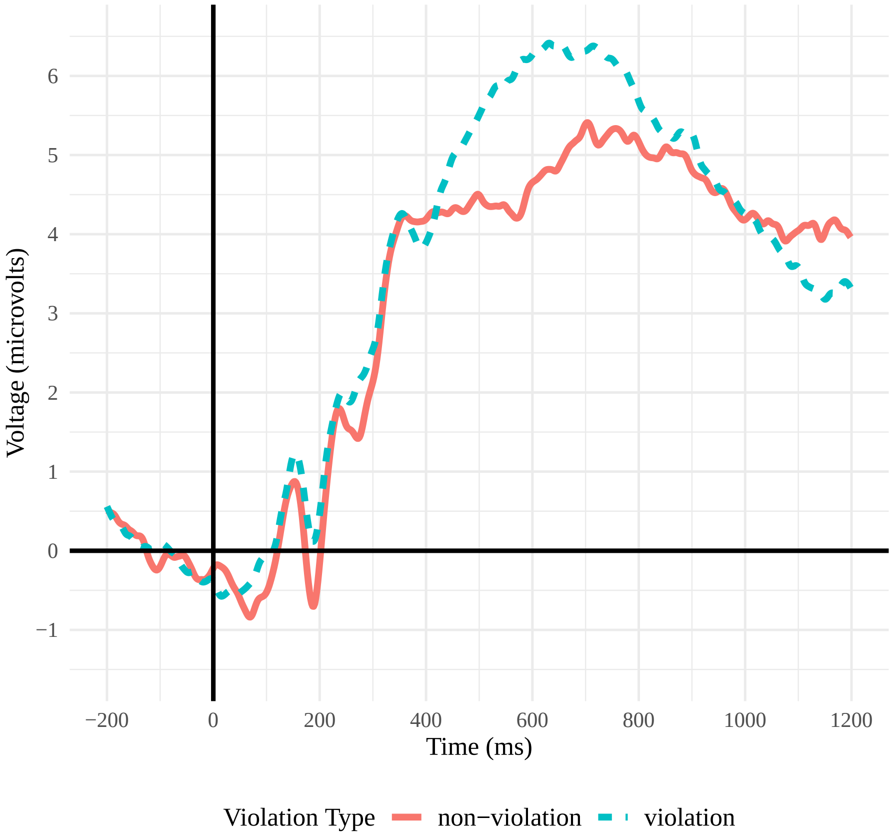
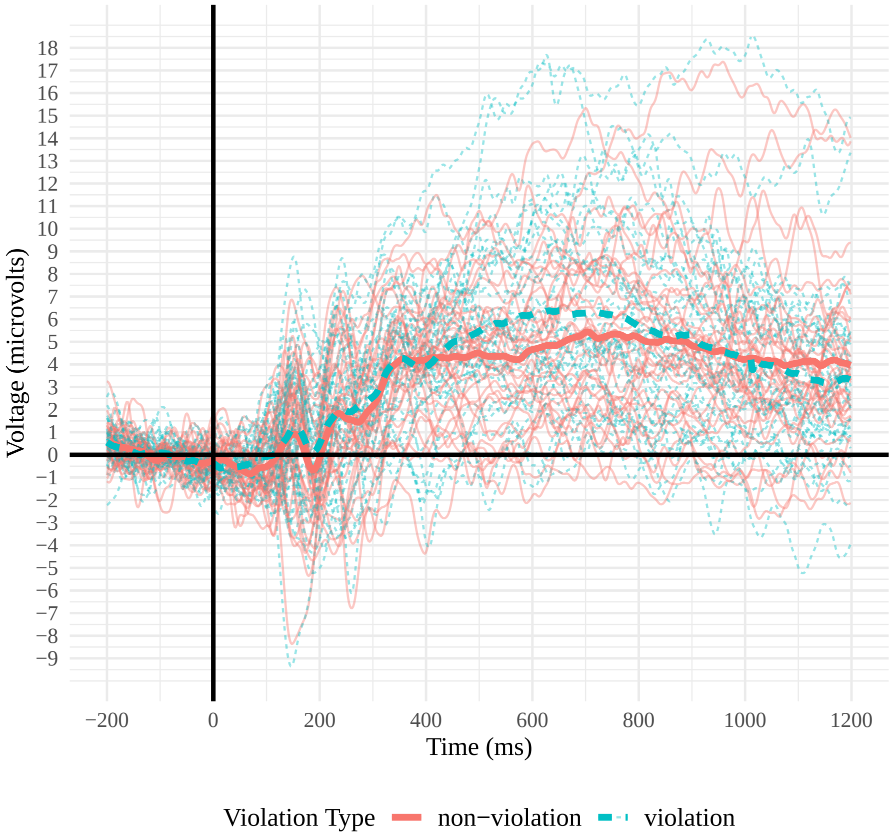
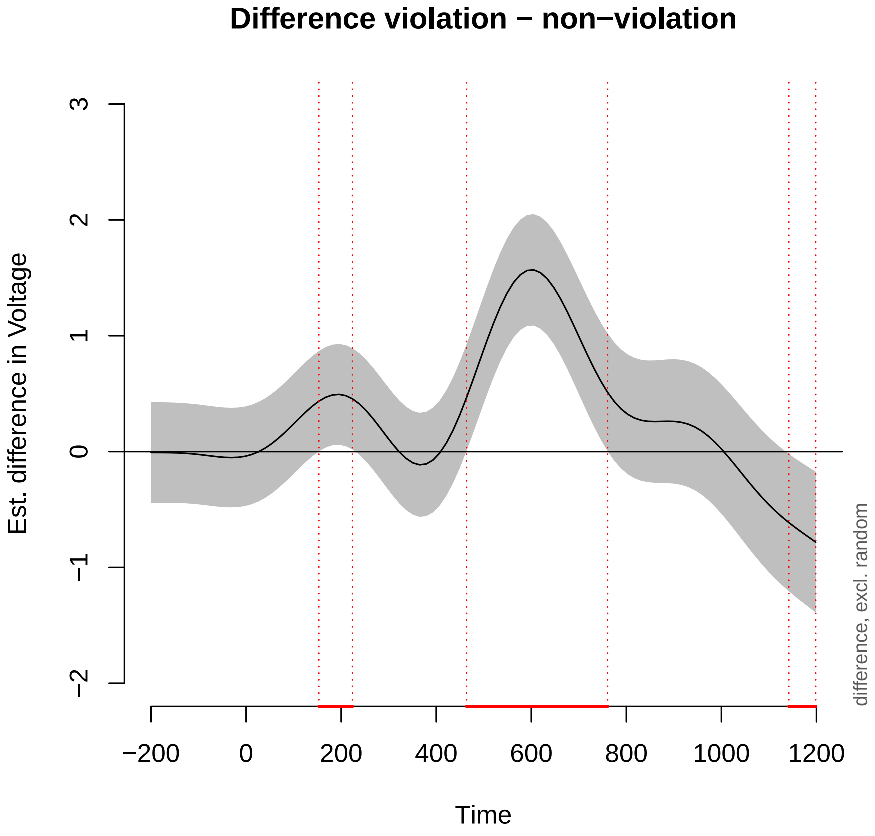


Figure 8.3.5: *Mean voltage amplitudes by violation type in centro-parietal regions (CP1, CP2, CP5, CP6, Pz, P3, P4, PO3, PO4) for n = 34. The thicker lines represent average voltage amplitudes by violation type, whereas the finer lines represent by-violation type averages for each participant.*



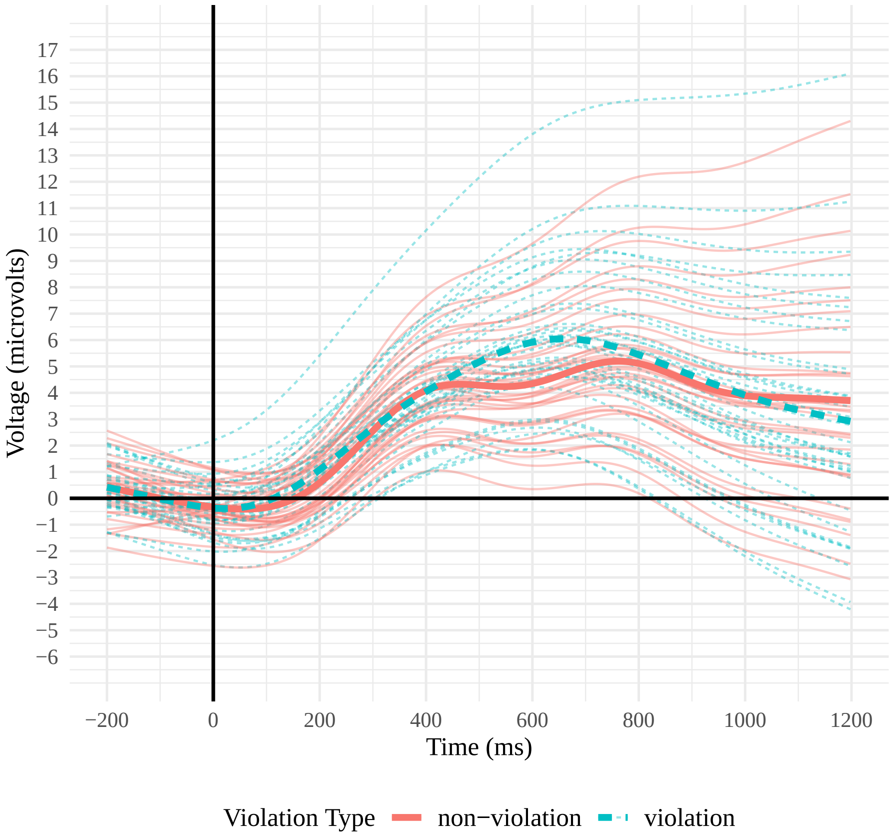
Model parameters for linear and non-linear (smooth) terms are reported in Appendix 8.F. Critically, the model yielded a significant interaction effect of *violation type* and *time* with F = 3489.47, p < 0.001 for non-violation trials compared to violation trials. This indicated a statistical difference 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.*

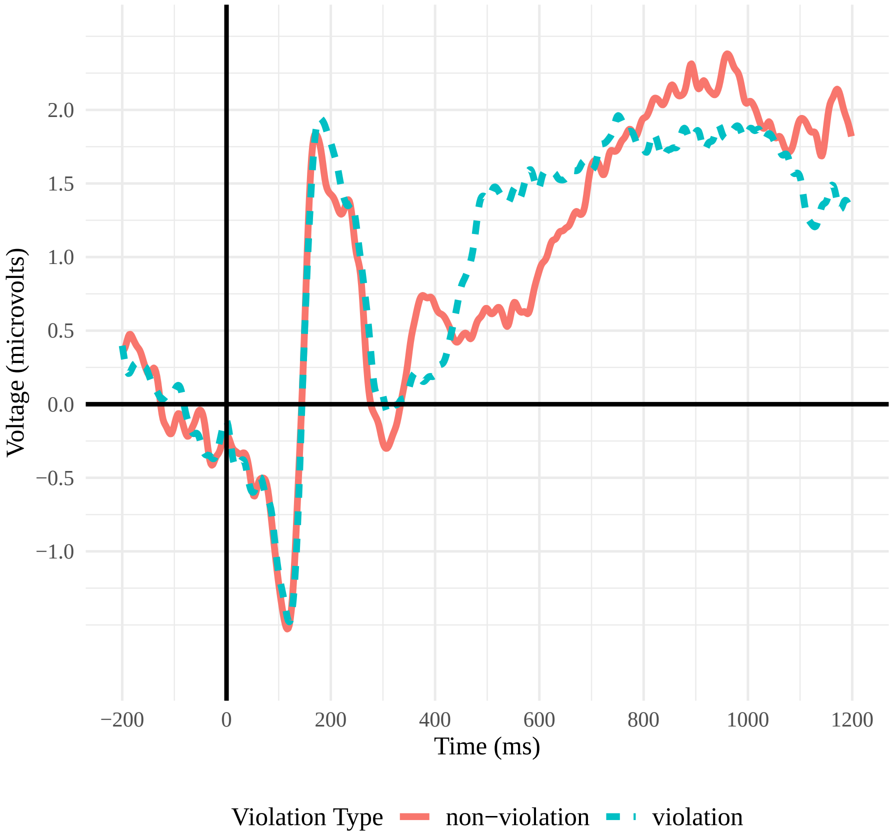
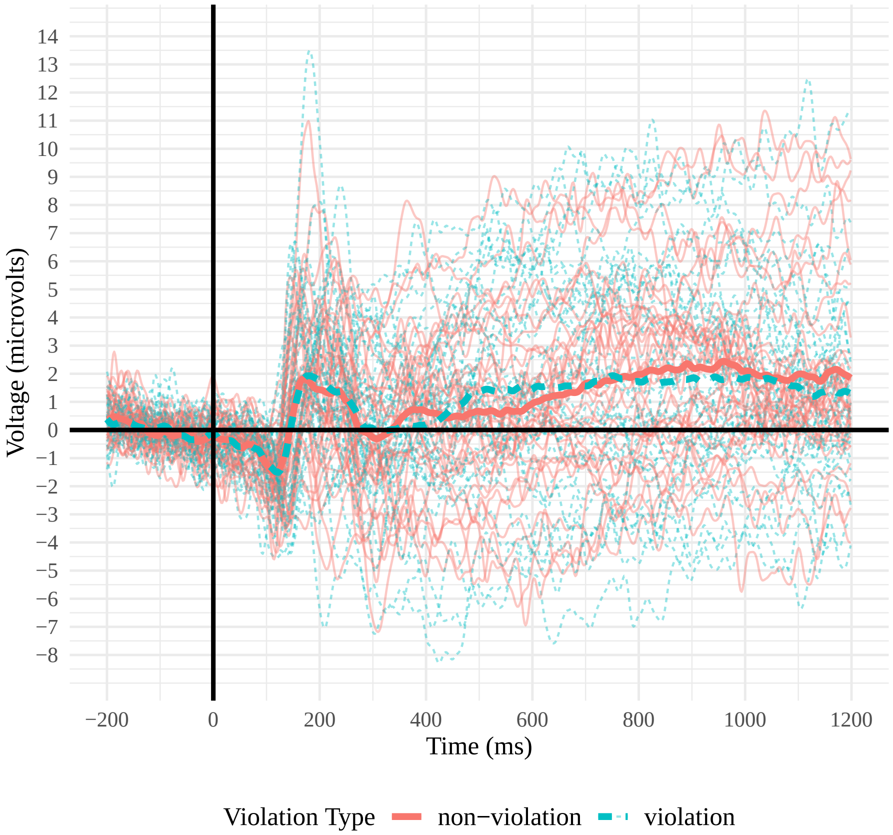
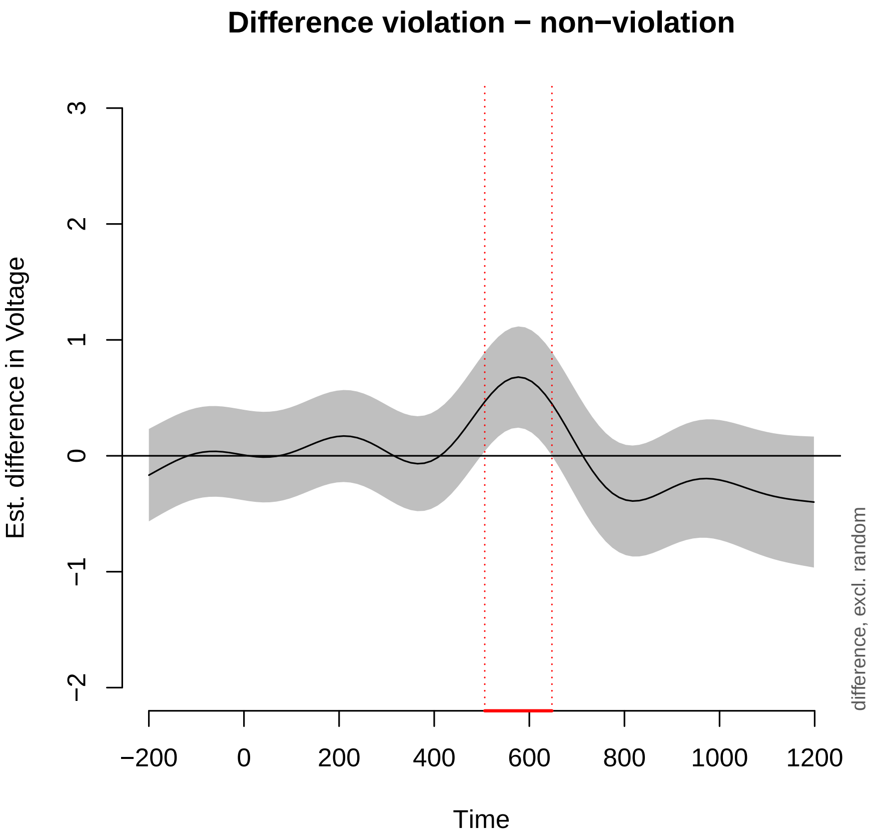


Figure 8.3.9: *Mean voltage amplitudes by violation type for left anterior regions (Fp1, AF3, F3, F7) for n = 34. The solid line represents average voltage amplitudes by violation type, whereas the finer lines represent by-violation type averages for each participant.*



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.

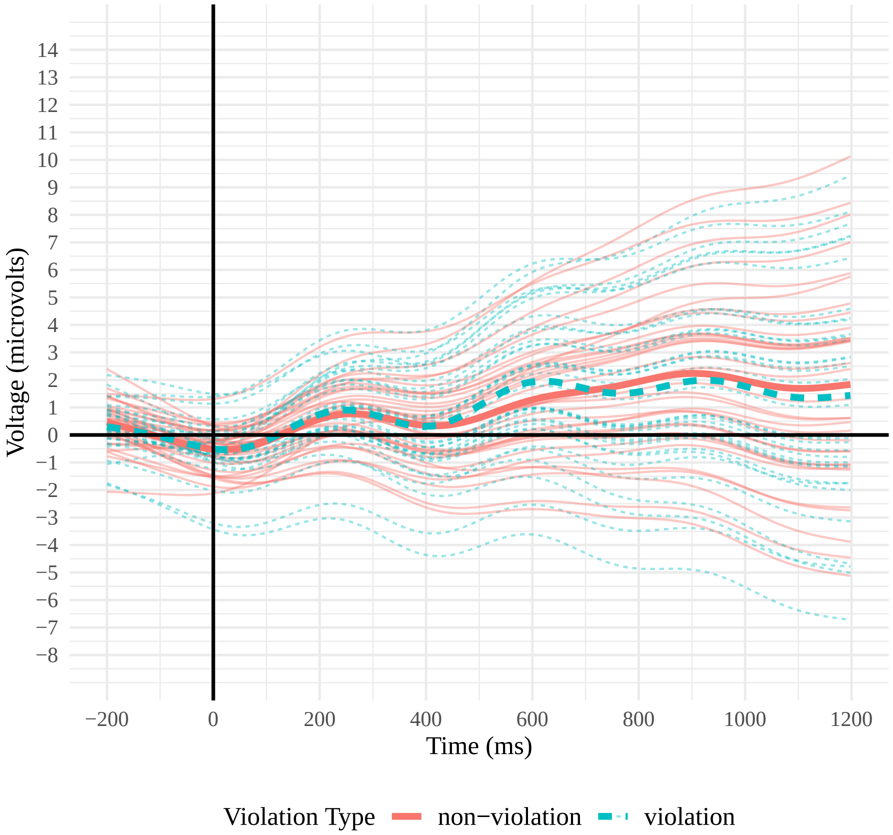
Figure 8.3.10: *Predicted differences in voltage amplitudes (μV) for violation and non-violation trials for channels Fp1, AF3, F3 and F7 (n = 34.). Note that random effects are excluded here.*



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

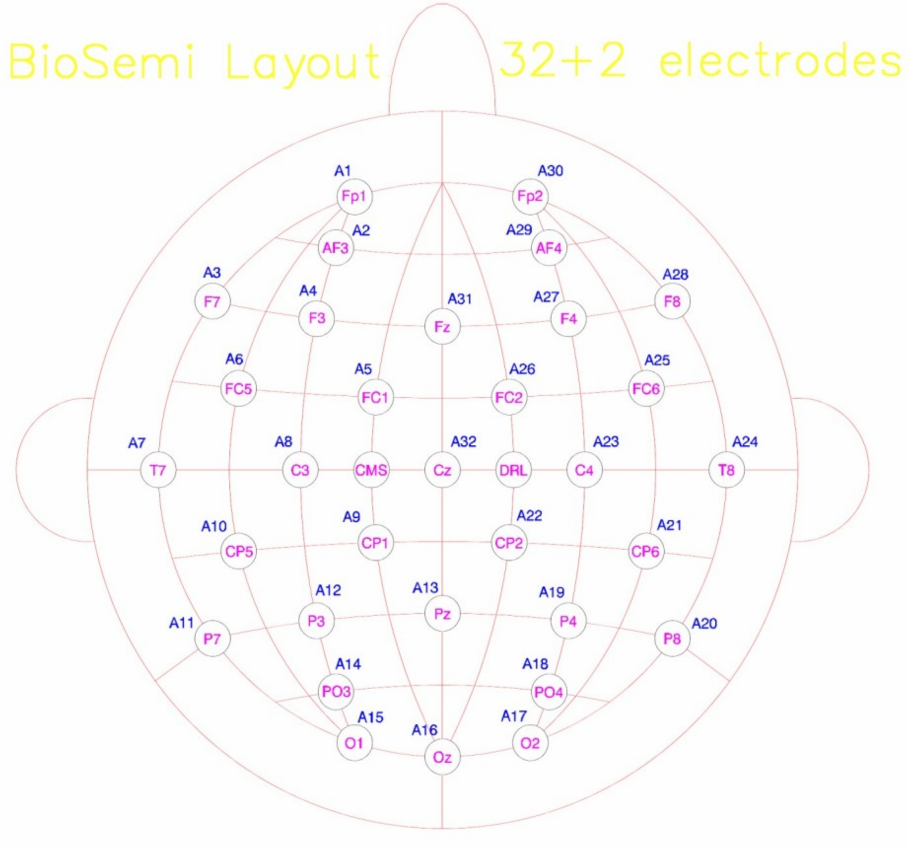
8.A Linguistic profile of participants

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 $\sim$ 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		
τ_{00Item}	0.27		
$\tau_{00Participant}$	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

CHAPTER 9

General discussion

The overarching question of this thesis was concerned with how the brain simultaneously manages both the native language (L1) and an additional non-native language. We focused on characterising this multilingual experience in the context of late language learners, i.e., individuals who acquired a non-native language later in development after fourteen years of age. Across several experiments, we dug deep into the behavioural and neural correlates of (non-)native comprehension and non-native production in order to obtain a nuanced picture of the underlying processing mechanisms. Specifically, we studied six critical issues in more depth: first, we quantified cross-linguistic influence (CLI) in non-native comprehension and non-native production. Second, we examined CLI from a neural perspective. Third, we explored native speakers and late language learners and their sensitivity to "faulty" linguistic input. Fourth, we studied the individual non-native production stages and the locus of target language selection in the context of CLI. Fifth, we investigated the role of language similarity between the L1 and the non-native language on non-native comprehension and production. Finally, we examined the modulating impact of language similarity on domain-general inhibitory control.

In this chapter, we first recapitulate the main research questions and critical findings from each chapter and discuss each chapter's relevance and theoretical contributions to the current literature. Next, we synthesise the research findings to provide a broader picture and an outlook on future research. Moreover, we provide an overview of some unanswered questions in our work. In this, we also touch upon the relevance of this thesis in terms of methodological and statistical advances as well as Open Science. We conclude this chapter with the discussion of future steps from here, and with potential limitations of this thesis.

9.1 Discussion of chapter findings

In **Chapter 2** of this thesis, we closely studied the role of gender congruency and cognate status during syntactic violation processing in the non-native language Spanish. We probed the presence of a gender congruency effect and a cognate facilitation effect. Further, we investigated whether these two linguistic features had a joint effect on syntactic violation processing. In addition, we characterised their influence on the neural correlate of syntactic violation processing, the P600 component. Studies have robustly reported a P600 effect as an index for syntactic violation processing and sensitivity to syntactic irregularities in highly proficient speakers (Foucart & Frenck-Mestre, 2011). Thus far, a large portion of research has focused on highly proficient non-native speakers. Contrastingly, studies on CLI including late language learners with lower proficiency levels are scarce and reported contradictory findings with respect to the elicitation of the P600 effect (Hahne, 2001; S. Rossi et al., 2006; Tokowicz & MacWhinney, 2005; Weber-Fox & Neville, 1996). Subsequently, the nature of CLI effects and the P600 component elicited in late language learners with lower proficiency levels remained unclear. These issues were therefore put to test in this chapter.

Our results indicated the following: we provided evidence for a gender congruency effect in our German late learners of Spanish, with a processing advantage for gender congruent items compared to incongruent items. Critically, we found a reverse cognate facilitation effect, with a processing advantage for non-cognates over cognates. In turn, this indicated that gender congruency was the more salient cue to our late language learners during the syntactic violation paradigm compared to similarities at the orthographic and phonological level. Next, we found no evidence for a joint interaction effect of gender congruency and cognate status on behavioural or neural measures of syntactic violation processing. This notion supports the view of gender congruency as the primary modulating feature of non-native comprehension in our study. Moreover, these findings promote the theoretical view of a shared representation of gender between the two languages as opposed to separate gender representations (Bordag & Pechmann, 2007; Lemhöfer et al., 2008; Morales et al., 2016). A further critical finding was the presence of a P600 effect, as reflected in a significant difference in voltage amplitudes between syntactic violations and non-violations. This implied that the P600 effect was not only found at higher proficiency levels, but also at lower proficiency levels such as in the late language learners of our study. In other words, our late language learners were sensitive to syntactic irregularities at the level of gender. This highlighted the critical role of the P600 component in this context. Yet, neither gender congruency nor cognate status modulated P600 effect sizes. Therefore, while late language learners with limited proficiency showed clear sensitivity to syntactic violations, other linguistic features seemed to have a limited effect at the neural level. The findings from this particular chapter push the theoretical boundaries of the current literature for several reasons: first, they suggest that gender congruency, but not cognate status, is a significant modulator of behaviour in non-native comprehension. Moreover, there is no evidence for a traceable joint effect of gender congruency and cognate status on P600 amplitudes. This has implications for the respective salience of these two linguistic features during non-native processing. Second, the P600 effect was confirmed to be linked to syntactic violation processing also in language learners with lower non-native proficiency levels. However, neither gender congruency nor cognate status significantly modulated P600 effect sizes. Finally, the study re-directed the focus of current research of non-native comprehension to the less studied population of late language learners.

In **Chapter 3** of this thesis, we shifted the focus from non-native comprehension to non-native production. In the multilingual literature, it is well-established that speakers are often slower in naming pictures in their non-native language than in their L1 (Hanulová et al., 2011). Yet, the nature of this discrepancy is poorly understood, particularly with respect to the timing of the individual production stages as were described in the LRM model of word production (Levelt et al., 1999). One critical influencing factor we investigated in this study is CLI, which results from the parallel activation of the L1 and the non-native language(s). Speakers must first overcome CLI to select a target language prior to articulation. In turn, this has implications for the temporal unfolding of the underlying production mechanisms (Costa et al., 2009; Hanulová et al., 2011; Strijkers et al., 2010). Research on this issue in multilingual populations and in late language learners is scarce (Costa et al., 2009; Hoshino & Thierry, 2011; Strijkers et al., 2010), but some studies have provided a testable theoretical framework to examine the time course of non-native production in more detail (Bürki & Laganaro, 2014; Indefrey, 2011; Levelt et al., 1999). Another related issue is the locus of target language selection, which is characterised by two contrasting theoretical accounts on target language selection: one account claiming that lexical entries from both languages are activated, but only the lexical entry from the target language is selected (Gollan et al., 2005; Lee & Williams, 2001). The second account suggests that lexical entries from both languages are selected and subject to subsequent phonological encoding (Christoffels et al., 2007; Colomé, 2001; Hoshino & Thierry, 2011; Rodriguez-Fornells et al., 2005). Subsequently, the main goals of the study were the following: first, to explore the effects of CLI on the time course of non-native production in late language learners (German-Spanish speakers) from a behavioural and neural perspective. More specifically, we exploited both the gender congruency effect and the cognate facilitation effect to trace CLI during the individual production stages. Next, we probed the locus of target language selection to get an insight into when during non-native production speakers select the target language over the non-target language. Finally, we placed a special focus on the P300 component, which is typically linked to cognitive control, but more recently also to working memory load (Barker & Bialystok, 2019; González Alonso et al., 2020; Polich, 2007). We predicted that the P300 component would be a critical index for CLI and resolution of CLI in order to succeed at a non-native production task such as a picture-naming task. Taken together, our main questions of this chapter were the following: first, how can we use CLI to characterise the modulation of the non-native production processes? Second, when is the target language selected during non-native production?

Our findings were as follows: with respect to our first research question, we found the classical gender congruency effect at the behavioural level, suggesting CLI between the gender systems in German-Spanish speakers. In contrast, we found no behavioural effect of cognate status on non-native production. This suggested that gender congruency, but not cognate status significantly altered the time course of non-native production. In turn, this resulted in a measurable delay in producing gender incongruent items compared to congruent items. As for our second research question, we found evidence for a P300 component. We interpreted this as an index for the mitigation of CLI to select the appropriate target language. In addition, there was a small modulatory effect of CLI of the P300 component amplitudes. This suggested that German-Spanish speakers still faced CLI during the phonological encoding stage of the LRM model (Indefrey, 2011; Levelt et al., 1999). Subsequently, this implied that CLI continued into advanced production stages, and that the target language had not been selected before lexical retrieval (Christoffels et al., 2007; Colomé, 2001; Hoshino & Thierry, 2011; Rodriguez-Fornells et al., 2005). This particular study therefore contributes novel evidence to both the characterisation of the time course of non-native production, as well as the description of the neural correlates linked to CLI and non-native production in light of the selection of the target language. These results are further relevant for describing the challenges faced by multilingual speakers in early language acquisition stages. However, the results could also be highly relevant for characterising multilingual populations with inherent language production deficits, e.g., patients with primary progressive aphasia (Calabria, Grunden, Serra, García-Sánchez & Costa, 2019; Kuzmina, Goral, Norvik & Weekes, 2019). Obtaining a clearer picture of the exact difficulties encountered by these populations during the individual production stages could be a critical milestone in the investigation of their clinical symptoms.

In **Chapter 4**, the questions we asked were identical to those in Chapters 2 and 3. However, the critical difference was that we investigated these questions in a linguistically highly similar language pair, namely Italian and Spanish. In terms of non-native comprehension, we asked first, whether and how gender congruency and cognate status impacted syntactic violation processing. Second, we investigated whether speakers of highly similar languages would also show a P600 effect comparable to the German-Spanish speakers. Third, we wanted to know whether the P600 effect was influenced by the two linguistic features of gender congruency and cognate status. Our results were remarkably similar to the results of the German-Spanish speakers from Chapter 2: first, we provided evidence for the classical gender congruency effect (Lemhöfer et al., 2008). In turn, this promoted the notion of shared gender systems across Italian and Spanish, which supported the gender-integrated representation hypothesis (Bordag & Pechmann, 2007; Lemhöfer et al., 2008). Moreover, cognate status had a similar reverse effect as in the German-Spanish speakers, with the Italian-Spanish speakers being more accurate for non-cognates compared to cognates. This supported the view of gender congruency emerging as the more salient feature during syntactic violation processing at this specific proficiency level. Crucially, we again provided evidence for a P600 effect in our Italian-Spanish speakers. Therefore, speakers of linguistically highly similar languages also demonstrated sensitivity to syntactic anomalies. Interestingly, and in contrast to the German-Spanish speakers, the results suggested a small modulation of RTs as a function of CLI at the behavioural level. In other words, this suggested that speakers of linguistically similar languages were sensitive to a joint influence of gender congruency and cognate status during syntactic violation processing. We did not find evidence for this notion in speakers of linguistically less similar languages. Therefore, these particular results provided us with a tentative preview of potentially significant differences in non-native comprehension in terms of behavioural and neural patterns as a function of language similarity.

In terms of non-native production, we examined first, how CLI originating from gender congruency and cognate status impacted the time course of non-native production. We placed a special focus on both behavioural measures and P300 component amplitudes. Second, we explored when during the non-native production process our Italian-Spanish speakers faced CLI, and during which production stage the target language was selected. With respect to our first research question, behavioural data suggested a significant modulation of the time course of non-native production as a function of cognate status. This was in contrast to the findings from the German-Spanish speakers in Chapter 3. The emerging picture was that speakers of linguistically less similar languages (German-Spanish) were behaviourally more sensitive to similarities at the level of gender, whereas speakers of linguistically similar languages (Italian-Spanish) were more sensitive to similarities at the level of orthography and phonology, i.e., cognates. Critically, the EEG results for the Italian-Spanish speakers demonstrated that the P300 component was elicited in connection with mitigatory processes linked to CLI. Yet, these results did not show evidence for a modulation of P300 amplitudes as a function of CLI. In turn, this result did not allow for a more nuanced investigation of the locus of target language selection, as per our second research question. Subsequently, the locus of target language selection in speakers of linguistically similar languages remains an open question for future studies. The EEG findings therefore showed a clear contrast to the German-Spanish speakers. Subsequently, the results from this task suggested quantitative differences in terms of behavioural and neural measures of non-native production for a linguistically similar language pair and a linguistically less similar language pair, particularly in terms of the EEG data. For both non-native comprehension and non-native production, our findings brought about the bigger question whether there were indeed measurable differences in non-native comprehension and production processes as a function of language similarity. This question subsequently served as our main foundation in the following two chapters.

Chapter 5 of this thesis was the first chapter to directly tackle the issue about the role of language similarity in non-native comprehension. Here, we asked the fundamental question of whether a higher similarity between the L1 and the non-native language would lead to a measurable processing advantage in our late language learners relative to a lower similarity between languages. Previous studies have reported increased CLI as well as differential ERP effects for linguistically more similar languages compared to less similar languages (Sabourin & Stowe, 2008; Tolentino & Tokowicz, 2011; Zawiszewski & Laka, 2020). More importantly, studies have tentatively suggested a processing advantage for speakers of highly similar languages (Zawiszewski & Laka, 2020). However, research on this issue is limited and required a more in-depth examination. Therefore, in Chapter 5 we compared the performance during the syntactic violation paradigm between the linguistically less similar language pair (German-Spanish) from Chapter 2, and the linguistically highly similar language pair (Italian-Spanish) from Chapter 4. More specifically, we investigated the effect of language similarity on behavioural measures and on voltage amplitudes in the form of the P600 component. Next, we also tested for an effect of language similarity on CLI. Our main research questions were the following: first, whether the P600 effect was larger for the linguistically similar compared to the linguistically more dissimilar group; and second, whether language similarity influenced CLI across groups. These questions were in addition to the question of whether language learners were sensitive to syntactic violations, which we had already established in the previous chapters. On the basis of our theoretical framework, we predicted a processing advantage for the Italian-Spanish speakers compared to the German-Spanish speakers.

Our results clearly demonstrated that language similarity was indeed traceable both at the behavioural and at the neural level. Interestingly, we found that this effect was pointing in different directions: on the one hand, we found that the linguistically less similar group (German-Spanish) yielded an overall behavioural processing advantage in terms of processing speed as well as smaller CLI effects during this task. On the other hand, the EEG results strongly suggested a processing advantage for the linguistically highly similar group (Italian-Spanish) in the form of a larger, more native-like P600 effect. Moreover, voltage amplitudes connected to CLI were also overall larger for the linguistically highly similar group. Taken together, our results from this chapter demonstrated once again the intricate interplay between the L1 and the non-native language. However, while evidence demonstrated a broad modulatory effect of language similarity on non-native comprehension, the exact nature of this effect remained ambiguous. Based on our results, we argued that language similarity may in fact be an accumulation of smaller similarity effects operating at different linguistic levels such as morphosyntactic similarity, orthographic similarity or phonological similarity. Nevertheless, our study provides critical insights into the underlying processing mechanisms across groups with different language constellations and paves the way for future research into this direction. A clear line of future research is the deconstruction of the "language similarity effect" and to perhaps abandon the notion of a *general* language similarity effect. In other words, we propose that the investigation of the language similarity effect may be more fruitful if language similarity was investigated from the perspective of different linguistic domains.

Chapter 6 marked another shift from non-native comprehension to non-native production. The question at the core of this chapter was concerned with whether speakers of linguistically similar languages possessed a production advantage over speakers of less similar languages. Previous theoretical frameworks suggested that increased CLI may result in a significant enhancement of the cognitive control network over time for linguistically similar languages (Stocco et al., 2014; Yamasaki et al., 2018). In turn, this control network is not only pivotal for the successful generation of non-native output in multilinguals, but also for the mitigation of CLI effects (D. W. Green, 1998). Therefore, in this chapter we systematically examined whether speakers of linguistically similar languages (Italian-Spanish) from Chapter 4 had a production advantage compared to the speakers of linguistically less similar languages (German-Spanish) from Chapter 3. In other words, we probed whether speakers of linguistically similar languages had effectively developed their control network to an extent that would result in an advantage in mitigating CLI effects during non-native production. In line with our theoretical framework, we predicted a production advantage for speakers of more similar languages compared to speakers of less similar languages. Moreover, akin to the previous chapters, we studied the P300 component as a potential index for the CLI in non-native production.

The behavioural findings from this study indicated the typical gender congruency effect as well as the cognate facilitation effect in both groups. Therefore, our behavioural results suggested a production advantage for gender congruent and cognate items compared to gender incongruent and non-cognate items. Interestingly, we found no behavioural evidence for an effect of language similarity on CLI or non-native production as a whole. This suggested that mitigatory strategies of CLI were similarly successful across both groups. At the neural level, we found evidence for an ERP correlate consistent with the P300 component. Critically, this P300 component was impacted by CLI: we found that cognates elicited larger P300 component amplitudes compared to non-cognates. Therefore, CLI appeared a significant modulatory factor of non-native production. Contrastingly, we found no evidence that the neural signatures during non-native production differed as a function of language similarity. Therefore, we found a modulation of voltage amplitudes by CLI on the one hand, but no overall difference between the two groups in terms of the P300 component on the other hand. Subsequently, we were unable to provide support for the notion of a production advantage for speakers of linguistically similar languages at the behavioural or at the neural level. In turn, this has implications for the relevance of language similarity during non-native production. However, it also leaves open the question of whether other factors, e.g., non-native proficiency, could have masked any potential language similarity effects. Therefore, a logical follow-up from this work is first, the investigation of the relationship between non-native proficiency levels and language similarity effects; and second, whether language similarity effects are be more pronounced at lower vs. higher proficiency stages.

In **Chapter 7**, we temporarily moved away from non-native comprehension and production and instead focused on higher cognitive functioning. More specifically, we investigated whether language similarity had a direct impact on domain-general inhibitory control performance. Previous research has previously suggested that speakers of highly similar languages may enhance their cognitive control network to a different degree than speakers of less similar languages (Stocco et al., 2014; Yamasaki et al., 2018). In this chapter, we therefore tested whether speakers of linguistically similar languages (Italian-Spanish) had developed superior inhibitory control skills through their prolonged experience with multiple similar languages compared to speakers of linguistically less similar languages (Dutch-Spanish). Our working hypothesis was that these enhanced inhibitory control skills would be reflected at the behavioural level in a simple spatial Stroop task. In line with this theoretical framework, we predicted a processing advantage and a smaller Stroop effect for the linguistically similar language pair (Italian-Spanish) compared to the more dissimilar language pair (Dutch-Spanish). This would be reflective of a better inhibitory control performance.

Our results indicated the classical Stroop effect in both groups. However, we did not find evidence that the Stroop effect was larger for one group compared to the other. In other words, we did not find evidence that inhibitory control performance was different across groups. Therefore, our findings suggested a limited effect of language similarity on inhibitory control performance. Contrasting with our hypothesis, we also found that speakers of linguistically less similar languages (Dutch-Spanish) were overall faster during the task compared to the speakers of linguistically similar languages (Italian-Spanish). Crucially, questions remain as to first, whether the difference in proficiency between the L1 and the non-native language may have an impact; and second, whether these results would be applicable to different non-native proficiency levels. Further, another emerging question is the extent to which language similarity influenced more domain-general cognitive control vs. language-specific control mechanisms. Particularly, managing two highly similar languages may have consequences for the language control network, but the implications for the domain-general control network may be more complex. Nevertheless, our study was one of the first to make a critical contribution in examining language similarity as a potential factor in driving the functional adaptations of the multilingual mind. However, additional research into this topic is needed to obtain a more complete picture on this issue, especially in combination with neuroimaging and electrophysiological methods.

Chapter 8 of our thesis was the only chapter concerned with native language processing as opposed to non-native language processing. Moreover, it is a prime example for the importance of constructive feedback from external researchers during the empirical research cycle: the main incentive for this study emerged from feedback by an anonymous reviewer during the journal submission of Chapter 2. The background of this study was the controversial nature of the neural correlates connected to gender agreement processing in Spanish, in particular in isolated determiner-noun phrases (Barber & Carreiras, 2003, 2005). A connected question was whether different ERP correlates would be elicited for Spanish compared to other languages, and whether processing of agreement violations in isolated structures vs. in context (e.g., in sentences) was supported by inherently different mechanisms. Therefore, we conducted a study to thoroughly examine the underlying neural correlates of gender agreement processing with native speakers of Spanish. We placed a particular focus on the P600 component as the classical index of syntactic violation processing. In addition, we also probed for the elicitation of an N400 effect and a LAN effect, in line with the disparity of results in the current literature. A novel aspect of this study was the combination of conventional ERP paradigms with advanced, data-driven statistical analyses in order to maximise the power of our findings.

Our results clearly indicated the typical P600 effect for gender agreement processing in isolated noun-phrases in native speakers of Spanish. This emphasised the P600 component as the primary index for syntactic violation processing. Contrastingly, we found neither evidence for an N400 effect, nor for a LAN effect. Subsequently, these results neither support the notion of differential neural mechanisms for processing gender agreement in Spanish compared to other languages, nor that there are differences between processing gender agreement violation in isolated noun-phrases compared to noun-phrases embedded in linguistic structures with more context, e.g., sentences. Finally, our advanced statistical analysis proved to be a viable approach to analyse complex EEG data. We therefore put forward a strong recommendation to consider this technique for future studies investigating native and non-native language-related neural phenomena.

9.2 The broader picture

Taking all chapters together, in the current thesis we studied the multilingual experience in several different populations, from German-Spanish speakers and Italian-Spanish speakers to Dutch-Spanish speakers. In addition, we examined native Spanish speakers to obtain a more detailed picture of the neural correlates supporting language comprehension at a more general level. We employed a range of different experimental tasks, among which a syntactic violation paradigm, a picture-naming task, a Stroop task and the LexTALE-Esp, to tackle some of the fundamental issues in multilingual language processing.

With respect to our first critical issue outlined at the beginning of this chapter and in the introduction, we learnt that CLI plays a significant role in non-native comprehension and production. In other words, the late language learners we examined across our studies faced CLI to an extent which was traceable both in the behavioural and in the neural patterns. Further, they were remarkably successful in the mitigation of CLI effects. Different linguistic features representing CLI, such as gender congruency and cognate status, emerge as influential factors in driving CLI in non-native comprehension and production. Interestingly, cognate status was found to be both a facilitatory and a hindering factor during non-native processing, with differential effects found across comprehension and production. This has direct implications for both models of non-native processing as well as non-native acquisition.

For our second critical issue about the neural correlates of CLI, we found no neural evidence for traceable CLI effects during non-native comprehension. In contrast, small CLI effects were found in non-native production. This indicated that CLI may exert differential influences at the neural level as a function of whether the target domain is comprehension or production. In turn, the corresponding findings have implications for the co-existence of two or more languages in the multilingual brain and their functional interplay. The comparison of CLI effects in comprehension vs. production is beyond the scope of this thesis. However, our results suggest a notable asymmetry in terms of CLI effects. This should be subject to future investigations, as it may provide critical insights into the complex relationship between comprehension and production mechanisms.

Speaking directly to the third issue under investigation in this thesis, it emerged throughout this thesis that our language learners were indeed sensitive to syntactic irregularities both at the behavioural and the neural level. Therefore, this sensitivity does not only manifest itself in highly proficient speakers, but also in less proficient speakers such as the late language learners in our studies. This was a robust finding across a linguistically similar language pair (Italian-Spanish) and a linguistically less similar language pair (German-Spanish). This notion has direct implications for the characterisation of language processing from the perspective of the neural architecture in late language learners. Critically, the P600 effect remains a pivotal neural marker in non-native comprehension and can therefore be used as a reliable index in similar studies tackling non-native language processing.

In terms of the fourth critical issue of this thesis about the potential of CLI to directly characterise the non-native production process, we demonstrate that CLI is an ideal lens through which to tackle more nuanced aspects of multilingual language processing. Among these aspects is the time course of non-native production and the subsequent implications for the selection of target language. Research on this topic is particularly limited and challenging from a methodological and electrophysiological perspective. It is nevertheless critical to continue the efforts to quantify all dimensions of the multilingual production mechanisms. From our findings, we learnt that CLI indeed significantly impacts the non-native production process at the neural level in some learners, and that they face CLI until advanced production stages. With this in mind, non-native production is an even more remarkable process because speakers have to continuously overcome the challenges brought about by CLI to succeed at an every-day task such as a simple conversation.

With respect to the fifth issue of the impact of language similarity on non-native comprehension and production, we discovered that language similarity plays distinct roles in non-native comprehension and non-native production. On the one hand, we found clear indications of the importance of language similarity in modulating syntactic violation processing and CLI during comprehension, but we did not find comparable evidence in non-native production. These results strongly suggest differential language similarity effects across the two domains, as briefly touched upon above in the discussion of the second issue. On the other hand, our results also suggest that the language similarity effect may need to be investigated in a more refined manner. More specifically, we argue that different driving forces may be at play for morphosyntactic similarity, for orthographic similarity, for phonological similarity etc.

which dynamically influence the size and direction of an overall language similarity effect. This is a critical aspect which should be subject to future investigations. Nevertheless, given the scarcity of research, our studies on language similarity provide novel insights into the multilingual experience, both from the point of view of the relevance of the native language, and from the point of view of the challenges faced during non-native comprehension and production by our speakers.

Finally, our sixth critical issue of this thesis was whether language similarity had direct consequences for higher cognitive functions such as the cognitive control network. Here, we learnt that speaking two highly similar languages may not directly translate into increased inhibitory control skills at a domain-general cognitive level. Instead, our results favour the working hypothesis that speaking highly similar languages first and foremost trains language-related mechanisms, such as the language control network, or mechanisms related to the mitigation of CLI. The connection to the cognitive control network therefore remains somewhat of a mystery, but it is a fascinating topic that should be examined more closely.

Pooling the evidence presented throughout this thesis, it is undeniable that multilingual language comprehension and production are extraordinarily complex processes. While we contribute novel evidence to guide the characterisation of the multilingual experience of late language learners, there is much more research to be conducted in this direction. The next section provides some concrete ideas on how to translate the insights gained throughout this thesis into tangible future research.

9.3 Where to go from here

As is not uncommon in empirical research, our studies provided as many fascinating novel findings as they brought about new research avenues. Chapter 2 interestingly suggested no interactive effect of two linguistic features previously shown to be subject to

CLI on non-native comprehension. However, this notion of additive facilitatory or obstructing effects needs to be the subject of a more thorough examination. Chapter 3 showed that speakers faced CLI until advanced processing stages. From our results, we were able to derive that CLI occurred *at least* until phonological encoding, however we could not make any claims beyond this stage. Therefore, we were unable to determine during which exact processing stage CLI was resolved, and in turn, when precisely the target language was selected. Subsequently, this issue also needs to be subject to future research. We also argued that this is a highly relevant issue also for clinical populations with distinct language production difficulties. Chapter 4 took the important step to apply the fundamental debates around CLI to a more similar language pair and tentatively tapped into the notion of language similarity in language processing. While the language pairs we investigated in this thesis were appropriate with respect to our research questions, it is also crucial to move away from the more "standard" language combinations typically found in current research. Instead, studies should explicitly target language combinations which show more pronounced differences across linguistic domains, for example English and Mandarin, Portuguese and Makhuwa or any other language combinations with significantly deviating syntactic, orthographic and phonological systems. How do CLI effects manifest themselves in those language combinations? We are convinced that particularly those language combinations will be critical in unveiling a piece of the broader puzzle around the multilingual experience. Chapter 5 directly probed the effect of language similarity in non-native comprehension, but brought about several more general questions. For example, *is* there an overall language similarity effect? This chapter tentatively suggested that the investigation of language similarity effects may be more productive if it was divided into smaller similarity effects. In other words, here we propose that research should individually tackle morphosyntactic, orthographic or phonological similarity effects to get a better overview of the role of language similarity. Another emerging question was the extent to which language proficiency is a driving force behind language similarity effects. For example, would our results be applicable to lower or higher non-native proficiency levels? Beyond the theoretical contributions, this chapter also explicitly encourages an alternative analysis approach for EEG in the form of generalised additive mixed models (GAMMs). Critically, we do not claim that this approach is suitable for all datasets and designs. However, we strongly encourage researchers to consider this approach for similar studies. In Chapter 6, we failed to show a modulation of the behavioural and neural correlates as a function of language similarity in non-native production. Therefore, this issue as a whole warrants a more in-depth investigation, as these results do not preclude a potentially more subtle language similarity effect. In addition, this chapter called for a thorough comparison of language similarity effects in non-native comprehension and production, as there appeared to be a disparity across the two domains with respect to the modulating power of language similarity. More specifically, does language similarity rely on different mechanisms if the linguistic domain is comprehension vs. production? Chapter 7 introduced the important question about the relationship between language similarity and the potential enhancement of cognitive control networks as a function of high similarity between languages. Here, questions remain as to whether prolonged experience with highly similar languages directly translates to measurable changes in the cognitive control network. Given that the study described in this chapter was a behavioural study, we propose that the combination with neuroimaging or EEG would provide additional critical insights into this matter. Moreover, we argue that a crucial factor to consider here is again non-native proficiency, as it was shown to be intimately linked to language control mechanisms. Finally, Chapter 8 settles some of the controversy surrounding the neural correlates of gender agreement processing in Spanish. Yet, questions remain as to whether ERP correlates can in fact be distinct for different languages, and whether context is a sufficiently salient factor to elicit distinct ERP responses. Similar to Chapter 5, this chapter was also characterised by the successful application of GAMMs for the EEG data. Subsequently, this shows the flexibility of this specific analysis approach and its potential to be applied to different research designs and datasets.

9.3.1 From ANOVAs to GAMMs

As outlined throughout this thesis, we used a range of experimental paradigms and statistical approaches in order to obtain a more complete picture of the critical issues in question. Importantly, we placed a special emphasis on the appropriate selection of statistical methods for a particular experimental design and dataset. With this in mind, a clear development in terms of statistical approaches can be seen throughout this thesis, in particular for the EEG analysis: in Chapters 2, 3, 4 and 6, we used linear mixed effects models, or LMMs (Baayen et al., 2008) to examine voltage amplitudes in pre-determined regions of interests and time windows of interest. As outlined in the introduction, this is a powerful approach because it allows for the analysis of unbalanced datasets, i.e., dataset with a diverging number of observations per participant or experimental condition. In addition, LMMs take into account differential effects for participants and individual experimental items without the need to calculate by-condition voltage grand averages (Frömer et al., 2018). This represents a major improvement from more conventional ANOVA-type analyses, which are still routinely featured in the most recent EEG literature on non-native language processing (Antúnez et al., 2021; Y. Cheng, Cunnings, Miller & Rothman, 2021; Pereira Soares, Kupisch & Rothman, 2022).

The EEG analyses reported in Chapters 5 and 8 represented a further advancement in terms of the statistical modelling of EEG data. Here, we transitioned towards models which extend the LMM framework, namely generalised additive mixed models, or GAMMs (Meulman et al., 2015; Tremblay & Newman, 2015). As previously described, GAMMs allow for the inclusion of (penalised) non-linear terms to flexibly model the oscillatory trend of voltage amplitudes. These non-linear terms are a collection of several functions to capture the full complexity of voltage amplitudes over time (De Cat et al., 2015; Meulman et al., 2015). A critical advantage of this particular approach is that researchers are not bound by a priori predictions about the time window of interest for a given ERP effect. Instead, this approach uses the observed and fitted data to determine the time window of the ERP effect. Therefore, GAMMs represent a robust and powerful approach to model complex EEG data. Subsequently, they may be particularly suited for less studied multilingual populations where ERPs effects are likely to take on more unconventional or novel topographic characteristics. We argue that conventional methods such as ANOVA-type analyses are generally less powerful in these situations to capture the true effects within the data.

Within the scope of this thesis, we emphasise that both LMMs and GAMMs are suitable alternatives to more conventional by-condition averaging approaches. Naturally, we urge caution in the application of these methods, because they may not be appropriate for all experimental designs and datasets. Furthermore, detailed knowledge about both the assumptions of these approaches and their implementation in statistical software are needed to prevent accidental misuse. At this point, we would like to highlight that the current status quo is that statistical advice is often sought after the completion of the data collection and the data analysis. This is reflected in a famous quote by R. A. Fisher: "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of." (Fisher, 1938). Therefore, we propose that the collaboration with statistical consultants should be routinely integrated into the empirical research cycle: researchers and statistical consultants should work together from the beginning of an experimental study to devise the most optimal design and analysis plan given a particular research question and available resources. We argue that collaborations between researchers and statistical consultants would represent a major development in the field of neurolinguistics and all its connected fields in terms of the overall research quality and the credibility of science as a whole. We go one step further and argue that this could be at least a partial mitigatory strategy for the (mis)use of statistical techniques, which was outlined as one problematic aspect of current published research (Sönning & Werner, 2021). Therefore, this notion of close collaboration with statisticians is highly relevant to the ongoing replication crisis and Open Science (Munafò et al., 2017; Sönning & Werner, 2021).

9.3.2 Open Science

Connected to the notion of adopting more suitable statistical approaches and increasing the collaboration with statistical consultants, we also placed a special emphasis on making our work accessible to the scientific community. Therefore, anonymised data from each of our studies as well as the relevant analysis scripts are openly available in the Open Science Framework (Foster & Deardorff, 2017). These data can be freely used by other researchers to validate our results or to pursue other research avenues. Moreover, throughout our chapters we also put a special effort into providing detailed descriptions of our participants, tasks and stimuli selection procedures, design choices and analysis procedures. This is consistent with the core principles of the Open Science movement (Mirowski, 2018; Vicente-Saez & Martinez-Fuentes, 2018). Fundamentally, Open Science should be seen as a critical building stone of future research. Subsequently, the empirical research cycle needs to be embedded within the principles proposed by Open Science for transparent, reproducible, robust and ethical research.

9.4 Limitations

There are some limitations to our work. First and foremost, we acknowledge that our results are likely most applicable to the specific populations tested in our studies. As indicated throughout this thesis, our findings overwhelmingly suggest that multilingualism is a highly complex, multidimensional phenomenon. Research into this field is only beginning to capture the full complexity of the multilingual experience and the corresponding behavioural and neural correlates. Therefore, broader generalisations should only be made once more research has corroborated our work and the work from previous studies in the literature.

Another limitation of our study is concerned with the more general debate around the quantification of language similarity between different languages (Cavalli-Sforza et al., 1994; Schepens et al., 2012; Van der Slik, 2010). In the introduction, we defined language similarity as the structural morphosyntactic, orthographic word form and phonological word form overlap across languages (Foote, 2009; Rothman & Cabrelli Amaro, 2010). However, we are aware that this definition may not encompass all possible dimensions of similarity between languages. Most critically, it does not include the subjective experience of similarity between languages, as briefly discussed in Ringbom and Jarvis (2009). Their work suggests that language learners differ in their perception of what constitutes a "similar" linguistic construct or a "similar" language compared to their native language. Therefore, more research is needed to develop an objective measure of language similarity. Importantly, such a measure should (at least) include quantitative ways of calculating language similarity, such as phonological similarity, orthographic similarity as well as draw from research on genetic distance. Critically, the measure should also incorporate the subjective experience of language similarity of each individual speaker. We are optimistic that future research will be successful in this respect, especially if joint efforts are made across the fields of neurolinguistics, cognitive neuroscience, language genetics and linguistics.

The third limitation is the evident need for a reliable and objective measure of multilingualism, and/or the multilingual experience. Several methods for quantifying different aspects of multilingualism are currently available, see Marian and Hayakawa (2021) for an overview. One example is the LEAP-Q (Kaushanskaya et al., 2020; Marian et al., 2007), as used in this study. Other examples include the Language History Questionnaire (P. Li, Zhang, Tsai & Puls, 2014), the Language Exposure Assessment Tool for infants and children (DeAnda, Bosch, Poulin, Zesiger & Friend, 2016), or the Language and Social Background Questionnaire, LSBQ (Anderson, Mak, Keyvani Chahi & Bialystok, 2018). We acknowledge that one current challenge lies in the fact that researchers are only beginning to understand the depths of the multilingual experience. Yet, there is a distinct lack for an all-encompassing, standardised measure to capture all aspects of multilingualism in its full complexity. There is promising research attempting to unify different measures and to generate an objective measure for multilingualism, see for example Marian and Hayakawa (2021). Therefore, there are some recent initiatives to devise the ultimate measure of the multilingual experience. With respect to the studies reported here, we are aware that the measures used throughout this thesis may not encompass all aspects of multilingualism. However, we are confident that current research is in the process of solving this particular challenge.

The fourth limitation is connected to the interpretation of absent effects in light of statistical power, specifically of behavioural effects. As is evident throughout this thesis, we have not included a priori power analyses to motivate our sample sizes but instead opted to base them on prior work. We know from previous literature that underpowered studies pose a threat to the scientific advancement in that their results should be interpreted with caution (Brysbaert, 2019; Brysbaert & Stevens, 2018; Westfall, Kenny & Judd, 2014). More recently, there have been concrete efforts to mitigate this long-standing issue, for example by means of detailed accounts on how to perform power analyses with specific experimental designs and statistical analyses in mind. For example, Brysbaert (2019) and Brysbaert and Stevens (2018) describe how to achieve 80% power in balanced designs using t-tests, one-way ANOVAs with three-between group levels, one-way repeated-measures ANOVAs with three levels or two-way repeated-measures ANOVAs. More recently, there has been work on how to perform power analysis on designs with one fixed effect with two levels and two random effects (Westfall et al., 2014), and more complex designs warranting mixed effects modelling with two-level fixed factors and multiple random effects (Brysbaert, 2019). However, when it comes to more complex designs where the analysis entails several fixed factors, interactions between factors and more complex random effects structures, as were presented throughout the current thesis, there is little consensus on how to calculate effect sizes or statistical power. In other words, guidance on calculating statistical power for the designs featuring in this thesis is currently difficult to obtain, in particular with respect to EEG studies. One potential solution is to employ simulations (Brysbaert & Stevens, 2018) using dedicated software such as the R package *simr* (P. Green, MacLeod & Alday, 2022). However, this requires highly specialised skills and is currently not straightforward to implement. Nevertheless, we acknowledge and strongly support the efforts to conduct adequately powered studies in order to build a stable foundation for future studies in the name of science.

The fifth and final limitation is the lack of a standardised approach to pre-processing EEG data, which represents a more general challenge to the fields of psycholinguistics, neurolinguistics and cognitive neuroscience. Current EEG studies on the neural correlates of non-native language processing are characterised by the large variation in terms of EEG recording and data pre-processing (Alday et al., 2017; Barber & Carreiras, 2005; Bürki & Laganaro, 2014; Costa et al., 2009; Eulitz et al., 2000; Foucart & Frenck-Mestre, 2011; Hahne, 2001; Midgley et al., 2011; Molinaro et al., 2008; Paolieri et al., 2020; S. Rossi et al., 2006; Strijkers et al., 2010; Von Grebmer Zu Wolfsthurn et al., 2021b). One connected challenge is the need for individually tailored EEG data pre-processing procedures due to the large differences in participants' EEG within and across studies. Yet, this is not an issue unique to electroencephalographic research: for example, this issue has long been known to researchers using functional magnetic resonance imaging (fMRI), see Waller et al. (2022). Researchers in that field have been explicitly pushing towards the standardisation of fMRI data collection and data analysis approaches (Pauli et al., 2016; Waller et al., 2022). For work involving EEG, there have also been relatively recent attempts at advancing and standardising EEG pre-processing pipelines, see for example Motamedi-Fakhr, Moshrefi-Torbati, Hill, Hill and White (2014), or Rodrigues, Weiß, Hewig and Allen (2021). However, standardised EEG data pre-processing procedures do not exist *yet*. Nevertheless, we are hopeful that the standardisation of the EEG procedures, at least to a certain extent, will be implemented in the near future. Subsequently, this will give researchers increased confidence in their work and will have critical implications for the replication crisis.

Taken together, the core take-away messages from our limitations are the following: we need more research on a larger pool of diverse multilingual populations; we need exhaustive measures of language similarity and the multilingual experience; we need guidelines on conducting power analyses for complex experimental designs; and finally, we are in need of a standardised approach to EEG data pre-processing and analysis. These are complex issues to address, but even small progress on any of these five issues will represent a radical step forward in the field and drive the advancement of science.

References

Abdel Rahman, R. & Sommer, W. (2008). Seeing what we know and understand: How knowledge shapes perception. *Psychonomic Bulletin & Review*, *15*(6), 1055–1063. doi: 10.3758/PBR.15.6.1055

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica*, *128*(3), 466–478. doi: 10.1016/j.actpsy.2008.03.014

Abutalebi, J., Cappa, S. F. & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism: Language and Cognition*, *4*(2), 179–190. doi: 10.1017/S136672890100027X

Abutalebi, J., Della Rosa, P. A., Ding, G., Weekes, B., Costa, A. & Green, D. W. (2013). Language proficiency modulates the engagement of cognitive control areas in multilinguals. *Cortex*, *49*(3), 905–911. doi: 10.1016/j.cortex.2012.08.018

Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., … Costa, A. (2012). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral Cortex*, *22*(9), 2076–2086. doi: 10.1093/cercor/bhr287

Abutalebi, J. & Green, D. W. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*(3), 242–275. doi: 10.1016/j.jneuroling.2006.10.003

Abutalebi, J. & Green, D. W. (2016). Neuroimaging of language control in bilinguals: Neural adaptation and reserve. *Bilingualism: Language and Cognition*, *19*(4), 689–698. doi: 10.1017/S1366728916000225

Acheson, D. J., Ganushchak, L. Y., Christoffels, I. K. & Hagoort, P. (2012). Conflict monitoring in speech production: Physiological evidence from bilingual picture naming. *Brain and Language*, *123*(2), 131–136. doi: 10.1016/j.bandl.2012.08.008

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Alario, F.-X. & Caramazza, A. (2002). The production of determiners: Evidence from French. *Cognition*, *82*(3), 179–223. doi: 10.1016/S0010-0277(01)00158-5

Alday, P. M., Schlesewsky, M. & Bornkessel-Schlesewsky, I. (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: Event-related potentials reflect continuous model updates. *eNeuro*, *4*(6), ENEURO.0311-16.2017. doi: 10.1523/ENEURO.0311-16.2017

Alemán Bañón, J., Fiorentino, R. & Gabriele, A. (2012). The processing of number and gender agreement in Spanish: An event-related potential investigation of the effects of structural distance. *Brain Research*, *1456*, 49–63. doi: 10.1016/j.brainres.2012.03.057

Amengual, M. (2012). Interlingual influence in bilingual speech: Cognate status effect in a continuum of bilingualism. *Bilingualism: Language and Cognition*, *15*(3), 517–530. doi: 10.1017/S1366728911000460

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia*, *49*(5), 970–983. doi: 10 .1016/j.neuropsychologia.2011.01.003

Anderson, J. A. E., Mak, L., Keyvani Chahi, A. & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, *50*(1), 250–263. doi: 10.3758/s13428-017-0867-9

Antúnez, M., Mancini, S., Hernández-Cabrera, J. A., Hoversten, L. J., Barber, H. A. & Carreiras, M. (2021). Cross-linguistic semantic preview benefit in Basque-Spanish bilingual readers: Evidence from fixation-related potentials. *Brain and Language*, *214*, 104905. doi: 10.1016/j.bandl.2020.104905

Aristei, S., Melinger, A. & Abdel Rahman, R. (2011). Electrophysiological chronometry of semantic context effects in language production. *Journal of Cognitive Neuroscience*, *23*(7), 1567–1586. doi: 10.1162/jocn.2010.21474

Baayen, H. R., Davidson, D. & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items.

Journal of Memory and Language, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baayen, H. R., Vasishth, S., Kliegl, R. & Bates, D. M. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. doi: 10.1016/j.jml.2016.11.006

Babou, C. A. & Loporcaro, M. (2016). Noun classes and grammatical gender in Wolof. *Journal of African Languages and Linguistics*, *37*(1), 1–57. doi: 10.1515/jall-2016-0001

Badecker, W., Miozzo, M. & Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender. *Cognition*, *57*(2), 193–216. doi: 10.1016/0010-0277(95)00663-J

Bakos, S., Mehlhase, H., Landerl, K., Bartling, J., Schulte-Körne, G. & Moll, K. (2020). Naming processes in reading and spelling disorders: An electrophysiological investigation. *Clinical Neurophysiology*, *131*(2), 351–360. doi: 10.1016/j.clinph.2019.11.017

Barber, H. A. & Carreiras, M. (2003). Integrating gender and number information in Spanish word pairs: An ERP study. *Cortex*, *39*(3), 465–482. doi: 10.1016/S0010-9452(08)70259-4

Barber, H. A. & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience*, *17*(1), 137–153. doi: 10.1162/0898929052880101

Barker, R. M. & Bialystok, E. (2019). Processing differences between monolingual and bilingual young adults on an emotion n-back task. *Brain and Cognition*, *134*, 29–43. doi: 10.1016/j.bandc.2019.05.004

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 1–2. doi: 10.3389/fpsyg.2013.00328

Barry, R. J., Steiner, G. Z., Blasio, F. M. D., Fogarty, J. S., Karamacoska, D. & MacDonald, B. (2020). Components in the P300: Don't forget the Novelty P3! *Psychophysiology*, *57*(7), e13371. doi: 10.1111/psyp.13371

Bates, D. M., Kliegl, R., Vasishth, S. & Baayen, H. R. (2018).

Parsimonious mixed models. *arXiv:1506.04967 [stat]*.

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

Bates, D. M., Mächler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Krivitsky, P. N. (2020). *Package 'lme4*.

Beatty-Martínez, A. L., Bruni, M. R., Bajo, M. T. & Dussias, P. E. (2021). Brain potentials reveal differential processing of masculine and feminine grammatical gender in native Spanish speakers. *Psychophysiology*, *58*(3), e13737. doi: 10.1111/psyp.13737

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., ... Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children*. *Bilingualism: Language and Cognition*, *15*(3), 616–629. doi: 10.1017/S1366728912000090

Berthele, R. (2021a). The extraordinary ordinary: Re-engineering multilingualism as a natural category. *Language Learning*, *71*(S1), 80–120. doi: 10.1111/lang.12407

Berthele, R. (2021b). Introduction: What's special about multilingualism? *Language Learning*, *71*(S1), 5–11. doi: 10.1111/lang.12436

Bialystok, E. (2010). Bilingualism. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 559–572. doi: 10.1002/wcs.43

Bialystok, E., Craik, F. & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 859–873. doi: 10.1037/0278-7393.34.4.859

Bialystok, E., Craik, F. I., Grady, C., Chau, W., Ishii, R., Gunji, A. & Pantev, C. (2005). Effect of bilingualism on cognitive control in the Simon task: Evidence from MEG. *NeuroImage*, *24*(1), 40–49. doi: 10.1016/j.neuroimage.2004.09.044

Bialystok, E., Craik, F. I. M., Klein, R. & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence From the Simon Task. *Psychology and Aging*, *19*(2), 290–303. doi: 10.1037/0882-7974.19.2.290

Bialystok, E., Craik, F. I. M. & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250. doi: 10.1016/j.tics.2012.03.001

Bialystok, E. & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, *7*(3), 325–339. doi: 10.1111/j.1467-7687.2004.00351.x

Blumenfeld, H. K. & Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes*, *22*(5), 633–660. doi: 10.1080/01690960601000746

Blumenfeld, H. K. & Marian, V. (2013). Parallel language activation and cognitive control during spoken word recognition in bilinguals. *Journal of Cognitive Psychology*, *25*(5), 547–567. doi: 10.1080/20445911.2013.812093

Bonneville, E. F. (2020). *Package "BVAtoR".*

Bordag, D. & Pechmann, T. (2007). Factors influencing L2 gender processing. *Bilingualism: Language and Cognition*, *10*(3), 299–314. doi: 10.1017/S1366728907003082

Bosch, J. E. & Unsworth, S. (2020). Cross-linguistic influence in word order: Effects of age, dominance and surface overlap. *Linguistic Approaches to Bilingualism*, 1–34. doi: 10.1075/lab.18103.bos

Bosma, E., Blom, E., Hoekstra, E. & Versloot, A. (2019). A longitudinal study on the gradual cognate facilitation effect in bilingual children's Frisian receptive vocabulary. *International Journal of Bilingual Education and Bilingualism*, *22*(4), 371–385. doi: 10.1080/13670050.2016.1254152

Bosma, E. & Pablos, L. (2020). Switching direction modulates the engagement of cognitive control in bilingual reading comprehension: An ERP study. *Journal of Neurolinguistics*, *55*, 100894. doi: 10.1016/j.jneuroling.2020.100894

Branzi, F. M., Della Rosa, P. A., Canini, M., Costa, A. & Abutalebi, J. (2016). Language control in bilinguals: Monitoring and response selection. *Cerebral Cortex*, *26*(6), 2367–2380. doi: 10.1093/cercor/bhv052

Brauer, M. (1998). Stroop interference in bilinguals: The role of similarity between the two languages. In A. F. Healy & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (First ed., pp. 317–337). Mahwah, New Jersey, US: Lawrence Erlbaum Associates Publishers.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi: 10.1016/j.tics.2011.12.010

Brenders, P. E., Van Hell, J. G. & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology*, *109*, 383–396. doi: 10.1016/j.jecp.2011.03.012

Broersma, P. & Weenink, D. (2019). *Praat: Doing phonetics by computer.*

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. doi: 10.5334/joc.72

Brysbaert, M. & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 9. doi: 10.5334/joc.10

Bull, W. (1965). *Spanish for teachers: Applied linguistics.* Ronald Press.

Bultena, S., Dijkstra, T. & van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, *67*(6), 1214–1241. doi: 10.1080/17470218.2013.853090

Bürki, A. & Laganaro, M. (2014). Tracking the time course of multi-word noun phrase production with ERPs or on when (and why) cat is faster than the big cat. *Frontiers in Psychology*, *5*, 1–13. doi: 10.3389/fpsyg.2014.00586

Bürki, A., Sadat, J., Dubarry, A.-S. & Alario, F.-X. (2016). Sequential processing during noun phrase production. *Cognition*, *146*, 90–99. doi: 10.1016/j.cognition.2015.09.002

Calabria, M., Grunden, N., Serra, M., García-Sánchez, C. & Costa, A. (2019). Semantic processing in bilingual aphasia: Evidence of language dependency. *Frontiers in Human Neuroscience*, *13*.

Calabria, M., Hernandez, M., Branzi, F. & Costa, A. (2012). Qualitative differences between bilingual language control and executive control: Evidence from task-switching. *Frontiers in Psychology*, *2*. doi: 10.3389/fpsyg.2011.00399

Camen, C., Morand, S. & Laganaro, M. (2010). Re-evaluating the time course of gender and phonological encoding during silent monitoring tasks estimated by ERP: Serial or parallel processing? *Journal of Psycholinguistic Research*, *39*(1), 35–49. doi: 10.1007/s10936-009-9124-4

Cantone, K. F. & Müller, N. (2008). Un nase or una nase? What gender marking within switched DPs reveals about the architecture of the bilingual language faculty. *Lingua*, *118*(6), 810–826. doi: 10.1016/j.lingua.2007.05.007

Cárdenas-Hagan, E., Carlson, C. D. & Pollard-Durodola, S. D. (2007). The cross-linguistic transfer of early literacy skills: The role of initial L1 and L2 skills and language of instruction. *Language, Speech, and Hearing Services in Schools*, *38*(3), 249–259. doi: 10.1044/0161-1461(2007/026)

Casaponsa, A., Antón, E., Pérez, A. & Duñabeitia, J. A. (2015). Foreign language comprehension achievement: Insights from the cognate facilitation effect. *Frontiers in Psychology*, *6*, 1–15. doi: 10.3389/fpsyg.2015.00588

Casaponsa, A. & Duñabeitia, J. A. (2016). Lexical organization of language-ambiguous and language-specific words in bilinguals. *Quarterly Journal of Experimental Psychology*, *69*(3), 589–604. doi: https://doi-org.ezproxy.leidenuniv.nl/10.1080/17470218.2015.1064977

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994). *The history and geography of human genes.* Princeton, NJ: Princeton University Press.

Cenoz, J. (2001). The effect of linguistic distance, L2 status and age on cross-linguistic influence in third language acquisition. In J. Cenoz, B. Hufeisen & U. Jessner (Eds.), *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives* (pp. 8–20). Multilingual Matters. doi:

10.21832/9781853595509-002

Cenoz, J. (2013). Defining multilingualism. *Annual Review of Applied Linguistics*, *33*, 3–18. doi: 10.1017/S026719051300007X

Chen, J., Zhao, Y., Zhaxi, C. & Liu, X. (2020). How does parallel language activation affect switch costs during trilingual language comprehension? *Journal of Cognitive Psychology*, *32*(5-6), 526–542. doi: 10.1080/20445911.2020.1788036

Chen, P., Bobb, S. C., Hoshino, N. & Marian, V. (2017). Neural signatures of language co-activation and control in bilingual spoken word comprehension. *Brain Research*, *1665*, 50–64. doi: 10.1016/j.brainres.2017.03.023

Cheng, X., Schafer, G. & Akyürek, E. G. (2010). Name agreement in picture naming: An ERP study. *International Journal of Psychophysiology*, *76*(3), 130–141. doi: 10.1016/j.ijpsycho.2010.03.003

Cheng, Y., Cunnings, I., Miller, D. & Rothman, J. (2021). Double-number marking matters for both L1 and L2 processing of nonlogical agreement similarity: An ERP investigation. *Studies in Second Language Acquisition*, 1–21. doi: 10.1017/S0272263121000772

Christoffels, I. K., Firk, C. & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research*, *1147*, 192–208. doi: 10.1016/j.brainres.2007.01.137

Clahsen, H. & Felser, C. (2006a). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*, 3–42.

Clahsen, H. & Felser, C. (2006b). How native-like is non-native language processing? *Trends in Cognitive Sciences*, *10*(12), 564–570. doi: 10.1016/j.tics.2006.10.002

Clegg, J. H. (2011). A frequency-based analysis of the norms for Spanish noun gender. *Hispania*, *94*(2), 303–319.

Coderre, E. L. & Van Heuven, W. J. B. (2014). The effect of script similarity on executive control in bilinguals. *Frontiers in Psychology*, *5*, 1–16. doi: 10.3389/fpsyg.2014.01070

Coderre, E. L., Van Heuven, W. J. B. & Conklin, K. (2013). The timing and magnitude of Stroop interference and facilitation in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *16*(2), 420–441. doi: 10.1017/S1366728912000405

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language*, *45*(4), 721–736. doi: 10.1006/jmla.2001.2793

Comesaña, M., Sánchez-Casas, R., Soares, A. P., Pinheiro, A. P., Rauber, A., Frade, S. & Fraga, I. (2012). The interplay of phonology and orthography in visual cognate word recognition: An ERP study. *Neuroscience Letters*, *529*(1), 75–79. doi: 10.1016/j.neulet.2012.09.010

Comesaña, M., Soares, A. P., Ferré, P., Romero, J., Guasch, M. & García-Chico, T. (2014). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The role of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 614–635.

Constantinidis, C. & Luna, B. (2019). Neural substrates of inhibitory control maturation in adolescence. *Trends in Neurosciences*, *42*(9), 604–616. doi: 10.1016/j.tins.2019.07.004

Corbett, G. G. (1991). *Gender*. Cambridge University Press. doi: 10.1017/CBO9781139166119

Cornips, L. & Hulk, A. (2008). Factors of success and failure in the acquisition of grammatical gender in Dutch. *Second Language Research*, *24*(3), 267–295. doi: 10.1177/0267658308090182

Costa, A., Albareda, B. & Santesteban, M. (2008). Assessing the presence of lexical competition across languages: Evidence from the Stroop task. *Bilingualism: Language and Cognition*, *11*(1), 121–131. doi: 10.1017/S1366728907003252

Costa, A., Caramazza, A. & Sebastián-Gallés, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283–1296. doi: 10.1037/0278-7393.26.5.1283

Costa, A., Heij, W. L. & Navarrete, E. (2006). The dynamics of bilingual lexical access. *Bilingualism: Language and Cognition*, *9*(2), 137–151. doi: 10.1017/S1366728906002495

Costa, A., Hernández, M. & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*(1), 59–86. doi: 10.1016/j.cognition.2006.12.013

Costa, A., Kovacic, D., Franck, J. & Caramazza, A. (2003). On the autonomy of the grammatical gender systems of the two languages of a bilingual. *Bilingualism: Language and Cognition*, *6*(3), 181–200. doi: 10.1017/S1366728903001123

Costa, A. & Pickering, M. J. (2019). The role of learning on bilinguals' lexical architecture: Beyond separated vs. integrated lexicons. *Bilingualism: Language and Cognition*, *22*(4), 685–686. doi: 10.1017/S1366728918000809

Costa, A. & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, *50*(4), 491–511. doi: 10.1016/j.jml.2004.02.002

Costa, A., Santesteban, M. & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language*, *94*(1), 94–103. doi: 10.1016/j.bandl.2004.12.002

Costa, A., Santesteban, M. & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1057–1074. doi: 10.1037/0278-7393.32.5.1057

Costa, A., Strijkers, K., Martin, C. & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, *106*(50), 21442–21446. doi: 10.1073/pnas.0908921106

Coulson, S., King, J. W. & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*(1), 21–58. doi: 10.1080/016909698386582

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Cutler, A., Weber, A. & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284. doi: 10.1016/j.wocn.2005.06.002

Davies, M. & Davies, K. H. (2017). *A frequency dictionary of Spanish* (2edition ed.). London ; New York: Routledge.

Davis, C., Sánchez-Casas, R., García-Albea, J. E., Guasch, M., Molero, M. & Ferré, P. (2010). Masked translation priming: Varying language experience and word type with Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *13*(2), 137–155. doi: 10.1017/S1366728909990393

DeAnda, S., Bosch, L., Poulin, D. D., Zesiger, P. & Friend, M. (2016). The Language Exposure Assessment Tool: Quantifying language exposure in infants and children. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1346–1356. doi: 10.1044/2016_JSLHR-L-15-0234

De Bot, K. (2004). The multilingual lexicon: Modelling selection and control. *International Journal of Multilingualism*, *1*(1), 17–32. doi: 10.1080/14790710408668176

De Cat, C., Klepousniotou, E. & Baayen, R. H. (2015). Representational deficit or processing effect? An electrophysiological study of noun-noun compound processing by very advanced L2 speakers of English. *Frontiers in Psychology*, *6*, 77. doi: 10.3389/fpsyg.2015.00077

Declerck, M., Koch, I., Duñabeitia, J. A., Grainger, J. & Stephan, D. N. (2019). What absent switch costs and mixing costs during bilingual language comprehension can tell us about language control. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 771–789. doi: 10.1037/xhp0000627

Declerck, M., Meade, G., Midgley, K. J., Holcomb, P. J., Roelofs, A. & Emmorey, K. (2021). On the connection between language control and executive control - an ERP study. *Neurobiology of Language*, *2*(4), 628–646. doi: 10.1162/nol{_}a{_}00032

De Diego Balaguer, R., Sebastián-Gallés, N., Díaz, B. & Rodríguez-Fornells, A. (2005). Morphological processing in early bilinguals: An ERP study of regular and irregular verb processing. *Cognitive Brain Research*, *25*(1), 312–327. doi: 10.1016/j.cogbrainres.2005.06.003

De Groot, A. M. B. (2017). Bi-and Multilingualism. In Y. Y. Kim & K. McKay-Semmler (Eds.), *The International Encyclopedia of Intercultural Communication* (pp. 1–10). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118783665

De Groot, A. M. B. & Nas, G. L. J. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*(1), 90–123. doi: 10.1016/0749-596X(91)90012-9

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135–168. doi: 10.1146/annurev-psych-113011-143750

Diependaele, K., Lemhöfer, K. & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. doi: 10.1080/17470218.2012.720994

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M. & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, *62*(3), 284–301. doi: 10.1016/j.jml.2009.12.003

Dijkstra, T. & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. doi: 10.1017/S1366728902003012

Dijkstra, T., Van Heuven, W. J. B. & Grainger, J. (1998). Simulating cross-language competition with the bilingual interactive activation model. *Psychologica Belgica*, *38*(3-4), 177–196.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E. & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. doi: 10.1080/17470218.2017.1310261

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T. & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *bioRxiv*, 2020.01.03.894378. doi: 10.1101/2020.01.03.894378

Eddington, D. (2002). Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics*, *9*(1), 49–75. doi: 10.1076/jqul.9.1.49.8482

Eriksen, B. A. & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. doi: 10.3758/BF03203267

Eulitz, C., Hauk, O. & Cohen, R. (2000). Electroencephalographic activity over temporal brain areas during phonological encoding in picture naming. *Clinical Neurophysiology*, *111*(11), 2088–2097. doi: 10.1016/S1388-2457(00)00441-7

Festman, J., Rodriguez-Fornells, A. & Münte, T. F. (2010). Individual differences in control of language interference in late bilinguals are mainly related to general executive abilities. *Behavioral and Brain Functions*, *6*(1), 5. doi: 10 .1186/1744-9081-6-5

Fisher, R. A. (1938). Presidential address to the first Indian statistical congress. In *Sankhya* (pp. 14–17).

Foote, R. (2009). Transfer in L3 acquisition: The role of typology. In Y.-K. I. Leung (Ed.), *Third language acquisition and universal grammer* (pp. 89–114). Multilingual Matters. doi: 10.21832/9781847691323-008

Foster, E. D. & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, *105*(2), 203–206. doi: 10.5195/jmla.2017.88

Foucart, A. & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*(3), 379–399. doi: 10.1017/S136672891000012X

Foucart, A. & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language*, *66*(1), 226–248. doi: 10.1016/j.jml.2011.07.007

Franceschina, F. (2002). Case and φ-feature agreement in advanced L2 Spanish grammars. *EUROSLA Yearbook*, *2*(1), 71–86. doi:

10.1075/eurosla.2.07fra

Franceschina, F. (2005). *Fossilized second language grammars: The acquisition of grammatical gender.* John Benjamins Publishing.

Francis, W. S. & Gallard, S. L. K. (2005). Concept mediation in trilingual translation: Evidence from response time and repetition priming patterns. *Psychonomic Bulletin & Review*, *12*(6), 1082–1088. doi: 10.3758/BF03206447

Frenck-Mestre, C., Anton, J. L., Roth, M., Vaid, J. & Viallet, F. (2005). Articulation in early and late bilinguals?? two languages: Evidence from functional magnetic resonance imaging:. *NeuroReport*, *16*(7), 761–765. doi: 10.1097/00001756-200505120-00021

Friederici, A. D., Gunter, T., Hahne, A. & Mauth, K. (2004). The relative timing of syntactic and semantic processes in sentence comprehension. *NeuroReport*, *15*(1), 165–169.

Friederici, A. D., Hahne, A. & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research*, *31*(1), 45–63. doi: 10.1023/A:1014376204525

Friederici, A. D., Steinhauer, K. & Frisch, S. (1999). Lexical integration: Sequential effects of syntactic and semantic information. *Memory & Cognition*, *27*(3), 438–453. doi: 10.3758/BF03211539

Fröber, K., Stürmer, B., Frömer, R. & Dreisbach, G. (2017). The role of affective evaluation in conflict adaptation: An LRP study. *Brain and Cognition*, *116*, 9–16. doi: 10.1016/j.bandc.2017.05.003

Frömer, R., Maier, M. & Abdel Rahman, R. (2018). Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, *12*, 1–15. doi: 10.3389/fnins.2018.00048

Gamboa Arana, O. L., Palmer, H., Dannhauer, M., Hile, C., Liu, S., Hamdan, R., … Appelbaum, L. G. (2020). Intensity- and timing-dependent modulation of motion perception with transcranial magnetic stimulation of visual cortex. *Neuropsychologia*, *147*, 107581. doi: 10.1016/j.neuropsychologia.2020

.107581

Ganushchak, L. Y., Christoffels, I. K. & Schiller, N. O. (2011). The use of electroencephalography in language production research: A review. *Frontiers in Psychology*, *2*, 1–6. doi: 10.3389/fpsyg.2011.00208

Ganushchak, L. Y., Verdonschot, R. G. & Schiller, N. O. (2011). When leaf becomes neuter: Event-related potential evidence for grammatical gender transfer in bilingualism. *NeuroReport*, *22*(3), 106–110. doi: 10.1097/WNR.0b013e3283427359

Gillon-Dowens, M., Vergara, M., Barber, H. A. & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience*, *22*(8), 1870–1887. doi: 10.1162/jocn.2009.21304

Goldfarb, L. & Tzelgov, J. (2007). The cause of the within-language Stroop superiority effect and its implications. *Quarterly Journal of Experimental Psychology*, *60*(2), 179–185. doi: 10.1080/17470210600983415

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C. & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition*, *33*(7), 1220–1234. doi: 10.3758/BF03193224

Gonthier, C., Braver, T. S. & Bugg, J. M. (2016). Dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, *44*(5), 778–788. doi: 10.3758/s13421-016-0591-1

González Alonso, J., Alemán Bañón, J., DeLuca, V., Miller, D., Pereira Soares, S. M., Puig-Mayenco, E., … Rothman, J. (2020). Event related potentials at initial exposure in third language acquisition: Implications from an artificial mini-grammar study. *Journal of Neurolinguistics*, *56*, 100939. doi: 10.1016/j.jneuroling.2020.100939

Gouvea, A. C., Phillips, C., Kazanina, N. & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. doi: 10.1080/01690960902965951

Grafmiller, J. (2020). *Package JGmermod.*

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81.

doi: 10.1017/S1366728998000133

Green, D. W. & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*(5), 515–530. doi: 10.1080/20445911.2013.796377

Green, D. W. & Kroll, J. F. (2019). The neurolinguistics of bilingualism. In G. I. de Zubicaray & N. O. Schiller (Eds.), *The Oxford Handbook of Neurolinguistics.* Oxford University Press.

Green, P., MacLeod, C. J. & Alday, P. M. (2022). *Package 'simr': Power analysis for generalised linear mixed models by simulation.*

Grosjean, F. (2012). Bilingual and monolingual language modes. In *The Encyclopedia of Applied Linguistics.* Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal0090

Grözinger, B., Kornhuber, H. H. & Kriebel, J. (1975). Methodological problems in the investigation of cerebral potentials preceding speech: Determining the onset and suppressing artefacts caused by speech. *Neuropsychologia*, *13*(3), 263–270. doi: 10.1016/0028-3932(75)90002-0

Grundy, J., Anderson, J. & Bialystok, E. (2017). Neural correlates of cognitive processing in monolinguals and bilinguals. *Annals of the New York Academy of Sciences*, *1396*. doi: 10.1111/nyas.13333

Grüter, T., Lew-Williams, C. & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191–215. doi: 10.1177/0267658312437990

Gunter, T., Friederici, A. & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*, 556–568. doi: 10.1162/089892900562336

Guo, T. & Peng, D. (2006). Event-related potential evidence for parallel activation of two languages in bilingual speech production:. *NeuroReport*, *17*(17), 1757–1760. doi: 10.1097/01.wnr.0000246327.89308.a5

Habets, B., Jansma, B. M. & Münte, T. F. (2008). Neurophysiological correlates of linearization in language production. *BMC Neuroscience*, *9*(1), 1–8. doi: 10.1186/1471-2202-9-77

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, *15*(6), 883–899. doi: 10.1162/089892903322370807

Hagoort, P., Brown, C. & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. doi: 10.1080/01690969308407585

Hagoort, P. & Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research*, *28*(6), 715–728. doi: 10.1023/A:1023277213129

Hagoort, P. & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia*, *38*(11), 1531–1549. doi: 10.1016/S0028-3932(00)00053-1

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research*, *30*(3), 251–266.

Hahne, A. & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, *4*(2), 123–141. doi: 10.1017/S1366728901000232

Hamers, J. F. & Lambert, W. E. (1972). Bilingual interdependencies in auditory perception. *Journal of Verbal Learning and Verbal Behavior*, *11*(3), 303–310. doi: 10.1016/S0022-5371(72)80091-4

Hanulová, J., Davidson, D. J. & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language and Cognitive Processes*, *26*(7), 902–934. doi: 10.1080/01690965.2010.509946

Hartig, F. (2020). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models.*

Hartsuiker, R. J., Pickering, M. J. & Veltkamp, E. (2004). Is syntax separate or shared between languages?: Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, *15*(6), 409–414. doi: 10.1111/j.0956-7976.2004.00693.x

Hasting, A. S. & Kotz, S. A. (2008). Speeding up syntax: On the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *Journal of Cognitive Neuroscience*, *20*(7), 1207–1219. doi: 10.1162/jocn.2008.20083

Heidlmayr, K., Ferragne, E. & Isel, F. (2021). Neuroplasticity in the phonological system: The PMN and the N400 as markers for the perception of non-native phonemic contrasts by late second language learners. *Neuropsychologia*, *156*, 107831. doi: 10.1016/j.neuropsychologia.2021.107831

Heidlmayr, K., Moutier, S., Hemforth, B., Courtin, C., Tanzmeister, R. & Isel, F. (2014). Successive bilingualism and executive functions: The effect of second language use on inhibitory control in a behavioural Stroop Colour Word task. *Bilingualism: Language and Cognition*, *17*(3), 630–645. doi: 10.1017/S1366728913000539

Heim, S., Friederici, A. D., Schiller, N. O., Rüschemeyer, S.-A. & Amunts, K. (2009). The determiner congruency effect in language production investigated with functional MRI. *Human Brain Mapping*, *30*(3), 928–940. doi: 10.1002/hbm.20556

Hermans, D., Bongaerts, T., De Bot, K. & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*(3), 213–229. doi: 10 .1017/S1366728998000364

Hilbert, S., Nakagawa, T. T., Bindl, M. & Bühner, M. (2014). The spatial Stroop effect: A comparison of color-word and position-word interference. *Psychonomic Bulletin & Review*, *21*(6), 1509–1515. doi: 10.3758/s13423-014-0631-4

Hilchey, M. D. & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic*

Bulletin & Review, *18*(4), 625–658. doi: 10.3758/s13423-011-0116-7

Hoshino, N. & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, *106*(1), 501–511. doi: 10.1016/j.cognition.2007.02.001

Hoshino, N. & Thierry, G. (2011). Language selection in bilingual word production: Electrophysiological evidence for cross-language competition. *Brain Research*, *1371*, 100–109. doi: 10.1016/j.brainres.2010.11.053

Hruby, T. & Marsalek, P. (2003). Event-related potentials - the P3 wave. *Acta Neurobiologiae Experimentalis*, *63*, 55–63.

Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. (2001, June). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564. doi: 10.1093/biomet/88.2.551

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 1–16. doi: 10.3389/fpsyg.2011.00255

Indefrey, P. & Levelt, W. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144. doi: 10.1016/j.cognition.2002.06.001

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

Ivanova, I. & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, *127*(2), 277–288. doi: 10.1016/j.actpsy.2007.06.003

Izura, C., Cuetos, F. & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicologica*, *35*(1), 49–67.

Janyan, A. & Hristova, M. (2007). Gender congruency and cognate effect in Bulgarian-English bilinguals: Evidence from a word-translation task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29), 1121–1126.

Jarvis, S. (2011). Conceptual transfer: Crosslinguistic effects in categorization and construal. *Bilingualism: Language and Cog-*

nition, *14*(1), 1–8. doi: 10.1017/S1366728910000155

Jeong, H., Sugiura, M., Sassa, Y., Haji, T., Usui, N., Taira, M., ... Kawashima, R. (2007). Effect of syntactic similarity on cortical activation during second language processing: A comparison of English and Japanese among native Korean trilinguals. *Human Brain Mapping*, *28*(3), 194–204. doi: 10.1002/hbm.20269

Jiao, L., Grundy, J. G., Liu, C. & Chen, B. (2020). Language context modulates executive control in bilinguals: Evidence from language production. *Neuropsychologia*, *142*, 107441. doi: 10.1016/j.neuropsychologia.2020.107441

Johnson, T. & Fendrich, M. (2005). Modeling sources of self-report bias in a survey of drug use epidemiology. *Annals of Epidemiology*, *15*(5), 381–389. doi: 10.1016/j.annepidem.2004.09.004

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, *1*(6), 571–591. doi: 10.1111/j.1749-818X.2007.00037.x

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

Kaushanskaya, M., Blumenfeld, H. K. & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism: Language and Cognition*, *23*(5), 945–950. doi: 10.1017/S1366728919000038

Klassen, R. (2016). The representation of asymmetric grammatical gender systems in the bilingual mental lexicon. *Probus*, *28*(1), 9–28. doi: 10.1515/probus-2016-0002

Koester, D. & Schiller, N. O. (2008). Morphological priming in overt language production: Electrophysiological evidence from Dutch. *NeuroImage*, *42*(4), 1622–1630. doi: 10.1016/j.neuroimage.2008.06.043

Kornrumpf, B., Niefind, F., Sommer, W. & Dimigen, O. (2016). Neural correlates of word recognition: A systematic comparison of natural reading and rapid serial visual presentation. *Journal of Cognitive Neuroscience*, *28*(9), 1374–1391. doi: 10.1162/jocn{_}a{_}00977

Kotz, S. A., Holcomb, P. J. & Osterhout, L. (2008). ERPs reveal comparable syntactic sentence processing in native and nonnative readers of English. *Acta Psychologica*, *128*(3), 514–527. doi: 10.1016/j.actpsy.2007.10.003

Kroll, J. F. & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, *25*(5), 497–514. doi: 10.1080/20445911.2013.799170

Kroll, J. F., Bobb, S. C., Misra, M. & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*(3), 416–430. doi: 10.1016/j.actpsy.2008.02.001

Kroll, J. F., Bobb, S. C. & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, *9*(2), 119–135. doi: 10.1017/S1366728906002483

Kroll, J. F., Dussias, P. E., Bice, K. & Perrotti, L. (2015). Bilingualism, mind, and brain. *Annual Review of Linguistics*, *1*(1), 377–394. doi: 10.1146/annurev-linguist-030514-124937

Kroll, J. F. & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*, 149–174.

Kroll, J. F. & Tokowicz, N. (2005). *Models of bilingual representation and processing: Looking back and to the future.* Oxford University Press.

Kutas, M. & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M. & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. doi: 10.1016/0301-0511(80)90046-0

Kuzmina, E., Goral, M., Norvik, M. & Weekes, B. S. (2019). What influences language impairment in bilingual aphasia? A meta-analytic review. *Frontiers in Psychology*, *10*, 445. doi: 10

.3389/fpsyg.2019.00445

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. & Pødenphant-Jensen, S. (2020). *Package lmerTest.*

Laganaro, M., Morand, S., Schwitter, V., Zimmermann, C., Camen, C. & Schnider, A. (2009). Electrophysiological correlates of different anomic patterns in comparison with normal word production. *Cortex*, *45*(6), 697–707. doi: 10.1016/j.cortex.2008.09.007

Lago, S., Mosca, M. & Garcia, A. S. (2021). The role of cross-linguistic influence in multilingual processing: Lexicon versus syntax. *Language Learning*, *71*(S1), 163–192. doi: 10.1111/lang.12412

La Heij, W., Van der Heijden, A. H. C. & Plooij, P. (2001). A paradoxical exposure-duration effect in the Stroop task: Temporal segregation between stimulus attributes facilitates selection. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 622–632. doi: 10.1037/0096-1523.27.3.622

Lange, V. M., Perret, C. & Laganaro, M. (2015). Comparison of single-word and adjective-noun phrase production using event-related brain potentials. *Cortex*, *67*, 15–29. doi: 10.1016/j.cortex.2015.02.017

Lau, E. F., Phillips, C. & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. doi: 10.1038/nrn2532

Leckey, M. & Federmeier, K. D. (2019). Electrophysiological methods in the study of language processing. In G. I. de Zubicaray & N. O. Schiller (Eds.), *The Oxford Handbook of Neurolinguistics.* Oxford University Press. doi: 10.1093/oxfordhb/9780190672027.013.3

Lee, M.-W. & Williams, J. N. (2001). Lexical access in spoken word production by bilinguals: Evidence from the semantic competitor priming paradigm. *Bilingualism: Language and Cognition*, *4*(3), 233–248. doi: 10.1017/S1366728901000426

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343. doi: 10.3758/

s13428-011-0146-0

Lemhöfer, K. & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, *32*(4), 533–550. doi: 10.3758/BF03195845

Lemhöfer, K., Dijkstra, T. & Michel, M. (2004). Three languages, one ECHO: Cognate effects in trilingual word recognition. *Language and Cognitive Processes*, *19*(5), 585–611. doi: 10.1080/01690960444000007

Lemhöfer, K., Spalek, K. & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language*, *59*(3), 312–330. doi: 10.1016/j.jml.2008.06.005

Lenth, R., Buerkner, P., Herve, M., Love, J., Riebl, H. & Singmann, H. (2019). *Package "emmeans": Estimated marginal means, aka least-squares means.*

Lev-Ari, S. & Peperkamp, S. (2013). Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics*, *41*(5), 320–331. doi: 10.1016/j.wocn.2013.06.002

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences*, *22*, 1–38. doi: 10.1017/S0140525X99001776

Li, C. & Gollan, T. H. (2018). Cognates interfere with language selection but enhance monitoring in connected speech. *Memory & Cognition*, *46*(6), 923–939. doi: 10.3758/s13421-018-0812-x

Li, P., Zhang, F., Tsai, E. & Puls, B. (2014). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition*, *17*(3), 673–680. doi: 10.1017/S1366728913000606

Linck, J. A., Hoshino, N. & Kroll, J. F. (2005). Cross-language lexical processes and inhibitory control. *The Mental Lexicon*, *3*(3), 349–374. doi: 10.1075/ml.3.3.06lin

Lo, S. & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1–16. doi:

10.3389/fpsyg.2015.01171

Lu, C.-h. & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, *2*(2), 174–207. doi: 10.3758/BF03210959

Luck, S. J. (1998). Sources of dual-task interference: Evidence from human electrophysiology. *Psychological Science*, *9*(3), 223–227. doi: 10.1111/1467-9280.00043

Luo, C. & Proctor, R. W. (2013). Asymmetry of congruency effects in spatial Stroop tasks can be eliminated. *Acta Psychologica*, *143*(1), 7–13. doi: 10.1016/j.actpsy.2013.01.016

MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General*, *121*(1), 12–14. doi: 10.1037/0096-3445.121.1.12

MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches.* Oxford University Press.

Mardia, K. V., Southworth, H. R. & Taylor, C. C. (1999). On bias in maximum likelihood estimators. *Journal of Statistical Planning and Inference*, *76*(1), 31–39. doi: 10.1016/S0378-3758(98)00176-1

Marian, V., Blumenfeld, H. K. & Boukrina, O. V. (2008). Sensitivity to phonological similarity within and across languages. *Journal of psycholinguistic research*, *37*(3), 141–170. doi: 10.1007/s10936-007-9064-9

Marian, V., Blumenfeld, H. K. & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. doi: 10.1044/1092-4388(2007/067)

Marian, V., Blumenfeld, H. K., Mizrahi, E., Kania, U. & Cordes, A.-K. (2013). Multilingual Stroop performance: Effects of trilingualism and proficiency on inhibitory control. *International Journal of Multilingualism*, *10*(1), 82–104. doi: 10.1080/14790718.2012.708037

Marian, V. & Hayakawa, S. (2021). Measuring bilingualism: The quest for a “bilingualism quotient”. *Applied Psycholinguistics*, *42*(2), 527–548. doi: 10.1017/S0142716420000533

Marian, V. & Spivey, M. (2003a). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*(2), 173–193. doi: 10.1017/S0142716403000092

Marian, V. & Spivey, M. (2003b). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*(2), 97–115. doi: 10.1017/S1366728903001068

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Martín-Loeches, M., Muñoz, F., Casado, P., Melcón, A. & Fernández-Frías, C. (2005). Are the anterior negativities to grammatical violations indexing working memory? *Psychophysiology*, *42*(5), 508–519. doi: 10.1111/j.1469-8986.2005.00308.x

Martín-Loeches, M., Nigbur, R., Casado, P., Hohlfeld, A. & Sommer, W. (2006). Semantics prevalence over syntax during sentence processing: A brain potential study of noun–adjective agreement in Spanish. *Brain Research*, *1093*(1), 178–189. doi: 10.1016/j.brainres.2006.03.094

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H. R. & Bates, D. M. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. doi: 10.1016/j.jml.2017.01.001

McMahon, A. & McMahon, R. (2005). *Language Classification by Numbers.* Oxford University Press.

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A. & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLOS ONE*, *10*(12), e0143328. doi: 10.1371/journal.pone.0143328

Midgley, K. J., Holcomb, P. J. & Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *Journal of Neurolinguistics*, *22*(3), 281–300. doi: 10.1016/j.jneuroling.2008.08.001

Midgley, K. J., Holcomb, P. J. & Grainger, J. (2011). Effects of cognate status on word comprehension in second language learners: An ERP investigation. *Journal of Cognitive Neuroscience*, *23*(7), 1634–1647. doi: 10.1162/jocn.2010.21463

Miozzo, M. & Caramazza, A. (1999). The selection of determiners in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 907–922. doi: 10.1037/0278-7393.25.4.907

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science*, *48*(2), 171–203. doi: 10.1177/0306312718772086

Mishra, R. K. & Singh, N. (2016). The influence of second language proficiency on bilingual parallel language activation in Hindi–English bilinguals. *Journal of Cognitive Psychology*, *28*(4), 396–411. doi: 10.1080/20445911.2016.1146725

Misra, M., Guo, T., Bobb, S. C. & Kroll, J. F. (2012). When bilinguals choose a single word to speak: Electrophysiological evidence for inhibition of the native language. *Journal of Memory and Language*, *67*(1), 224–237. doi: 10.1016/j.jml.2012.05.001

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. doi: 10.1006/cogp.1999.0734

Molinaro, N., Barber, H. A., Caffarra, S. & Carreiras, M. (2015). On the left anterior negativity (LAN): The case of morphosyntactic agreement: A Reply to Tanner et al. *Cortex*, *66*, 156–159. doi: 10.1016/j.cortex.2014.06.009

Molinaro, N., Barber, H. A. & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930. doi: 10.1016/j.cortex.2011.02.019

Molinaro, N., Vespignani, F. & Job, R. (2008). A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research*, *1228*, 161–176. doi: 10.1016/j.brainres.2008.06.064

Morales, L., Paolieri, D., Dussias, P. E., Valdés Kroff, J. R., Gerfen, C. & Bajo, M. T. (2016). The gender congruency effect during bilingual spoken-word recognition. *Bilingualism: Language and Cognition*, *19*(2), 294–310. doi: 10.1017/S1366728915000176

Moreno, E. M., Rodríguez-Fornells, A. & Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *Journal of Neurolinguistics*, *21*(6), 477–508. doi: 10.1016/j.jneuroling.2008.01.003

Moreno, S., Bialystok, E., Wodniecka, Z. & Alain, C. (2010). Conflict resolution in sentence processing by bilinguals. *Journal of Neurolinguistics*, *23*(6), 564–579. doi: 10.1016/j.jneuroling.2010.05.002

Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P. & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations? *Cognition*, *118*(2), 286–292. doi: 10.1016/j.cognition.2010.11.006

Mosca, M. (2017). *Multilingual's language control* (Unpublished doctoral dissertation). University of Potsdam, Potsdam.

Mosca, M. & De Bot, K. (2017). Bilingual language switching: Production vs. recognition. *Frontiers in Psychology*, *8*, 934. doi: 10.3389/fpsyg.2017.00934

Motamedi-Fakhr, S., Moshrefi-Torbati, M., Hill, M., Hill, C. M. & White, P. R. (2014). Signal processing techniques applied to human sleep EEG signals—A review. *Biomedical Signal Processing and Control*, *10*, 21–33. doi: 10.1016/j.bspc.2013.12.003

Müller, N. & Hulk, A. (2001). Crosslinguistic influence in bilingual language acquisition: Italian and French as recipient languages. *Bilingualism: Language and Cognition*, *4*(1), 1–21. doi: 10.1017/S1366728901000116

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. doi: 10.1038/s41562-016-0021

Münte, T. F., Matzke, M. & Johannes, S. (1997). Brain activity associated with syntactic incongruencies in words and pseudowords. *Journal of Cognitive Neuroscience*, *9*(3), 318–329. doi:

10.1162/jocn.1997.9.3.318

Neath, A. A. & Cavanaugh, J. E. (2012). The Bayesian Information Criterion: Background, derivation, and applications. *WIREs Computational Statistics*, *4*(2), 199–203. doi: 10.1002/wics.199

Neuhaus, A. H., Trempler, N. R., Hahn, E., Luborzewski, A., Karl, C., Hahn, C., ... Dettling, M. (2010). Evidence of specificity of a visual P3 amplitude modulation deficit in schizophrenia. *Schizophrenia Research*, *124*(1-3), 119–126. doi: 10.1016/j.schres.2010.08.014

Neuhaus, A. H., Urbanek, C., Opgen-Rhein, C., Hahn, E., Ta, T. M. T., Koehler, S., ... Dettling, M. (2010). Event-related potentials associated with Attention Network Test. *International Journal of Psychophysiology*, *76*(2), 72–79. doi: 10.1016/j.ijpsycho.2010.02.005

Neville, H., Nicol, J. L., Barss, A., Forster, K. I. & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*(2), 151–165. doi: 10.1162/jocn.1991.3.2.151

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J. & Ullman, M. T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience*, *24*(5), 1205–1223. doi: 10.1162/jocn_a_00143

Nichols, E. S. & Joanisse, M. F. (2019). Individual differences predict ERP signatures of second language learning of novel grammatical rules. *Bilingualism: Language and Cognition*, *22*(1), 78–92. doi: 10.1017/S1366728917000566

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

Nozari, N. & Pinet, S. (2020). A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. *Journal of Neurolinguistics*, *53*, 1–19. doi: 10.1016/j.jneuroling.2019.100875

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning.* Cambridge University Press.

Osterhout, L. & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. doi: 10.1016/0749-596X(92)90039-Z

Osterhout, L., Mclaughlin, J., Pitkänen, I., Frenck-Mestre, C. & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning*, *56*, 199–230.

Osterhout, L. & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773. doi: 10.1006/jmla.1995.1033

Osterhout, L. & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes*, *14*(3), 283–317. doi: 10.1080/016909699386310

Padrón, I., Fraga, I. & Acuña-Fariña, C. (2020). Processing gender agreement errors in pleasant and unpleasant words: An ERP study at the sentence level. *Neuroscience Letters*, *714*, 134538. doi: 10.1016/j.neulet.2019.134538

Paolieri, D., Demestre, J., Guasch, M., Bajo, T. & Ferré, P. (2020). The gender congruency effect in Catalan–Spanish bilinguals: Behavioral and electrophysiological evidence. *Bilingualism: Language and Cognition*, *23*(5), 1–11. doi: 10.1017/S1366728920000073

Paolieri, D., Padilla, F., Koreneva, O., Morales, L. & Macizo, P. (2019). Gender congruency effects in Russian–Spanish and Italian–Spanish bilinguals: The role of language proximity and concreteness of words. *Bilingualism: Language and Cognition*, *22*(1), 112–129. doi: 10.1017/S1366728917000591

Pardo, J. V., Pardo, P. J., Janer, K. W. & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, *87*(1), 256–259. doi: 10.1073/pnas.87.1.256

Parker-Jones, Ō., Green, D. W., Grogan, A., Pliatsikas, C., Filippopolitis, K., Ali, N., ... Price, C. J. (2012). Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. *Cerebral Cortex*, *22*(4), 892–902. doi: 10.1093/cercor/bhr161

Pauli, R., Bowring, A., Reynolds, R., Chen, G., Nichols, T. E. & Maumet, C. (2016). Exploring fMRI results space: 31 variants of an fMRI analysis in AFNI, FSL, and SPM. *Frontiers in Neuroinformatics*, *10*.

Peeters, D., Dijkstra, T. & Grainger, J. (2013). The representation and processing of identical cognates by late bilinguals: RT and ERP effects. *Journal of Memory and Language*, *68*(4), 315–332. doi: 10.1016/j.jml.2012.12.003

Pereira Soares, S. M., Kupisch, T. & Rothman, J. (2022). Testing potential transfer effects in heritage and adult L2 bilinguals acquiring a mini grammar as an additional language: An ERP ppproach. *Brain Sciences*, *12*(5), 669. doi: 10.3390/brainsci12050669

Pereira Soares, S. M., Ong, G., Abutalebi, J., Del Maschio, N., Sewell, D. & Weekes, B. (2019). A diffusion model approach to analyzing performance on the Flanker task: The role of the DLPFC. *Bilingualism: Language and Cognition*, *22*(5), 1194–1208. doi: 10.1017/S1366728918000974

Pika, S., Nicoladis, E. & Marentette, P. F. (2006). A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer? *Bilingualism: Language and Cognition*, *9*(3), 319–327. doi: 10.1017/S1366728906002665

Pivneva, I., Palmer, C. & Titone, D. (2012). Inhibitory control and L2 proficiency modulate bilingual language production: Evidence from spontaneous monologue and dialogue speech. *Frontiers in Psychology*, *3*. doi: 10.3389/fpsyg.2012.00057

Pliatsikas, C. (2020). Understanding structural plasticity in the bilingual brain: The Dynamic Restructuring Model. *Bilingualism: Language and Cognition*, *23*(2), 459–471. doi: 10.1017/S1366728919000130

Poarch, G. J. & Van Hell, J. G. (2012a). Cross-language activation in children's speech production: Evidence from second language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *111*(3), 419–438. doi: 10.1016/j.jecp.2011.09.008

Poarch, G. J. & Van Hell, J. G. (2012b). Executive functions and inhibitory control in multilingual children: Evidence from second-language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *113*(4), 535–551. doi: 10.1016/j.jecp.2012.06.013

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Polich, J. (2012). Neuropsychology of P300. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components.* Oxford, UK: Oxford University Press. doi: 10.1093/oxfordhb/9780195374148.013.0089

Porcaro, C., Medaglia, M. T. & Krott, A. (2015). Removing speech artifacts from electroencephalographic recordings during overt picture naming. *NeuroImage*, *105*, 171–180. doi: 10.1016/j.neuroimage.2014.10.049

Puig-Mayenco, E., González Alonso, J. & Rothman, J. (2020). A systematic review of transfer studies in third language acquisition. *Second Language Research*, *36*(1), 31–64. doi: 10.1177/0267658318809147

Pulvermüller, F. (2007). Word processing in the brain as revealed by neurophysiological imaging. In *The Oxford Handbook of Psycholinguistics* (pp. 118–140). Oxford University Press. doi: 10.1093/oxfordhb/9780198568971.013.0008

Pulvermüller, F., Shtyrov, Y. & Hauk, O. (2009). Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, *110*(2), 81–94. doi: 10.1016/j.bandl.2008.12.001

Putnam, M. T., Carlson, M. & Reitter, D. (2018). Integrated, not isolated: Defining typological proximity in an integrated multilingual architecture. *Frontiers in Psychology*, *8*, 1–16. doi: 10.3389/fpsyg.2017.02212

Ringbom, H. (1987). *The role of the first language in foreign language learning* (Vol. 34). Clevedon, Avon, England: Multilingual Matters.

Ringbom, H. & Jarvis, S. (2009). The importance of cross-linguistic similarity in language learning. In M. H. Long & C. Doughty (Eds.), *The handbook of language teaching.* Chichester, U.K. ; Malden, MA: Blackwell Publishing Ltd.

Ritter, W. & Vaughan, H. G. (1969). Averaged evoked responses in vigilance and discrimination: A reassessment. *Science*, *164*(3877), 326–328. doi: 10.1126/science.164.3877.326

Rodrigues, J., Weiß, M., Hewig, J. & Allen, J. J. B. (2021). EPOS: EEG Processing Open-Source Scripts. *Frontiers in Neuroscience*, *15*.

Rodriguez-Fornells, A., Kramer, U., Lorenzo-Seva, U., Festman, J. & Münte, T. (2012). Self-Assessment of Individual Differences in Language Switching. *Frontiers in Psychology*, *2*, 388. doi: 10.3389/fpsyg.2011.00388

Rodriguez-Fornells, A., Schmitt, B. M., Kutas, M. & Münte, T. F. (2002). Electrophysiological estimates of the time course of semantic and phonological encoding during listening and naming. *Neuropsychologia*, *40*(7), 778–787. doi: 10.1016/S0028-3932(01)00188-9

Rodriguez-Fornells, A., van der Lugt, A., Rotte, M., Britti, B., Heinze, H.-J. & Münte, T. F. (2005). Second language interferes with word production in fluent bilinguals: Brain potential and functional imaging evidence. *Journal of Cognitive Neuroscience*, *17*(3), 422–433. doi: 10.1162/0898929053279559

Roelofs, A. (2021). Response competition better explains Stroop interference than does response exclusion. *Psychonomic Bulletin & Review*, *28*(2), 487–493. doi: 10.3758/s13423-020-01846-0

Roelofs, A., Piai, V., Garrido Rodriguez, G. & Chwilla, D. J. (2016). Electrophysiology of cross-language interference and facilitation in picture naming. *Cortex*, *76*, 1–16. doi: 10.1016/j.cortex.2015.12.003

Rosenman, R., Tennekoon, V. & Hill, L. G. (2011). Measuring bias in self-reported data. *International journal of behavioural & healthcare research*, *2*(4), 320–332. doi: 10.1504/IJBHR.2011

.043414

Rossi, E., Newman, S., Kroll, J. F. & Diaz, M. T. (2018). Neural signatures of inhibitory control in bilingual spoken production. *Cortex*, *108*, 50–66. doi: 10.1016/j.cortex.2018.07.009

Rossi, S., Gugler, M. F., Friederici, A. D. & Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience*, *18*(12), 2030–2048. doi: 10.1162/jocn.2006.18.12.2030

Rothman, J. (2015). Linguistic and cognitive motivations for the Typological Primacy Model (TPM) of third language (L3) transfer: Timing of acquisition and proficiency considered*. *Bilingualism: Language and Cognition*, *18*(2), 179–190. doi: 10.1017/S136672891300059X

Rothman, J. & Cabrelli Amaro, J. (2010). What variables condition syntactic transfer? A look at the L3 initial state. *Second Language Research*, *26*(2), 189–218. doi: 10.1177/0267658309349410

Runnqvist, E., Strijkers, K., Sadat, J. & Costa, A. (2011). On the temporal and functional origin of L2 disadvantages in speech production: A critical review. *Frontiers in Psychology*, *2*, 1–8. doi: 10.3389/fpsyg.2011.00379

Rust, N. C. & Mehrpour, V. (2020). Understanding image memorability. *Trends in Cognitive Sciences*, *24*(7), 557–568. doi: 10.1016/j.tics.2020.04.001

Sá-Leite, A. R., Fraga, I. & Comesaña, M. (2019). Grammatical gender processing in bilinguals: An analytic review. *Psychonomic Bulletin & Review*, *26*(4), 1148–1173. doi: 10.3758/s13423-019-01596-8

Sá-Leite, A. R., Luna, K., Fraga, I. & Comesaña, M. (2020). The gender congruency effect across languages in bilinguals: A meta-analysis. *Psychonomic Bulletin & Review*, *27*, 677–693. doi: 10.3758/s13423-019-01702-w

Sabourin, L. & Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, *24*(3), 397–430. doi: 10.1177/0267658308090186

Sabourin, L., Stowe, L. A. & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, *22*(1), 1–29. doi: 10.1191/0267658306sr259oa

Salamoura, A. & Williams, J. N. (2007). The representation of grammatical gender in the bilingual lexicon: Evidence from Greek and German. *Bilingualism: Language and Cognition*, *10*(3), 257–275. doi: 10.1017/S1366728907003069

Schendan, H. E. & Kutas, M. (2003). Time course of processes and representations supporting visual object identification and memory. *Journal of Cognitive Neuroscience*, *15*(1), 111–135. doi: 10.1162/089892903321107864

Schepens, J., Dijkstra, T. & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166. doi: 10.1017/S1366728910000623

Schepens, J., Dijkstra, T., Grootjen, F. & Van Heuven, W. J. B. (2013). Cross-language distributions of high frequency and phonetically similar cognates. *PLoS ONE*, *8*(5), 1–16. doi: 10.1371/journal.pone.0063006

Schiller, N. O. (2006). Lexical stress encoding in single word production estimated by event-related brain potentials. *Brain Research*, *1112*(1), 201–212. doi: 10.1016/j.brainres.2006.07.027

Schiller, N. O., Bles, M. & Jansma, B. M. (2003). Tracking the time course of phonological encoding in speech production: An event-related brain potential study. *Cognitive Brain Research*, *17*(3), 819–831. doi: 10.1016/S0926-6410(03)00204-0

Schiller, N. O. & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, *48*(1), 169–194. doi: 10.1016/S0749-596X(02)00508-9

Schiller, N. O. & Costa, A. (2006). Different selection principles of freestanding and bound morphemes in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1201–1207. doi: 10.1037/0278-7393.32.5.1201

Schneider, W., Eschman, A. & Zuccolotto, A. (2002). *E-prime User's Guide.*

Schriefers, H. (1992). Lexical access in the production of noun phrases. *Cognition*, *45*(1), 33–54. doi: 10.1016/0010-0277(92) 90022-A

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 841–850. doi: 10.1037/0278 -7393.19.4.841

Schriefers, H., de Ruiter, J. P. & Steigerwald, M. (1999). Parallelism in the production of noun phrases: Experiments and reaction time models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(3), 702–720. doi: 10.1037/0278 -7393.25.3.702

Schriefers, H. & Jescheniak, J. D. (1999). Representation and processing of grammatical gender in language production: A review. *Journal of Psycholinguistic Research*, *28*(6), 575–600. doi: 10.1023/A:1023264810403

Schwartz, B. D. & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, *12*(1), 40–72. doi: 10.1177/026765839601200103

Schwartz, M., Minkov, M., Dieser, E., Protassova, E., Moin, V. & Polinsky, M. (2015). Acquisition of Russian gender agreement by monolingual and bilingual children. *International Journal of Bilingualism*, *19*(6), 726–752. doi: 10.1177/ 1367006914544989

Schwieter, J. W. (2016). *Cognitive control and consequences of multilingualism.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Serratrice, L., Sorace, A., Filiaci, F. & Baldo, M. (2012). Pronominal objects in English–Italian and Spanish–Italian bilingual children. *Applied Psycholinguistics*, *33*(4), 725–751. doi: 10.1017/S0142716411000543

Sholl, A., Sankaranarayanan, A. & Kroll, J. F. (1995). Transfer between picture naming and translation: A test of asymmetries in bilingual memory. *Psychological Science*, *6*(1), 45–49. doi: 10.1111/j.1467-9280.1995.tb00303.x

Shor, R. E. (1970). The processing of conceptual information on spatial directions from pictorial and linguistic symbols. *Acta Psychologica*, *32*, 346–365. doi: 10.1016/0001-6918(70)90109-5

Shrout, P. E. & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. doi: 10.1146/annurev-psych-122216-011845

Simon, J. R. & Small, A. M. (1969). Processing auditory information: Interference from an irrelevant cue. *Journal of Applied Psychology*, *53*(5), 433–435. doi: 10.1037/h0028034

Skeide, M. A. & Friederici, A. D. (2017). Neurolinguistic studies of sentence comprehension. In *The Handbook of Psycholinguistics* (pp. 438–456). John Wiley & Sons, Ltd. doi: 10.1002/9781118829516.ch19

Sönning, L. & Werner, V. (2021). The replication crisis, scientific revolutions, and linguistics. *Linguistics*, *59*(5), 1179–1206. doi: 10.1515/ling-2019-0045

Squires, N. K., Squires, K. C. & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*(4), 387–401. doi: 10.1016/0013-4694(75)90263-1

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Strik, W., Wiest, R., . . . Dierks, T. (2012). Structural plasticity in the language system related to increased second language proficiency. *Cortex*, *48*(4), 458–465. doi: 10.1016/j.cortex.2010.10.007

Steinhauer, K. & Connolly, J. (2008). Event-related potentials in the study of language. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of the neuroscience of language* (pp. 191–203). Elsevier Ltd. doi: 10.1016/B978-0-08-045352-1.00009-4

Steinhauer, K. & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135–162. doi: 10.1016/j.bandl.2011.07.001

Steinhauer, K., White, E. J. & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research*, *25*(1), 13–41. doi: 10.1177/0267658308098995

Stocco, A., Yamasaki, B., Natalenko, R. & Prat, C. S. (2014). Bilingual brain training: A neurobiological framework of how bilingual experience improves executive function. *International Journal of Bilingualism*, *18*(1), 67–92. doi: 10.1177/1367006912456617

Stocker, L. & Berthele, R. (2020). The roles of language mode and dominance in French–German bilinguals' motion event descriptions. *Bilingualism: Language and Cognition*, *23*(3), 519–531. doi: 10.1017/S1366728919000294

Strijkers, K., Costa, A. & Thierry, G. (2010). Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, *20*(4), 912–928. doi: 10.1093/cercor/bhp153

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *8*(6), 643–662. doi: 10.1037/h0054651

Sunderman, G. & Kroll, J. F. (2006). First language activation during second language lexical processing: An investigation of lexical form, meaning, and grammatical class. *Studies in Second Language Acquisition*, *28*(3), 387–422.

Sung, J. E., Yoo, J. K., Lee, S. E. & Eom, B. (2017). Effects of age, working memory, and word order on passive-sentence comprehension: Evidence from a verb-final language. *International Psychogeriatrics*, *29*(6), 939–948. doi: 10.1017/S1041610217000047

Swaab, T. Y., Ledoux, K., Camblin, C. C. & Boudewyn, M. A. (2011). *Language-related ERP components* (E. S. Kappenman & S. J. Luck, Eds.). Oxford University Press. doi: 10.1093/oxfordhb/9780195374148.013.0197

Szewczyk, J. M. & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. doi: 10.1080/23273798.2017.1401101

Tackett, J. L., Brandes, C. M., King, K. M. & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, *15*(1), 579–604.

doi: 10.1146/annurev-clinpsy-050718-095710

Team, R. C. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

Thierry, G. & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, *104*(30), 12530–12535. doi: 10.1073/pnas.0609927104

Tokowicz, N. & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*(2), 173–204. doi: 10.1017/S0272263105050102

Tolentino, L. C. & Tokowicz, N. (2011). Across languages, space and time: A review of the role of cross-language similarity in L2 (morpho)syntactic processing as revealed by fMRI and ERP methods. *Studies in Second Language Acquisition*, *33*(1), 91–125. doi: 10.1017/S0272263110000549

Torres, L., Blevins, A. S., Bassett, D. S. & Eliassi-Rad, T. (2020). The why, how, and when of representations for complex systems. *arXiv:2006.02870 [cs, q-bio]*.

Tremblay, A. & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples: Modeling using mixed-effects regression. *Psychophysiology*, *52*(1), 124–139. doi: 10.1111/psyp.12299

Unsworth, S. (2008). Age and input in the acquisition of grammatical gender in Dutch. *Second Language Research*, *24*(3), 365–395. doi: 10.1177/0267658308090185

Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A. & Tsimpli, I. (2014). The role of age of onset and input in early child bilingualism in Greek and Dutch. *Applied Psycholinguistics*, *35*(4), 765–805. doi: 10.1017/S0142716412000574

Valente, A., Bürki, A. & Laganaro, M. (2014). ERP correlates of word production predictors in picture naming: A trial by trial multiple regression analysis from stimulus onset to response. *Frontiers in Neuroscience*, *8*, 1–13. doi: 10.3389/fnins.2014.00390

Valente, A., Pinet, S., Alario, F.-X. & Laganaro, M. (2016). "When" does picture naming take longer than word reading? *Frontiers in Psychology*, *0*. doi: 10.3389/fpsyg.2016.00031

Van der Meij, M., Cuetos, F., Carreiras, M. & Barber, H. A. (2011). Electrophysiological correlates of language switching in second language learners. *Psychophysiology*, *48*(1), 44–54. doi: 10.1111/j.1469-8986.2010.01039.x

Van der Slik, F. W. P. (2010). Acquisition of Dutch as a second language: The explanative power of cognate and genetic linguistic distance measures for 11 West European first languages. *Studies in Second Language Acquisition*, *32*(3), 401–432. doi: 10.1017/S0272263110000021

Van Hell, J. G. & Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research*, *26*(1), 43–74. doi: 10.1177/0267658309337637

Van Heuven, W. J. B., Conklin, K., Coderre, E. L., Guo, T. & Dijkstra, T. (2011). The influence of cross-language similarity on within- and between-language Stroop effects in trilinguals. *Frontiers in Psychology*, *2*, 1–15. doi: 10.3389/fpsyg.2011.00374

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M. & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*, *3*(2), 131–150. doi: 10.1162/jocn.1991.3.2.131

Van Rij, J., Wieling, M. & Baayen, H. R. (2020). *Itsadug: Interpreting time series and autocorrelated data using GAMMs.*

Van Turennout, M. & Hagoort, P. (1997). Electrophysiological evidence on the time course of semantic and phonological processes in speech productio. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 787–806.

Verbyla, A. P. (2019). A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics*, *61*(1), 39–50. doi: 10.1111/anzs.12254

Verdonschot, R. G., Middelburg, R., Lensink, S. E. & Schiller, N. O.

(2012). Morphological priming survives a language switch. *Cognition*, *124*(3), 343–349. doi: 10.1016/j.cognition.2012.05.019

Vicente-Saez, R. & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. doi: 10.1016/j.jbusres.2017.12.043

Viglione, G. (2020). Are women publishing less during the pandemic? Here's what the data say. *Nature*, *581*(7809), 365–366. doi: 10.1038/d41586-020-01294-9

Voeten, C. C. (2019). *Permutes: Permutation tests for time series data.*

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L. & Schiller, N. O. (2021a). Cross-linguistic interference in late language learners: An ERP study. *Brain and Language*, *221*, 104993. doi: 10.1016/j.bandl.2021.104993

Von Grebmer Zu Wolfsthurn, S., Pablos-Robles, L. & Schiller, N. O. (2021b). Noun-phrase production as a window to language selection: An ERP study. *Neuropsychologia*, *162*, 108055. doi: 10.1016/j.neuropsychologia.2021.108055

Walenski, M., Europa, E., Caplan, D. & Thompson, C. K. (2019). Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies. *Human Brain Mapping*, *40*(8), 2275–2304. doi: 10.1002/hbm.24523

Waller, L., Erk, S., Pozzi, E., Toenders, Y. J., Haswell, C. C., Büttner, M., . . . Veer, I. M. (2022). ENIGMA HALFpipe: Interactive, reproducible, and efficient analysis for resting-state and task-based fMRI data. *Human Brain Mapping*, *43*(9), 2727–2742. doi: 10.1002/hbm.25829

Weber-Fox, C. M. & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience*, *8*(3), 231–256. doi: 10.1162/jocn.1996.8.3.231

Weinreich, U. (1953). *Languages in contact: Findings and problems.* Berlin: Mouton de Gruyter.

Weissberger, G. H., Gollan, T. H., Bondi, M. W., Clark, L. R. & Wierenga, C. E. (2015). Language and task switching in the bilingual brain: Bilinguals are staying, not switching, experts. *Neuropsychologia*, *66*, 193–203. doi: 10.1016/j.neuropsychologia.2014.10.037

Westergaard, M., Mitrofanova, N., Mykhaylyk, R. & Rodina, Y. (2017). Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model. *International Journal of Bilingualism*, *21*(6), 666–682. doi: 10.1177/1367006916648859

Westfall, J., Kenny, D. A. & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. doi: 10.1037/xge0000014

White, E. J., Genesee, F. & Steinhauer, K. (2012). Brain responses before and after intensive second language learning: Proficiency based changes and first language background effects in adult learners. *PLOS ONE*, *7*(12), e52318. doi: 10.1371/journal.pone.0052318

White, E. J., Titone, D., Genesee, F. & Steinhauer, K. (2017). Phonological processing in late second language learners: The effects of proficiency and task. *Bilingualism: Language and Cognition*, *20*(1), 162–183. doi: 10.1017/S1366728915000620

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

Wicha, N. Y. Y., Moreno, E. M. & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. doi: 10 .1162/0898929041920487

Winter, B. (2019). *Statistics for linguists: An introduction using R*. New York, NY: Routledge. doi: 10.4324/9781315165547

Wiseheart, M., Viswanathan, M. & Bialystok, E. (2016). Flexibility in task switching by monolinguals and bilinguals*. *Bilingualism: Language and Cognition*, *19*(1), 141–146. doi: 10.1017/S1366728914000273

Wong, B., Yin, B. & O'Brien, B. (2016). Neurolinguistics: Structure, function, and connectivity in the bilingual brain. *BioMed Research International*, *2016*, 1–22. doi: 10.1155/2016/7069274

Wood, S. (2021). *Package 'mgcv'.*

Wood Bowden, H., Steinhauer, K., Sanz, C. & Ullman, M. T. (2013). Native-like brain processing of syntax can be attained by university foreign language learners. *Neuropsychologia*, *51*(13), 2492–2511. doi: 10.1016/j.neuropsychologia.2013.09.004

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials (ERPs) in studies of perception and attention. *Attention, perception & psychophysics*, *72*(8), 2031–2046. doi: 10.3758/APP.72.8.2031

Wu, Y. J. & Thierry, G. (2013). Fast modulation of executive function by language context in bilinguals. *Journal of Neuroscience*, *33*(33), 13533–13537. doi: 10.1523/JNEUROSCI.4760-12.2013

Xiong, K., Verdonschot, R. G. & Tamaoka, K. (2020). The time course of brain activity in reading identical cognates: An ERP study of Chinese - Japanese bilinguals. *Journal of Neurolinguistics*, *55*, 1–14. doi: 10.1016/j.jneuroling.2020.100911

Yamasaki, B. L., Stocco, A. & Prat, C. S. (2018). Relating individual differences in bilingual language experiences to executive attention. *Language, Cognition and Neuroscience*, *33*(9), 1128–1151. doi: 10.1080/23273798.2018.1448092

Yip, V. & Matthews, S. (2007). Relative clauses in Cantonese-English bilingual children: Typological challenges and processing motivations. *Studies in Second Language Acquisition*, *29*(2), 277–300.

Zawiszewski, A., Gutiérrez, E., Fernández, B. & Laka, I. (2011). Language distance and non-native syntactic processing: Evidence from event-related potentials. *Bilingualism: Lan-*

guage and Cognition, *14*(3), 400–411. doi: 10 .1017/ S1366728910000350

Zawiszewski, A. & Laka, I. (2020). Bilinguals processing noun morphology: Evidence for the language distance hypothesis from event-related potentials. *Journal of Neurolinguistics*, *55*, 100908. doi: 10.1016/j.jneuroling.2020.100908

Zhang, Q. & Damian, M. F. (2009). The time course of segment and tone encoding in Chinese spoken production: An event-related potential study. *Neuroscience*, *163*(1), 252–265. doi: 10.1016/j.neuroscience.2009.06.015

Zurn, P., Bassett, D. S. & Rust, N. C. (2020). The Citation Diversity Statement: A practice of transparency, a way of life. *Trends in Cognitive Sciences*, *24*(9), 669–672. doi: 10.1016/j.tics.2020.06.009

Nederlandse samenvatting

Het spreken van meer dan één taal heeft een diepgaande invloed op zowel de geest als de hersenen. Maar hoe beheren de meertalige hersenen zowel een moedertaal als een niet-moedertaal? In dit proefschrift heb ik geprobeerd om de meertalige ervaring van late taalleerders op drie manieren te karakteriseren. Ten eerste heb ik het effect van onderlinge beïnvloeding tussen de moedertaal en een andere taal (*cross-linguistic interference*; CLI) op het begrip en het productievermogen in de niet-moedertaal onderzocht. Hierbij was ik vooral geïnteresseerd in hoe overeenkomsten in grammaticale geslachtssystemen en de orthografische en fonologische woordvorm weerslag vinden in hun gedragscorrelaten en neurale correlaten bij late taalleerders. Ten tweede heb ik rechtstreeks verschillende meertalige populaties vergeleken om de effecten van taalgelijkenis op CLI en het begrip en productie van een andere taal te kwantificeren. Ten derde heb ik onderzocht of en hoe taalgelijkenis buiten taalverwerking om van invloed is op hogere cognitieve functies, zoals inhibitoire controle. Dit zijn cruciale vraagstukken omdat zij rechtstreeks betrekking hebben op de wijze waarop de moedertaal en een andere taal in de hersenen naast elkaar bestaan. Bovenden helpen de antwoorden op deze vragen ons om de functionele organisatie van deze talen in het meertalige brein te karakteriseren.

In verschillende studies hebben wij deze kwesties systematisch onderzocht aan de hand van diverse experimentele paradigmata en een combinatie van gedrags- en hersenmetingen. Wij gebruikten bijvoorbeeld elektro-encefalografie om een gedetailleerd inzicht te krijgen in de onderliggende mechanismen van begrip en productie in een niet-moedertaal. Daarnaast hebben we bijzondere aandacht besteed aan de transparantie van onze onderzoeksmethoden, experimentele ontwerpen en procedures in overeenstemming met de beginselen van Open Science. Verder hebben we de statistische aanpak die in dit proefschrift is gebruikt, grondig beschreven. Hierbij gingen we verder dan gebruikelijk voor de analyse van complexe datasets zoals elektro-encefalografische gegevens.

In **hoofdstuk 2** hebben we CLI onderzocht in het begrip van een niet-moedertaal bij Duitstaligen die op late leeftijd Spaans leren. We hebben onderzocht of bepaalde taalkundige overeenkomsten een cumulatief effect kunnen hebben op de taalverwerking, en of late taalleerders gevoelig zijn voor syntactische fouten in grammaticaal geslacht.

Hoofdstuk 3 karakteriseert binnen het kader van CLI de afzonderlijke stadia van niet-moedertalige productie van dezelfde Duits-Spaanse late leerders die in hoofdstuk 2 genoemd zijn. In het bijzonder hebben we onderzocht tot aan welke productiefase en in welke mate anderstaligen met CLI werden geconfronteerd. Hierop voortbouwend hebben we ook gekeken naar de locus van doeltaalselectie tijdens de anderstalige productie.

In **hoofdstuk 4** zijn we overgestapt op een andere meertalige populatie, namelijk moedertaalsprekers van het Italiaans die op latere leeftijd Spaans leren. Dezelfde vragen als in hoofdstukken 2 en 3 worden gesteld om te onderzoeken of de resultaten uit die vorige hoofdstukken ook van toepassing zijn op een combinatie van twee talen die meer op elkaar lijken. In het bijzonder wilden we de impact van CLI op zowel het begrip als de productie van niet-moedertaalsprekers onderzoeken, met behulp van een soortgelijke methodologie als in de vorige hoofdstukken.

In **hoofdstuk 5** hebben we de invloed van taalovereenkomst op het begrip van een niet-moedertaal onderzocht. Bij de vergelijking van begrip van een niet-moedertaal in de Duits-Spaanssprekenden en Italiaans-Spaanssprekenden vroegen we ons af of sprekers van talen die sterk op elkaar lijken een inherent verwerkingsvoordeel hadden ten opzichte van sprekers van minder op elkaar lijkende talen. Hierbij hebben we vooral getracht om een potentiële impact van taalgelijkenis op het begrip van een niet-moedertaal vanuit een neuraal perspectief te beschrijven.

Hoofdstuk 6 is een aanvulling op het vorige hoofdstuk, omdat het het productievermogen in een niet-moedertaal direct vergelijkt met betrekking tot de taalgelijkenis in het geval van de Duits-Spaanse sprekers en de Italiaans-Spaanse sprekers uit het vorige hoofdstuk. We hebben gekeken of sprekers van talen die sterk op elkaar lijken een meetbaar productievoordeel hadden in hun niet-moedertaal in vergelijking met sprekers van talen die minder op elkaar lijken.

In **hoofdstuk 7** gingen we verder dan de rol van taalovereenkomst in begrip en productie in niet-moedertalen. In het bijzonder hebben we namelijk onderzocht of sprekers van sterk op elkaar lijkende talen, bijvoorbeeld Italiaans-Spaans, een algemeen voordeel ontwikkelden met betrekking tot inhibitoire controle in vergelijking met sprekers van talen die minder op elkaar lijken, zoals het Nederlands en het Spaans. Met behulp van een eenvoudige gedragstaak hebben we de prestaties van deze twee groepen op het gebied van inhibitoire controle onderzocht om zo een mogelijke impact van taalgelijkenis op hogere cognitieve functies te onderzoeken.

In **hoofdstuk 8** hebben we de neurale correlaten van taalbegrip bij moedertaalsprekers van het Spaans verkend in het kader van de controverse in de huidige literatuur over de neurale correlaten van de verwerking van grammaticaal geslacht in het Spaans. We hebben hebben zo de primaire neurale correlaten bestudeerd van de verwerking van grammaticale fouten in combinaties van determinatoren en zelfstandige naamwoorden in moedertaalsprekers van het Spaans. Verder hebben we geavanceerde statistische technieken toegepast om robuust bewijs te leveren voor dit specifieke debat.

Curriculum Vitae

Sarah Von Grebmer Zu Wolfsthurn was born on the 23^{rd} of October 1994 in Bruneck/Brunico in the northern Italian province of Bozen/Bolzano. After graduating from the "Sprachen- and Realgymnasium Nikolaus Cusanus" in her hometown, she completed a BSc in Psychology at the University of Bristol (United Kingdom) in 2016. Next, she obtained an MSc in Neuropsychology at the University of Bristol in 2017, taught as a joint programme between the university and the North Bristol National Health Service trust. She then completed several research assistant positions at the University of Bristol, and a research intern position at the Max Planck Institute for Psycholinguistics in Nijmegen (The Netherlands). In September 2018, she joined the Horizon 2020 project "The Multilingual Mind" led by Prof. dr. Theo Marinis as a Marie Skłodowska-Curie Fellow and PhD candidate at the Leiden University Centre for Linguistics (The Netherlands). She was supervised by Prof. dr. Niels O. Schiller and Dr. Leticia Pablos-Robles on her project on the neural correlates of cross-language interactions in the multilingual brain. Her research is primarily focused on non-native language production and comprehension in multilinguals, the role of language similarity on cross-language interactions, as well as language control and executive functions in the multilingual brain. These research interests are also reflected in the present thesis, which contains the main findings of the research conducted within her PhD project.