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.