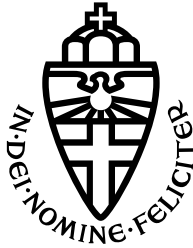


RADBOUD UNIVERSITY NIJMEGEN



FACULTY OF SOCIAL SCIENCES

**A neural network simulation of
event-related potentials:
Bilinguals' response to syntactic
violations differing in
cross-language similarity**

RESEARCH PROJECT MSc ARTIFICIAL INTELLIGENCE

Daily supervisors:
dr. Stefan FRANK
Yung Han KHOE

Author:
Stephan VERWIJMEREN

Internal supervisor:
dr. Hartmut FITZ

Second reader:
dr. Martijn BENTUM

February 2024

Abstract

Event-related potentials (ERPs) are used to study how language is processed in the brain, including differences between native (L1) and second-language (L2) processing. One specific use of ERPs is measuring native-like processing in bilinguals: when the same ERP appears for the same condition in L2 as in L1, it is considered native-like. A P600 effect can be measured in proficient L2 learners in response to an L2 syntactic violation, indicating native-like processing. Cross-language similarity seems to be a factor that modulates P600 effect size. This manifests in a reduced P600 effect in response to a syntactic feature in the L2, where its syntactic construction is different from the syntactic construction in the L1. The precise functional interpretation of ERPs remains a matter of debate. Fitz and Chang (2019) proposed a theory where ERPs reflect learning signals that arise from mismatches in predictive processing. These signals are propagated across the language system to make future predictions more accurate. We test if this theory can account for the mentioned P600 effect reduction in late bilinguals, by implementing a model capable of simulating the P600. We perform an experiment containing three types of syntactic constructions differing in cross-language similarity, designed to elicit a P600 effect in simulated L2 learners progressing through learning stages. Simulated English-Spanish participants displayed a P600 when encountering constructions similar in cross-language similarity. Conversely, simulated English-Spanish participants displayed a reduced P600 when encountering constructions different in cross-language similarity. The difference between these ERP responses of our simulated participants is similar to the difference between ERP responses of participants in human ERP studies. Simulated participants did not however show a clear P600 in response to constructions that were unique to the L2, which is due to the model not being sensitive to the specific construction used, namely a violation in grammatical gender. Our findings partially further support the viability of error propagation as an account of ERP effects, and brought an inability of our model to light.

Keywords: Event-related potential; cross-language similarity; P600; prediction error; bilingualism.

Introduction

Psycholinguistic studies investigating neural mechanisms underlying adult second-language (L2) learning and processing often use electroencephalography (EEG), a technique for recording electrical voltage potentials produced by neural activity. Recorded potentials can be analyzed in relation to cognitive events, and can yield interpretable patterns called event-related potentials (ERPs) (Morgan-Short, 2014). ERP effects have been observed in response to syntactic violations in first language (L1) processing, as an increased positivity in the ERP waveform that starts around 600 ms after observing an anomalous word, as compared to its correct counterpart (Osterhout and Mobley, 1995). This effect is called a P600.

ERP research using the P600 has been done to find out if L2 learners show similar ERP effects as native speakers for morpho-syntactic processing. Research has shown that L2 learners can show native-like ERP effects for L2 grammatical features that are present in their L1 as well as for grammatical features unique to their L2 (Morgan-Short, 2014). ERP effects and their magnitude in L2 learners have been shown to be primarily determined by proficiency (Antonicelli and Rastelli, 2022; Caffarra et al., 2015; McLaughlin et al., 2010; Morgan-Short, 2014).

In bilingual ERP review studies, proficiency is the most important factor determining P600 size, but among other factors, support for influence of L1-L2 similarity is present, often modulating the found effect of proficiency (Antonicelli and Rastelli, 2022; McLaughlin et al., 2010; Morgan-Short, 2014). In Osterhout et al. (2006) for example, English speakers learning French were tested on French linguistic anomalies at three moments in time. The anomalies include subject-verb agreement which is phonologically realized in French, and determiner number agreement which is not phonologically realized and differs in construction between French and English. Regarding their ERP responses, their learning pattern was different. At the first testing moment, French learners showed a N400 for subject-verb agreement violations and were not sensitive to determiner number agreement violations. From the second testing moment forward, French learners showed a consistent P600 for subject-verb agreement. However, even at the third testing moment, after 8 months of French instruction, participants did not show reliable sensitivity to the determiner number agreement violations.

However, Caffarra et al. (2015) performed an empirical review study using logistic regression on multiple factors over ERP studies, and found no significant effect of L1-L2 similarity on participants' ERPs. Caffarra et al. (2015) point out that cross-linguistic similarity cannot always account for the ERP effects observed in L2 syntactic processing. They name the study of Dowens et al. (2011) which stresses the roles of proficiency in L2 syntax processing. In this study, Chinese speakers in their third or fourth year of Spanish show a clear P600 effect in response to Spanish sentences containing violations of number and grammatical gender agreement. The Chinese language is an isolating language in which morphosyntactic features such as gender and number are not computed, having

no similarity with Spanish syntactic rules of gender and number agreement.

To what extent L1-L2 similarity affects ERP effects in bilinguals is unclear. Some individual ERP studies showed reduced P600 effects, or no P600 effect, for syntactic features that are instantiated differently between languages (Antoncelli and Rastelli, 2022; Liu et al., 2017; Morgan-Short, 2014), while other ERP studies have shown P600 effects for syntactic L2 features regardless of L1-L2 similarity (Caffarra et al., 2015; McLaughlin et al., 2010; Morgan-Short, 2014). When L1-L2 similarity is influential, it seems to be complex. It appears that native-like processing, here meaning showing a native-like P600 response, of syntactic L2 features that are unique to the L2 is possible (Foucart and Frenck-Mestre, 2012; McLaughlin et al., 2010; Morgan-Short, 2014), as is native-like processing of syntactic L2 features that are expressed similarly in the L1 and L2 (Foucart and Frenck-Mestre, 2011; McLaughlin et al., 2010; Morgan-Short, 2014). But when a syntactic feature is present in the L1 and the L2, and the syntactic feature is expressed differently in the L2, the P600 seems to be less sensitive to L2 violations of these differently expressed syntactic features (Sabourin and Stowe, 2008; Tokowicz and MacWhinney, 2005).

In Tokowicz and MacWhinney (2005), English speaking Spanish learners were presented with Spanish sentences containing syntactic violations. There were three types of syntactic violations differing in cross-language similarity, resulting in the following three similarity conditions: Similar, Unique and Different. The three types of syntactic violations were tense violation (Similar), determiner gender violation (Unique) and determiner number violation (Different). See Table 1 for example sentences. A sentence with a tense violation contained a verb in the progressive tense without an auxiliary verb. The syntactic construction for this tense agreement is similar between Spanish and English. In a sentence with a determiner gender violation, the gender of a noun phrase was switched to the incorrect opposite gender, resulting in a violation at the following noun. This syntactic construction is unique to Spanish compared to English, since the English language does not express grammatical gender. In a sentence with a determiner number violation, the number of the determiner was switched to the incorrect number, resulting in a violation at the following noun. In both languages, plurality of a noun is expressed by an inflectional morpheme -s at the end of a noun. However, plurality in Spanish is also expressed in the determiner preceding a noun, which makes the syntactic construction different from English.

Computational models of ERP effects

ERPs and factors like L1-L2 similarity that influence ERPs are a useful research tool in psycholinguistic research studying language processing. The precise functional interpretation of ERPs however has remained unclear (Beres, 2017; Kaan, 2007). Computational cognitive models have been proposed to give a functional interpretation of ERPs (Eddine et al., 2022). Many of these models resemble a neural network, specifically a Simple Recurrent Network containing a connection from the previous hidden layer to the hidden layer (Elman, 1990), where input

Table 1: Constructions containing syntactic violations with Spanish example sentences and their English translation. Violations are indicated with an asterisk and critical words are underlined. Table adopted from Tokowicz and MacWhinney (2005).

Construction	Similarity	Example sentence Spanish	English translation
Tense	Similar	Su abuela <u>*cocinando</u> muy bien	His grandmother <u>*cooking</u> very well
Determiner gender	Unique	Ellos fueron a <u>*un fiesta</u>	They went to <u>*a-MASC party</u>
Determiner number	Different	*El <u>niños</u> están jugando	*The-SING boys are playing

words are fed through a network, in a forward direction, to predict the next word. To learn and make future predictions more accurate, the neural networks use backpropagation, where calculated error between the output prediction and the target prediction is backpropagated through a network, changing network connection weights. With this method of error-based learning, models are able to learn syntactic roles and constructions (Chang, 2009; John and McClelland, 1990). Consequently, error-based learning can be a possible explanation of syntactic acquisition.

The computational cognitive models by Brouwer et al. (2017) and Fitz and Chang (2019) are able to model the P600. Specifically, Fitz and Chang (2019) used Chang’s (2002) Dual-path model to show that backpropagated error corresponds to P600 size across a wide range of studies, providing support for the hypothesis that ERPs might reflect learning signals. This account of ERPs is known as the Error Propagation account.

The Dual-path model is a connectionist model of sentence production and syntactic development. The model has two pathways. The first pathway is the sequencing system that learns how words are ordered in a sentence and is based on the Simple Recurrent Network (Elman, 1990). The second pathway is a meaning system that learns how to map message content onto words in a target language. Previously, the Dual-path model was used to explain a wide range of sentence production phenomena in a number of different languages (Chang et al., 2006; Chang, 2009; Chang et al., 2015). For our studies, we used a bilingual extension of the Dual-path model, the Bilingual Dual-path model (Tsoukala et al., 2017).¹ Janciauskas and Chang (2018) used the Bilingual Dual-path model to investigate whether this model’s learning mechanism and the age of L2 acquisition could explain a decreasing performance on English grammaticality tasks by Korean speaking English learners. Using the Bilingual Dual-path model, Tsoukala et al. (2021b) were able to produce Spanish-to-English and English-to-Spanish code-switches. In previous work, the Bilingual Dual-path model specifically was used to simulate ERP responses to syntactic violations in second-language learning (Verwijmeren et al., 2023), where ERP effects of the Bilingual Dual-path resembled those in human subjects, adding further support to the Error Propagation account.

¹https://gitlab.com/yhkhoe/bilingual-dual-path/-/tree/erp_11-12_similarity

The present study

We perform a computational modelling experiment to investigate whether simulated L2 learners are sensitive to cross-language similarity in sentence processing. We do this to further test the viability of Error Propagation as an account of ERPs. Sensitivity will be investigated via measuring a simulated P600 effect in the Bilingual Dual-Path Model (Tsoukala et al., 2021a). In our study specifically we investigate sensitivity to cross-language mismatches, where these mismatches are grammatical structures that do not match across the languages. We simulate the experiment of Tokowicz and MacWhinney (2005), and test their findings on ERP effects corresponding with cross-language similarity. We simulate native speakers of English (L1) who start learning Spanish (L2) from a later age, using the Bilingual Dual-Path Model. At every L2 learning stage, we run a grammaticality agreement experiment similar to the experiment in Tokowicz and MacWhinney (2005), presenting simulated participants with three types of stimuli containing different sentences with grammaticality violations that elicit a P600 in native human speakers, differing in cross-language similarity, and we present simulated participants with control sentences without such violations.

We expect that simulated L2 learners will show sensitivity to grammaticality violations for constructions that are similar in the two languages, and to grammaticality violations that are unique to the L2, because participants in human ERP studies show sensitivity to these L1-L2 similar grammaticality violations and to L2 unique violations as well (Foucart and Frenck-Mestre, 2011, 2012; McLaughlin et al., 2010; Morgan-Short, 2014). Specifically, we expect a clear P600 effect to grammaticality violations expressed similarly in L1 and L2, and a clear P600 effect to grammaticality violations expressed uniquely in L2. We expect that simulated L2 learners will show less sensitivity to violations of constructions that are present in both the L1 and L2 but differ in their expression between the L1 and L2, especially in early learning stages, in line with the findings of Sabourin and Stowe (2008) and Tokowicz and MacWhinney (2005). Specifically, we expect a reduced P600 effect or the absence of a P600 effect to grammaticality violations that are expressed differently between L1 and L2, compared to the two previously mentioned expected P600 effects.

Methods

To simulate late Spanish-English bilinguals, we train the Bilingual Dual-path model (Figure 1) to learn English from “infancy” and Spanish as L2 at a later stage. The training input to the model consisted of sentences from two artificial languages (modelled on Spanish and English) that were paired with messages that encoded their meaning. The model learned to express messages as sentences in the target language (Spanish or English) by predicting the next word. After each training epoch, the model is evaluated to measure language proficiency, and tested in experimental trials to measure simulated ERPs.

For exploratory purpose and L1 control group, we also simulate monolingual

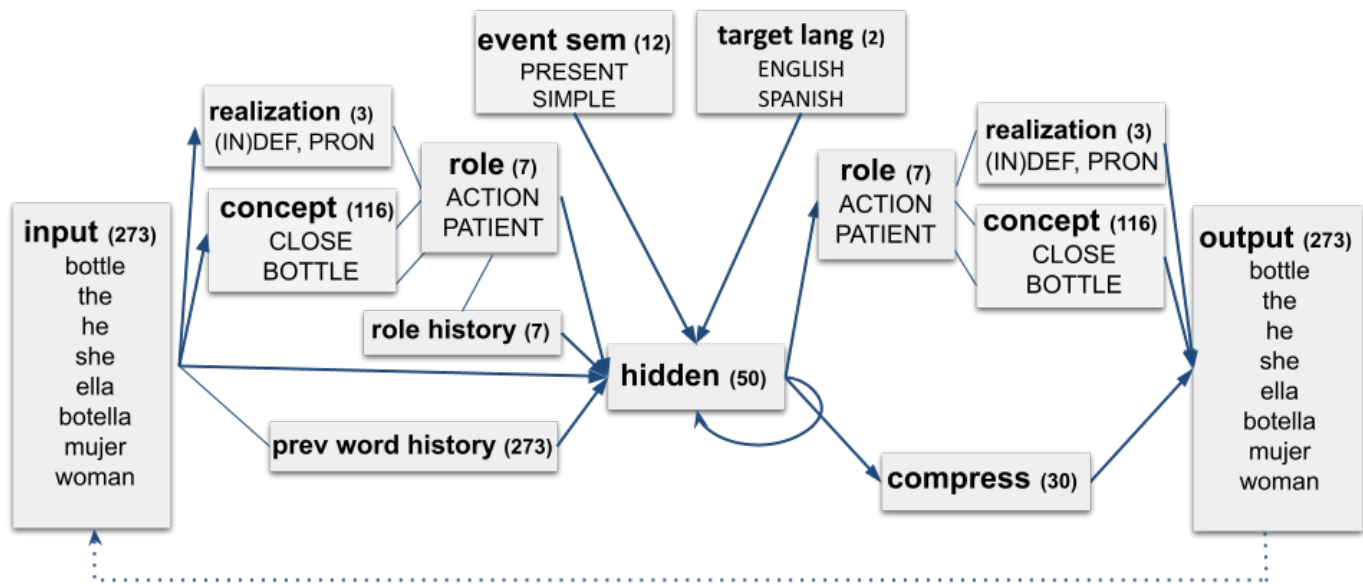


Figure 1: Architecture of the Bilingual Dual-path model. The model learns to map messages onto sentences in different languages by predicting the next word in its input. The sequencing system (lower path) maps from the input through a hidden layer to the output via a compression layer. The meaning system (upper path) uses information about thematic roles, concepts, and the realization of concept (e.g., by a pronoun or with an (in)definite determiner). The number of units per layer are shown in parentheses. Arrows between layers indicate trainable weight connections, and lines between layers indicate fixed-weight connections. Figure adapted from Tsoukala et al. (2021a).

Spanish participants. The model subjects are trained, except that they received only Spanish. The model subjects are evaluated and tested in experimental trials in the same fashion as the bilingual model subjects.

Artificial languages

Table 2 shows the different constructions in the artificial languages. Taken together, the two artificial languages consisted of 259 lexical items: 121 nouns, 11 adjectives, 6 pronouns, 6 determiners, 12 prepositions, 87 verbs, 8 auxiliary verbs, 6 verb inflectional morphemes, 1 plural noun marker, and the period. The inflectional morphemes were used to generate verbs with simple, progressive and perfect aspect in present or past tense. The plural noun marker was used to generate plural nouns.

The meaning space had 116 concepts and 7 thematic roles. Thematic roles are similar to those from Chang et al. (2006). For example, the semantic message: AGENT: LADY; ACTION-LINKING: CARVE; PATIENT: CAKE; AGENT-MODIFIER: OLD; TARGET-LANGUAGE: EN would be expressed in English by the

Table 2: Constructions with English example sentences. In the artificial language modelled on English, inflectional morphemes -prg, -prf and -ss are used for verb conjugations in progressive, perfect, and 3rd-person present simple tense, respectively.

Construction	Example sentence English	Example sentence Spanish
Animate intransitive	The woman is play -prg	La mujer está jugar -ger
Animate with intransitive	The woman is play -prg with a dog	La mujer está jugar -ger con un perro
Inanimate intransitive	The apple is fall -prg	La manzana está caer -ger
Locative	The boy is walk -prg around the school	El niño está caminar -ger alrededor la escuela
Theme-experiencer (active)	The uncle surprise -ss the grandfather	El tío sorprender -a-e el abuelo
Theme-experiencer (passive)	The grandfather is surprise -prf by the uncle	El abuelo está sorprender -par por el tío
Transitive (active)	The girl bake -ss a cake	La niña hornear -a-e un pastel
Transitive (passive)	The cake is bake -prf by the girl	El pastel está hornear -par por la niña
Cause-motion	The hostess is put -prg a cactus into the office	La anfitriona está poner -ger un cacto dentro la oficina
Benefactive transitive	The grandmother repair -ss the cup for the girl	La abuela reparar -a-e la taza para la niña
State-change	The waiter is fill -prg the cup with water	El camarero está llenar -ger la taza con agua
Locative alternation	The man spray -ss the sink with water	El hombre rociar -a-e el lavabo con agua

sentence: “the old lady carves a cake”. The semantic message AGENT: ORANGE, PL; ACTION-LINKING: DISAPPEAR; TARGET-LANGUAGE: ES would be expressed in Spanish by the sentence: “las naranja -s desaparecer -an-en”.

Model configuration

For our simulations, we modified the Bilingual Dual-path model to resemble the architecture used in Fitz and Chang (2019): Previous word-history and role-history layers were added to the Bilingual Dual-path model by Tsoukala et al. (2017), which kept a running average of the activation of the input layer and role layer, respectively, and were connected to the hidden layer. The input layer is directly connected to the hidden layer. In configurations of the Bilingual Dual-path model in other works (Tsoukala et al., 2017, 2021a,b; Khoe et al., 2023), the input layer is not directly connected to the hidden layer, but rather connected to another compress layer, which is then connected to the hidden layer.

As pre-registered², all models used 50 hidden-layer units and 30 compress-layer units. Internal layer units used the logistic activation function; the output layer units used a softmax activation function. Weights were initialized randomly, uniformly between ± 1 . Thematic roles are implemented by introducing fixed-weight connections between role units and concept units (see Figure 1). Fixed weights for concept-to-role connections and realization-to-role connections were set to a value of 6. The concept layer had a set bias of -3 . Following Fitz and Chang (2019), the input for the model is set to the single highest activation value of the sum of the produced vector and the target vector, to emphasize correct word prediction.

Model training

As pre-registered², for each of 60 model subjects, for English and Spanish combined, we generated 10,000 unique message-sentence pairs for bilingual training and a novel set of 200 message-sentence pairs for testing. The sentences are approximately equally divided over the two languages, where the percentage of English sentences was sampled from a uniform distribution between 48% and 52% and the rest was Spanish. Sentence constructions were distributed uniformly in the training input. Following Fitz and Chang (2019), the message was excluded from 70% of the training items. Each model first iterated five times over its monolingual English training set, followed by 45 epochs over its bilingual training set. The training set’s order was randomized at the beginning of each of these 50 epochs. The model learned by steepest descent backpropagation, with momentum set to 0.9. Initially, the learning rate was set to 0.1, it decreased linearly to 0.02 over the 5 epochs of monolingual training, and then stayed constant during bilingual training.

Model evaluation

After each epoch, model accuracy was tested using a 200-sentence test set. The model’s L1 and L2 proficiency was evaluated with two measures. Following Tsoukala et al. (2021a), syntactic accuracy was measured as the percentage of sentences for which all words had the correct part of speech. Second, meaning accuracy was measured as the percentage of syntactically correct sentences that also conveyed the target message without additions. As pre-registered², we excluded the 20 subjects with the lowest meaning accuracy, leaving data from 40 model subjects.

Experimental trials

We generated 30 Spanish control sentences to obtain simulated ERPs on. For each of the control sentences we generated a sentences for every violation type, and compared them to their respective control sentence. For examples, see Table 3. The control sentence was a syntactically correct active transitive sentence.

²The pre-registration can be accessed here: <https://aspredicted.org/iu4x7.pdf>

Table 3: Example sentences for the experimental trials. The bold morphemes indicate the sentence position where prediction error was measured. To calculate error differences, prediction error in the control sentence was calculated at the same positions as in the violated sentences.

Example sentence	Violation type	L1-L2 similarity
El padre hacer -a-e una bañera	Control	Control
El padre hacer -ger una bañera	Tense	Similar
Los padre hacer -a-e una bañera	Determiner number	Different
La padre hacer -a-e una bañera	Determiner gender	Unique

We had three violation types: tense, determiner number and determiner gender. For the tense violation, the verb did not agree in tense. Tense violations were created by changing the inflectional marker for singular verbs (-a-e) to the inflectional marker for progressive verbs (-ger). For the determiner number violation, the first determiner did not agree in number with the following noun. Determiner number violations were created by changing the singular determiner to a plural determiner. For the determiner gender violation, the first determiner did not agree in grammatical gender with the following noun. Determiner gender violations were created by changing the determiner to a determiner of the opposite grammatical gender.

Model subject differences

Weights are initialized randomly, and differed between subjects. The percentage of English versus Spanish (training and testing) sentences varied between subjects, ranging from 48/52 to 52/48. The distribution of constructions is the same for all subjects. Training, testing and experimental trial sentences in the same language with the same constructions can differ between subjects in two ways. Firstly, sentences can differ in content-words resulting in different meaning of sentences. Consequently a different content-word can result in a different grammatical gender of a noun phrase. Secondly, singular nouns that are direct objects can differ in definiteness of the article.

Measuring model ERPs

After every training epoch, the model was tested on the experimental sentences. As in Fitz and Chang (2019), learning was turned on in the model during processing, but connection weights were reset to the weights of the respective training epoch after each test sentence in order to exclude learning effects during the experiment. The state in which the model encountered each trial was thus the same for all of the sentences.

We measured the prediction error at the hidden layer (see Fitz and Chang, 2019, for details). The prediction error of output unit j is the difference between its activation y_j and the target activation t_j , or: $\delta_j = y_j - t_j$, with $y_j \in [0, 1]$ and

$t_j \in \{0, 1\}$. This error was backpropagated through the network, as happens during training, to generate error at deeper layers. Error for units connected to the output layer was calculated as shown in Eq. 1, where k indexes the units connected to the output layer with weight w_{kj} , and j references the units that are backpropagating error.

$$\delta_k = y_k(1 - y_k) \sum_{j=1}^n \delta_j w_{kj} \quad y_k \in [0, 1] \quad (1)$$

Error was calculated the same for other layers backpropagating error through the network. The simulated P600 sizes are the sums over $|\delta|$ of the hidden-layer units, respectively. The error in a violated sentence was collected at the first position where its word results in a violation. The error in a control sentence was collected at the same position as the violated sentence for comparison.

Results

Monolingual model subjects

Mean Spanish meaning accuracy and mean Spanish syntactic accuracy were 99.98% and 99.99%, respectively, at the end of training.

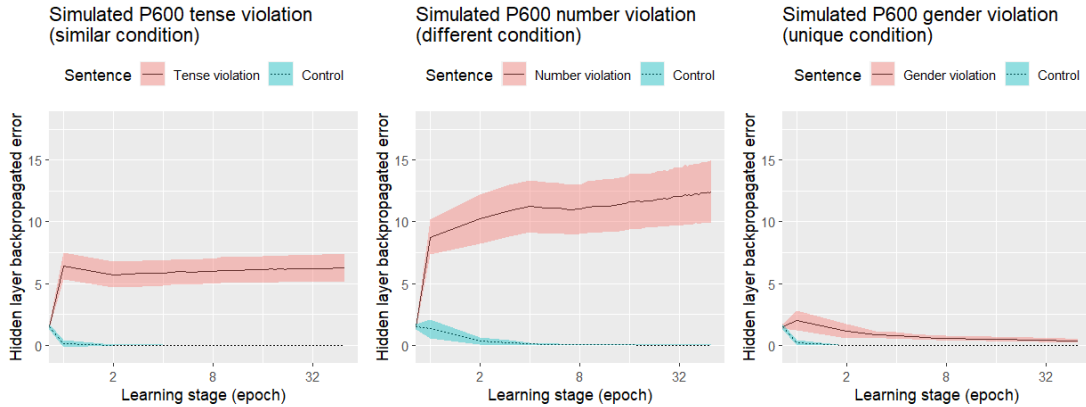


Figure 2: Mean backpropagated error (averaged over all monolingual trained model subjects) as a function of learning stage in the hidden layer, split between the three violation types. Learning stage is log-scaled. Shaded areas represent the 95% CI computed over items.

The mean prediction errors over L2 learning stages at the hidden layer are displayed in Figure 2. In response to a tense violation, the mean error increased to 6.26 at the end of monolingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a tense violation was 1.51 at the start of monolingual training, which declined to 0.00 at the end of monolingual training.

In response to a determiner number violation, the mean error increased to 12.45 at the end of monolingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a tense violation was 1.51 at the start of monolingual training, which declined to 0.00 at the end of monolingual training. In response to a violation, error size is larger in the Different condition compared to the Similar condition. In response to a control item, for both conditions, error declined to 0.00.

In response to a determiner gender violation, the mean error increased to 2.02 after the first epoch, and declined to 0.35 the end of monolingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a tense violation was 1.51 at the start of monolingual training, which declined to 0.00 at the end of monolingual training. In response to a violation, error size is very low compared to the other two conditions. Therefore, the difference in error between control and violation items is much smaller compared to our other two conditions.

Table 4: Summary of the components in the generalized additive mixed-effects model fit on data from monolingual participants, comparing L1-L2 similarity conditions Different and Similar.

Component	Predictor	Est.	SE	<i>t</i> -value	Pr(> <i>t</i>)
parametric coefficients	(Intercept)	8.74	0.34	26.04	<0.001
	NOT_SIMILAR	-3.83	0.47	-8.07	<0.001
Component	Predictor	edf	Ref.df	<i>F</i> -value	Pr(> <i>t</i>)
smooth terms	s(LEARNING_STAGE)	8.89	8.90	1141.37	<0.001
	s(LEARNING_STAGE:NOT_SIMILAR)	8.96	9.00	488.73	<0.001
	s(LEARNING_STAGE, participant)	314.73	359.00	30.75	<0.001
	s(NOT_SIMILAR, participant)	77.95	78.00	1655.87	<0.001

Table 5: Summary of the components in the generalized additive mixed-effects model fit on data from monolingual participants, comparing L1-L2 similarity conditions Different and Unique.

Component	Predictor	Est.	SE	<i>t</i> -value	Pr(> <i>t</i>)
parametric coefficients	(Intercept)	5.97	0.31	19.15	<0.001
	NOT_UNIQUE	7.74	0.44	17.57	<0.001
Component	Predictor	edf	Ref.df	<i>F</i> -value	Pr(> <i>t</i>)
smooth terms	s(LEARNING_STAGE)	8.84	8.86	301.10	<0.001
	s(LEARNING_STAGE:NOT_UNIQUE)	8.99	9.00	1864.80	<0.001
	s(LEARNING_STAGE, participant)	313.84	359.00	25.76	<0.001
	s(NOT_UNIQUE, participant)	77.96	78.00	1911.26	<0.001

Exploratory analysis

Similar to our pre-registered analysis, we analyzed the data from our experiment with two generalized additive mixed-effects models (GAMMs) (Hastie, 2017), using the `bam` function from the package `mgcv` (Wood and Wood, 2015) in R (R Core Team, 2013). We fit a GAMM to determine if participants respond differently between L1-L2 similarity conditions Similar and Different, and we fit a second GAMM to determine if model participants respond differently between L1-L2 similarity conditions Unique and Different. Both GAMMs fit the difference in prediction error, a numerical value, from the Bilingual Dual-path model, here trained only on Spanish input. The first model³ included the predictors of interest: `NOT_SIMILAR`, `LEARNING_STAGE`, and their interaction. `NOT_SIMILAR` and was dummy-coded. `NOT_SIMILAR` levels Similar and Different were coded 0 and 1, respectively. The number of training epochs is indicated by the `LEARNING_STAGE` predictor, which was standardized. We include by-participant random slopes for `NOT_SIMILAR` and by-participant random smooths for `LEARNING_STAGE`. See Table 4 for a summary of the GAMM. We report estimates, standard errors, t -values and p -values for the parametric coefficients. We report estimated degrees of freedom, reference degrees of freedom, F -values and p -values for the smooth terms. Error difference shows a significantly different from 0 trend over `LEARNING_STAGE` ($F = 1141.37$, $\text{edf} = 8.61$, $p < .001$), and a significantly different from 0 interaction between `LEARNING_STAGE` and `NOT_SIMILAR` ($F = 488.73$, $\text{edf} = 8.39$, $p < .001$). There is a larger simulated P600 effect for the Different condition compared to the Similar condition.

The second GAMM³ included the predictors of interest: `NOT_UNIQUE`, `LEARNING_STAGE`, and their interaction. `NOT_UNIQUE` and was dummy-coded. `NOT_UNIQUE` levels Unique and Different were coded 0 and 1, respectively. The number of training epochs is indicated by the `LEARNING_STAGE` predictor, which was standardized. We include by-participant random slopes for `NOT_SIMILAR` and by-participant random smooths for `LEARNING_STAGE`. See Table 5 for a summary of the GAMM. We report estimates, standard errors, t -values and p -values for the parametric coefficients. We report estimated degrees of freedom, reference degrees of freedom, F -values and p -values for the smooth terms. Error difference shows a significantly different from 0 trend over `LEARNING_STAGE` ($F = 301.10$, $\text{edf} = 7.44$, $p < .001$), and a significantly different from 0 interaction between `LEARNING_STAGE` and `NOT_UNIQUE` ($F = 1864.80$, $\text{edf} = 8.79$, $p < .001$). With respect to error values in the other two conditions, the simulated P600 effect in the Unique condition is very weak.

Bilingual model subjects

Figure 3 displays the proficiency of the model at the start and the end of bilingual training. At the start of bilingual training, mean English meaning accuracy and mean English syntactic accuracy are 77.99% and 92.32%, respectively.

³The script for the GAMMs can be accessed here: https://osf.io/nbxu6/?view_only=f0edeab7fd8e40bfa5310bb234de940e

These increase to 86.59% and 93.53%, respectively, at the end of bilingual training. At the start of bilingual training, mean Spanish meaning accuracy and mean Spanish syntactic accuracy are 0.29% and 1.84%, respectively. These increase to 61.36% and 85.65% respectively at the end of bilingual training.

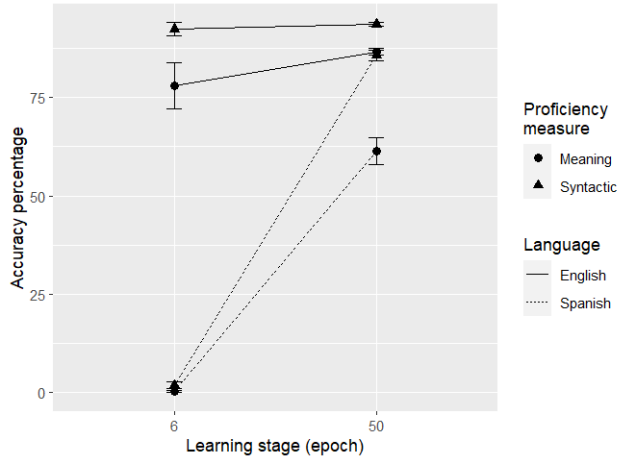


Figure 3: Mean proficiency of the bilingual model. The syntactic and meaning accuracy are displayed for the first and last epoch of bilingual training. The error bars show the 95% confidence interval.

The mean prediction error over L2 learning stages at the hidden layer are displayed in Figure 4. In response to a tense violation, the mean error was 2.63 at the start of bilingual training, and increased to 15.22 at the end of bilingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a tense violation was 1.79, which increased to 3.40 at epoch 22, and then declined to 1.15 at the end of bilingual training.

In response to a determiner number violation, the mean error was 5.45 at the start of bilingual training, and increased to 12.83 at epoch 20. At epoch 39 the mean error was 13.13, where after it declined to 11.49 at the end of bilingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a determiner number violation was 1.29, which increased to 4.11 at the end of bilingual training. In response to a violation in later learning stages, error size is smaller in the Different condition compared to the Similar condition. In response to a control item, error size increases in the Different condition, where in the Similar condition error size declines after the 22nd epoch.

In response to a gender violation, the mean error was 0.91 at the start of bilingual training, and increased to 4.24 at the end of bilingual training. In a control sentence, the mean error in response to the word at the same position error is measured for a gender violation was 0.59, which declined to 0.36 at

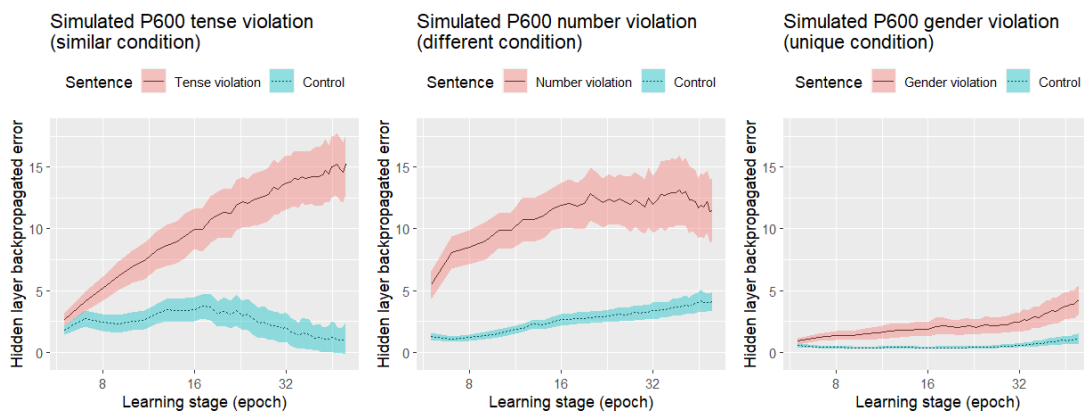


Figure 4: Mean backpropagated error (averaged over all bilingual trained model subjects) as a function of learning stage in the hidden layer, split between the three violation types. Learning stage is log-scaled. Shaded areas represent the 95% CI.

epoch 16, and then increased 1.09 at the end of bilingual training. In response to a violation, error size is low compared to the other two conditions, similar to the error size in the experiment with monolingually trained participants. The difference in error between control and violation items is smaller compared to our other two conditions.

Table 6: Summary of the components in the generalized additive mixed-effects model fit on data from bilingual participants, comparing L1-L2 similarity conditions Different and Similar.

Component	Predictor	Est.	SE	<i>t</i> -value	Pr(> <i>t</i>)
parametric coefficients	(Intercept)	9.12	0.27	33.30	<0.001
	NOT_SIMILAR	0.70	0.39	1.81	0.0701
Component	Predictor	edf	Ref.df	<i>F</i> -value	Pr(> <i>t</i>)
smooth terms	s(LEARNING_STAGE)	8.61	8.72	33.60	<0.001
	s(LEARNING_STAGE:NOT_SIMILAR)	8.39	8.89	2202.45	<0.001
	s(LEARNING_STAGE, participant)	295.03	359.00	48.34	<0.001
	s(NOT_SIMILAR, participant)	77.83	78.00	447.96	<0.001

Pre-registered analysis

As pre-registered, we analyzed the data from our experiment with two generalized additive mixed-effects models (GAMMs) (Hastie, 2017), using the bam function from the package mgcv (Wood and Wood, 2015) in R (R Core Team, 2013). We fit a GAMM to determine if participants respond differently between L1-L2 similarity conditions Similar and Different, and we fit a second GAMM

Table 7: Summary of the components in the generalized additive mixed-effects model fit on data from bilingual participants, comparing L1-L2 similarity conditions Different and Unique.

Component	Predictor	Est.	SE	<i>t</i> -value	Pr(> <i>t</i>)
parametric coefficients	(Intercept)	5.26	0.267	20.83	<0.001
	NOT_UNIQUE	4.76	0.31	15.27	<0.001
Component	Predictor	edf	Ref.df	<i>F</i> -value	Pr(> <i>t</i>)
smooth terms	s(LEARNING_STAGE)	7.44	7.78	8.94	<0.001
	s(LEARNING_STAGE:NOT_UNIQUE)	8.79	8.98	334.19	<0.001
	s(LEARNING_STAGE, participant)	307.02	359.00	2748.53	0.05
	s(NOT_UNIQUE, participant)	68.57	78.00	283.45	<0.001

to determine if model participants respond differently between L1-L2 similarity conditions Unique and Different. Both GAMMs fit the difference in prediction error from the Bilingual Dual-path model, a numerical value. The first GAMM³ included the predictors of interest: NOT_SIMILAR, LEARNING_STAGE, and their interaction. NOT_SIMILAR and was dummy-coded. NOT_SIMILAR levels Similar and Different were coded 0 and 1, respectively. The number of L2 training epochs is indicated by the LEARNING_STAGE predictor, which was standardized. We include by-participant random slopes for NOT_SIMILAR and by-participant random smooths for LEARNING_STAGE. See Table 6 for a summary of the GAMM. We report estimates, standard errors, *t*-values and *p*-values for the parametric coefficients. We report estimated degrees of freedom, reference degrees of freedom, *F*-values and *p*-values for the smooth terms. Error difference shows a significantly different from 0 trend over LEARNING_STAGE ($F = 33.60$, edf = 8.61, $p < .001$), and a significantly different from 0 interaction between LEARNING_STAGE and NOT_SIMILAR ($F = 2302.45$, edf = 8.39, $p < .001$). We see a clear P600 effect in the Similar condition, and a P600 effect in the Different condition. The P600 effect in the Different condition is reduced compared to the P600 effect in the Similar condition, in line with our expectations, and shows a different trend over learning stages.

The second GAMM³ included the predictors of interest: NOT_UNIQUE, LEARNING_STAGE, and their interaction. NOT_UNIQUE and was dummy-coded. NOT_UNIQUE levels Unique and Different were coded 0 and 1, respectively. The number of L2 training epochs is indicated by the LEARNING_STAGE predictor, which was standardized. We include by-participant random slopes for NOT_SIMILAR and by-participant random smooths for LEARNING_STAGE. See Table 7 for a summary of the GAMM. We report estimates, standard errors, *t*-values and *p*-values for the parametric coefficients. We report estimated degrees of freedom, reference degrees of freedom, *F*-values and *p*-values for the smooth terms. Error difference shows a significantly different from 0 trend over LEARNING_STAGE ($F = 8.94$, edf = 7.44, $p < .001$), and a significantly different from 0 interaction between LEARNING_STAGE and NOT_UNIQUE ($F = 334.19$, edf = 8.79, $p < .001$). We see a weak P600 effect in the Unique condition, which is smaller than the

P600 effect in the Different condition. With respect to error values in the other two conditions, this simulated P600 effect in the Unique condition is weak. A smaller effect in the Unique condition compared to the Different condition is not in line with our expectations.

Discussion

In the present work, we investigated whether simulated L2 learners are sensitive to cross-language similarity in sentence processing. We used a cognitive computational model (Chang, 2002) to simulate English-Spanish bilinguals and exposed the model throughout L2 learning to three types of syntactic L2 violations differing in cross-language similarity. We recorded simulated P600s in response to these syntactically anomalous sentences from the model by calculating propagated prediction error at the hidden layer, similar to the Error Propagation account in Fitz and Chang (2019). On this account, ERPs are summary signals of brain activity that index the propagation of prediction error during comprehension whose functional role is to support learning. The results of our bilingual simulations are only partially in alignment with our expectations. Our results reveal a clear P600 effect in the Similar condition, where syntactically anomalous sentences in the L2 contain a tense violation, and a P600 effect in the Different condition, where syntactically anomalous sentences in the L2 contain a number violation. Prediction error in the Unique condition, where anomalous sentences in the L2 contain a gender violation, was higher than prediction error of the corresponding control sentence, but the difference in error can hardly be called a P600, especially compared to the error sizes in response to our other two types of syntactic violations.

A reduced P600 effect in response to a violation in the syntactic construction that is expressed differently (number violation) between L1 and L2, compared to the P600 effect in response to a violation in the syntactic construction that is expressed similarly (tense violation), is in line with the human ERP study by Sabourin and Stowe (2008), although languages and types of syntactic constructions are different. Comparing the error values in response to a tense violation and error values in response to a number violation, our results approach findings in the human ERP study by Tokowicz and MacWhinney (2005), namely that in our work a violation in the syntactic construction that is expressed differently between L1 and L2 (number violation) elicits a reduced P600 effect in comparison to a violation in the syntactic construction that is expressed similarly between L1 and L2 (tense violation). In our model, these results support the influence of cross-language similarity on L2 processing, as well as syntactic development. However, in Tokowicz and MacWhinney (2005) the mean activation amplitude in response to a violation in the syntactic construction that is expressed differently between L1 and L2 (number violation) is lower than the mean error value in response to control sentences, which indicates there is no P600 effect. In our work, the simulated P600 effect is still present, but reduced. Sabourin and Stowe (2008) touched upon the lack of L1 control group in Tokow-

icz and MacWhinney (2005), that could have been useful to test if stimuli elicit a consistent P600 in native L1 speakers. Sabourin and Stowe (2008) suggested that certain constructions are less consistent in eliciting a P600, and that it is possible P600 effects in some constructions are lost in L2 processing noise.

The violation in the syntactic construction that is expressed uniquely in L2 (gender violation) elicits a small P600 effect in our model. However, late bilingual learning is not the cause of this lack of a P600 effect. Looking at our monolingual simulations, only trained on Spanish language input, simulated participants also do not elicit a clear P600 in response to a sentence containing a gender violation. It is not entirely clear why prediction error is low in response to a gender violation. A possible explanation is the implementation of syntactic features in the model. While training, the model receives a message which the upper path of the model maps onto words. In this message, tense as well as plurality of nouns is implemented, but gender is not. Grammatical gender is present and expressed in our artificial language of Spanish, but there is no representation of gender in the event-semantics layer of the Bilingual Dual-path model. Verb conjugation indicating tense as well as plurality of nouns are implemented by including certain particles that follow verbs or nouns. In the model these particles are words on their own. We have no such particles for gender, only separate gendered determiners for Spanish. While during our experiment simulated L2 learners do not receive a message as input, connection weights can be influenced during learning with meaning containing messages.

The error propagation account has been able to reproduce key findings from a considerable number of monolingual ERP studies (Fitz and Chang, 2019). Previous work on simulating bilingual ERPs and how they change over development (Verwijmeren et al., 2023) added further support to his account. In our present work results vary. The increase in P600 size in response to a tense violation further supports a theory of stages of syntactic learning in L2 learners where the magnitude of different ERP components changes during acquisition (McLaughlin et al., 2010). The reduced P600 in the Different condition, where the syntactic construction that is expressed differently between L1 and L2, compared to the P600 in the similar condition, supports a theory of cross-language similarity affecting ERP effects in L2 learners. The model in its present state however is unable to produce a clear P600 in response to a grammatical gender violation, which is in contrast with human participants (Antonicelli and Rastelli, 2022; Caffarra et al., 2015; McLaughlin et al., 2010; Foucart and Frenck-Mestre, 2011; Frenck-Mestre et al., 2009; Morgan-Short, 2014; Tokowicz and MacWhinney, 2005). Further work is needed to determine if the Bilingual Dual-path model together with the Error Propagation account is able to simulate a P600 effect in response to a grammatical gender violation. Grammatical gender is not implemented in the event-semantics layer of the upper path of the Bilingual Dual-path model, while plurality and tense are implemented. Using the semantic gender of certain nouns in our artificial language modelled on Spanish is a setting that can be turned on in the Bilingual Dual-path model, where semantic gender can be used as message content to be mapped onto words. This setting was not turned on in our experiments. Performing a similar experiment with a

determiner gender violation, turning on the setting of semantic gender and having experimental sentences containing nouns with a semantic gender, can shed light on if the mapping of message content onto words plays an important role in backpropagated prediction error size. Another but similar approach would be to implement grammatical gender in the event-semantics layer of the upper path of the Bilingual Dual-path model, and to enhance our artificial language by implementing more inflectional particles for syntactic marking, including grammatical gender.

The error propagation account explains why ERPs elicited by lexical violations (N400) precede ERPs in response to syntactic violations (P600) and this account has been able to reproduce key findings from a considerable number of monolingual ERP studies (Fitz and Chang, 2019). One of the findings from our results on the influence of cross-language similarity on ERPs, adds further support for this account, while the rest of the results pose further questions on how syntactic rules are implemented in the Bilingual Dual-path model and if it is able to produce a clear P600 effect for grammatical gender violations. Apart from the error propagation account, the model of Brouwer et al. (2017) can also explain monolingual N400 and P600 effects but it remains to be tested whether this model would be able to simulate ERP effects in bilinguals and the change in size of these effects during second language acquisition. What is unique about the error propagation account is that it can naturally model and explain ERPs in development because on this account ERPs are directly linked to learning. Therefore, the magnitude of ERP effects is expected to change as different pieces of linguistic knowledge are acquired. One limitation of the model is that it currently does not account for differences in the precise onset of the N400 or P600 and that it does not model earlier ERP components such as the early left-anterior negativity (eLAN) which has been elicited in some bilingual studies (Caffarra et al., 2015).

References

- Antonicelli, G. and Rastelli, S. (2022). Event-related potentials in the study of L2 sentence processing: A scoping review of the decade 2010-2020. *Language Acquisition*, pages 1–38.
- Beres, A. M. (2017). Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research. *Applied psychophysiology and biofeedback*, 42:247–255.
- Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, 41:1318–1352.
- Caffarra, S., Molinaro, N., Davidson, D., and Carreiras, M. (2015). Second language syntactic processing revealed through event-related potentials: An empirical review. *Neuroscience & Biobehavioral Reviews*, 51:31–47.
- Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, 26(5):609–651.
- Chang, F. (2009). Learning to order words: A connectionist model of heavy np shift and accessibility effects in japanese and english. *Journal of Memory and Language*, 61(3):374–397.
- Chang, F., Baumann, M., Pappert, S., and Fitz, H. (2015). Do lemmas speak German? a verb position effect in German structural priming. *Cognitive Science*, 39(5):1113–1130.
- Chang, F., Dell, G. S., and Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2):234.
- Dowens, M. G., Guo, T., Guo, J., Barber, H., and Carreiras, M. (2011). Gender and number processing in chinese learners of spanish—evidence from event related potentials. *Neuropsychologia*, 49(7):1651–1659.
- Eddine, S. N., Brothers, T., and Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. *The Psychology of Learning and Motivation*, page 123.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.
- Fitz, H. and Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15–52.
- Foucart, A. and 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.

- Foucart, A. and 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.
- Frenck-Mestre, C., Foucart, A., Carrasco-Ortiz, H., and Herschensohn, J. (2009). Processing of grammatical gender in French as a first and second language: Evidence from ERPs. *Eurosla yearbook*, 9(1):76–106.
- Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.
- Janciauskas, M. and Chang, F. (2018). Input and age-dependent variation in second language learning: A connectionist account. *Cognitive science*, 42:519–554.
- John, M. F. S. and McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial intelligence*, 46(1-2):217–257.
- Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and linguistics compass*, 1(6):571–591.
- Khoe, Y. H., Tsoukala, C., Kootstra, G. J., and Frank, S. L. (2023). Is structural priming between different languages a learning effect? modelling priming as error-driven implicit learning. *Language, Cognition and Neuroscience*, 38(4):537–557.
- Liu, H., Dunlap, S., Tang, Y., Lu, Y., and Chen, B. (2017). The modulatory role of L1 and L2 morphosyntactic similarity during production of L2 inflected words: An ERP study. *Journal of Neurolinguistics*, 42:109–123.
- McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., and Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, 60:123–150.
- Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics*, 34:15–36.
- Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., and 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. and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34(6):739–773.
- R Core Team (2013). Core R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. Online: <http://www.R-project.org>*, 201.

- Sabourin, L. and Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, 24(3):397–430.
- Tokowicz, N. and 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.
- Tsoukala, C., Broersma, M., Van den Bosch, A., and Frank, S. L. (2021a). Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior*, 4(1):87–100.
- Tsoukala, C., Frank, S. L., and Broersma, M. (2017). “He’s pregnant”: Simulating the confusing case of gender pronoun errors in L2 English. In *the 39th Annual Meeting of the Cognitive Science Society*, pages 3392–3397. Cognitive Science Society.
- Tsoukala, C., Frank, S. L., Van Den Bosch, A., Kroff, J. V., and Broersma, M. (2021b). Modeling the auxiliary phrase asymmetry in code-switched spanish–english. *Bilingualism: Language and Cognition*, 24(2):271–280.
- Verwijmeren, S., Frank, S. L., Fitz, H., and Khoe, Y. H. (2023). A neural network simulation of event-related potentials in response to syntactic violations in second-language learning.
- Wood, S. and Wood, M. S. (2015). Package ‘mgcv’. *R package version*, 1(29):729.