EXPLICIT VS. IMPLICIT ACQUISITION OF GRAMMATICAL GENDER IN AN ARTIFICIAL LANGUAGE BY MONOLINGUAL NATIVE ENGLISH SPEAKERS: AN ERP STUDY

by

Emily M. Jarvis

Submitted in partial fulfillment of the requirements for the degree of

Masters of Science

at

Dalhousie University

Halifax, Nova Scotia

July 2017

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vii

ABSTRACT viii

LIST OF ABBREVIATIONS USED viiii

CHAPTER 1: INTRODUCTION 1

1.1 Second Language Acquisition 1

1.2 Grammatical Gender 4

1.3 Implicit vs. Explicit Grammar Instruction 8

1.4 Behavioural Methods of Assessing Gender Learning 10

1.5 Physiological Methods for Assessing Gender Learning 12

1.6 The current study 17

1.7 Hypotheses 18

CHAPTER 2: METHODS 21

2.1 Participants 21

2.2 Materials 22

2.2.1 Language and Education Survey 22

2.2.2 Gender Association Task 23

2.2.3 Two Alternative Forced Choice Learning Task 24

2.2.4 Match/ Mismatch ERP task 25

2.2.5 Naming Task 27

2.3 Procedure 28

2.4 EEG Acquisition and Data Analysis 28

CHAPTER 3: RESULTS 32

3.1 Learning Assessment 32

3.1.1 Naming Task 32

3.1.2 Match-Mismatch Task: Accuracy Ratings 32

3.2 EEG Analysis 34

3.2.1 300 - 500 ms 34

3.2.1.1 Nouns 34

3.2.1.2 Determiners 36

3.2.2: 700 - 900 ms 38

3.2.2.1 Nouns 38

3.2.2.2 Determiners 40

3.3 GAT Analyses 47

CHAPTER 4: DISCUSSION 49

4.1 Research Questions 49

4.1.1 Question 1 50

4.1.2 Question 2 52

4.1.3 Question 3 54

4.2 Limitations and Future Directions 55

4.3 Conclusions 57

References 59

Appendix A 66

Appendix B 69

Appendix C 74

LIST OF TABLES

Table 3.1. Summary of linear mixed effects model for nouns at 300-500ms……..35

Table 3.2. Summary of linear mixed effects model for determiners at 300-500ms…….37

Table 3.3. Summary of linear mixed effects model for nouns at 700-900ms……..39

Table 3.4. Summary of linear mixed effects model for nouns at 700 - 900 ms……...41

Table 3.5. Summary of ANOVA model for GAT pre- and post- training……...48

LIST OF FIGURES

Figure 2.1. Example of congruent and incongruent GAT pairs.....24

Figure 2.2. Example of a trial of the two alternative forced choice training task.....25

Figure 2.3. Example of each condition in the match/mismatch task.....27

Figure 3.1. Model Estimated results for accuracy scores on match/mismatch trials pre- and post- training for implicit and explicit training groups.....33

Figure 3.2. Model-Estimated Values for 300-500ms ERP effects on nouns..... 36

Figure 3.3. Model-Estimated Values for 300-500ms ERP effects on determiners..... 38

Figure 3.4. Model-Estimated Values for 700-900ms ERP effects on nouns 40

Figure 3.5. Model-Estimated Values for 700-900ms ERP effects on determiners.....42

Figure 3.6. Waveform differences across conditions pre- and post-training in response to determiners.....44

Figure 3.7. Waveform differences across conditions pre- and post-training in response to nouns..... 44

Figure 3.8. Overall waveform differences across conditions pre- and post-training..... 45

Figure 3.9. Topographical plots of average ERP scalp distribution at 400 ms for determiners..... 45

Figure 3.10. Topographical plots of average ERP scalp distribution at 400 ms for nouns.....46

Figure 3.11. Topographical plots of average ERP scalp distribution at 800 ms for determiners..... 46

Figure 3.12. Topographical plots of average ERP scalp distribution at 800 ms for nouns..... 47

ABSTRACT

Many adults wish to learn a second language to improve their employability, cognitive function, or social connections. However, learning a second language as an adult can seem very daunting. One particularly difficult feature of some second languages is that of grammatical gender; a noun classification system that separates nouns according to somewhat arbitrary classes which include, in many cases, feminine and masculine. This study's aim was to identify whether explicit and implicit teaching methods differ in their ability to teach grammatical gender in a second language to individuals whose first language does not contain the feature. A group of English native speakers underwent a two-day training paradigm using a two-alternative forced choice task to learn 44 artificial nouns and their genders either explicitly or implicitly. A match/mismatch ERP task was used in combination with a gender association task (Phillips & Boroditsky, 2002) to assess gender and determiner learning. ERP and behavioural results did not show group differences in gender learning, but showed small N400s and increased accuracy ratings indicative of early noun learning. Limitations and future directions are discussed.

LIST OF ABBREVIATIONS USED

L1	First Language
L2	Second Language
AL	Artificial Language
CPH	Critical Period Hypothesis
AOA	Age of Acquisition
ERP	Event Related Potential
EEG	Electroencephalography
GAT	Gender Association Task
GAM	Generalized Additive Modeling
LME	Linear Mixed Effects Modeling
ROI	Region of Interest
ICA	Independent Component Analysis

CHAPTER 1: INTRODUCTION

1.1 Second Language Acquisition

In a world of international development and collaboration, one of the most valuable skills a person can have is multilingualism. To be able to convey and express ideas in more than one language not only increases one's employability, it also has the potential to make one more successful in one's chosen field (Christofides & Swidinsky, 2008). Unfortunately, in Canada the rate of bilingualism and multilingualism amongst adult English native speakers is not high, at about eight percent (Statistics Canada, 2013). For adult learners wishing to broaden their linguistic diversity, second language (L2) learning can seem challenging or even impossible. In fact, many adults find learning a second language to be very challenging, and especially more challenging than learning a second language as a child. Investigating why L2 acquisition seems more difficult for adult learners could help lead to an amelioration of second language instruction and learning for late learners.

The most notable explanation as to why adults have a more difficult time acquiring a second language is the Critical Period Hypothesis (CPH) (Lennenberg, 1967). Proponents of the CPH suggest that, in order to learn a second language completely and with near-native proficiency, one must learn it before a particular (critical) age. This age is often cited as the onset of puberty when the brain begins to prune neurons, and neuroplasticity, the ability for the brain to change in significant ways, is believed to be reduced (e.g. Snow & Hoefnagel-Höhle, 1978). Proponents of CPH believe that language acquisition beginning after the critical period will result in a

lack of mastery of the second language, and, that native-like proficiency will not be achieved.

A number of studies have demonstrated that one aspect of language that might be subject to a critical period is grammar. Johnson and Newport (1989) administered a grammaticality judgment task to immigrant L2 speakers of English who had arrived in the US and achieved proficiency in English at different ages. They found a strong effect of age of immigration on test performance, such that before puberty performance was correlated with age of arrival in the US, with those who had arrived earlier having higher test scores. For the group who had arrived after puberty, test scores were low, and there was no correlation between age of arrival and performance, suggesting evidence for the CPH. Weber-Fox and Neville (1996) investigated grammatical acquisition in a second language by early and late bilinguals. They presented two groups - one who had learned a second language before 11 years, and the other after 11 years - with grammatical violations in the second language. They found, using neurophysiological methods of assessment called event related brain potentials (see below), that early learners were more sensitive to gender violations than late learners, suggesting that grammar is sensitive to a critical period.

The critical period hypothesis is a highly debated theory in the field of language research. While there has been some evidence of age-related decline in ability to acquire a second language, there is very little convincing evidence of a sharp decline following a so-called *critical age.* Birdsong and Molis (2001) conducted a replication of Johnson and Newport's (1989) study. They found that the negative correlation between age of acquisition (AoA) and proficiency continued past the "critical age," and also found

evidence of native-like late learners. They suggest that L2 proficiency may be more reliant on frequency of L2 use, as well as overlap between the L1 and L2, than on AoA.

Chiswick and Miller (2008) gathered census data from immigrants to the USA, and found that although English proficiency did decline as a function of age at migration, there was no sharp decline in proficiency at a specific age, suggesting a lack of critical period. Flege and Yemi-Komshian (1999) presented native korean speakers who were second-language learners of English with a series of English grammaticality judgements. Although they did find that correct grammaticality judgements decreased with the age of acquisition (AOA), they found that the effects of AOA disappeared when controlling for other confounding variables, such as exposure to English, and the amount of education received in the United States. The only aspect that did appear to decline consistently with a later AoA was the participants' ability to shed their foreign accent. Other research has shown similar results (i.e. Steinhauer, 2014; DeCarli et al, 2015; Brice & Brice, 2008), suggesting that while adult learners might be limited in their phonological learning in a new language, given the proper circumstances, they are likely able to learn a second language with native-like proficiency.

Second language acquisition requires learning new phonemes, an entirely new vocabulary and also new syntactic and grammatical rules. One of the most challenging aspects of second language acquisition can, in fact, be the acquisition of novel grammar not present in one's first language (Bobb, Kroll, & Jackson, 2015). Many languages have grammar features that are unique to the specific language, or to a group of languages. These grammar features can be difficult to grasp for learners who are not familiar with them. This study investigated the learnability of grammatical gender by

individuals whose first language does not contain the feature. Specifically, this study investigated the ERP components on a match/mismatch task completed by a group of individuals whose first language does not contain grammatical gender after a two-day language training paradigm.

1.2 Grammatical Gender

This project employed computer-based second language learning games to investigate how people face one of the most challenging steps in learning many new languages: grammatical gender. Many languages — such as French, Spanish, Italian, Dutch, German — categorize nouns according to specific, often arbitrary classes called genders, whereas other languages — such as English, Finnish, Japanese, Malay and Turkish — do not. In addition to gender classes being arbitrary, there is often little consistency in a noun's gender across those languages where grammatical gender classes are present. For example, in Spanish the word for "key" is feminine, whereas in German it is masculine. In German the word for "apple" is masculine, whereas in French it is feminine. Moreover, the gender of certain nouns often do not make sense on a semantic level. For example, in French the word "beard" — a word which one might think of as inherently masculine — has feminine gender, whereas the word for "mascara" — a conventionally feminine item — is masculine. While native speakers of gendered languages tend not to have any trouble identifying the genders of nouns in their native language, these inconsistencies and irregularities can make learning gender in a new language especially difficult for adults (Lemhöfer, Schriefers, & Hamique, 2010).

The gender of a noun influences many syntactic behaviours, such as which

determiner to use as well as other word- formation patterns such as verb conjugation. For example, in French, masculine words are given the determiners *le* (definite) or *un* (indefinite), whereas feminine nouns are given the determiner *la* (definite) or *une* (indefinite). Furthermore, the gender of a noun in a sentence can influence the form of adjectives; the phrase *la petite fille*, meaning "the little girl", differs in both determiner and adjective form from the phrase *le petit garcon,* meaning "the little boy", whereas in English, these phrases have the same determiner, *the*, and the adjective *little* does not differ in conjugation. To conjugate incorrectly according to gender is extremely uncommon among native speakers of gendered languages (Sabourin, 2001). Therefore, learning the grammatical genders of nouns is important for achieving native-like proficiency in a second language that contains grammatical gender features.

There is some indication that grammatical gender serves a cognitive purpose for native and proficient speakers and listeners of gendered languages. When speaking a language with grammatical gender, the determiner of a noun provides important information to the listener about the word that is going to follow. Grosjean, Dommergues, Cornu, Guillelmon, and Besson (1994), conducted a series of two experiments to determine the facilitatory effects of gender on lexical retrieval and word recognition in French native speakers. They found that nouns preceded by a gender article, such as a determiner or an adjective, were identified more quickly than those that were not preceded by gender articles, or by gender neutral or ambiguous articles. Guillelmon and Grosjean (2001) conducted a similar experiment with English-French bilinguals. They found that early learners of French (those who had become bilingual before the age of 13) had near-native like facilitatory effects of gendered articles on

word recognition, but late learners (those who had become bilingual at 24 years old or later) showed no effects of gender marking on word recognition. Bates, Devescovi, Hernandez, and Pizzamiglio (1996) presented participants with determiner-noun pairs that were either congruent or incongruent in terms of gender. Participants were asked to decide if the determiner-noun pairs were grammatical or not via a button press. Bates and colleagues found that participants were faster at identifying congruent determiner-noun pairs than incongruent pairs, suggesting a disruptive effect of the incongruity, and further offering support for the concept of facilitation and inhibition effects of gender on word recognition and language processing. This suggests that adult second language learners might not experience the cognitive benefits of gendered articles that are experienced by native and early learners.

However, it should be noted that learning grammatical gender in a second language is not impossible for speakers of a non-gendered language. Sabourin, Stowe, and de Haan (2006) found that second language speakers of Dutch were able to correctly identify the genders of nouns about eighty percent of the time despite whether their first language contained grammatical gender or not. This shows that, for adult native speakers of non-gendered languages, although the perceptual cognitive benefits of learning grammatical gender in a second language might not be present, it is not impossible for late learners of a gendered language to acquire near native-like proficiency when speaking in a second, gendered language. Similar results have also been obtained in other studies (e.g. Kurinski & Sera, 2011; Kempe, Brooks, Kharkhurin, 2010; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012).

Some research indicates that late learners whose native language contains

grammatical gender might face different challenges than those whose first language does not contain the linguistic feature. Due to the lack of gender congruence across different gendered languages, it has been shown that speakers of one gendered language will often assign the incorrect gender to a word in another gendered language when the genders in the two languages do not coincide. Lemhofer, Schriefer, and Hamique (2010) conducted a training study wherein German speakers were instructed on the genders of words in Dutch, and were asked to name images of pictures in Dutch. Even after intensive training including feedback on their performance, German speakers still misgendered Dutch words that had incongruent genders in German. Sabourin and colleagues (2006) found that when identifying gender agreement in Dutch, native speakers of German (which has a high rate of overlapping gender with Dutch) performed better than speakers of Romance languages (which have a lower rate of gender overlap), suggesting the transfer of grammatical knowledge from one's first language to the new language . Interestingly, Sabourin and colleagues also found that English L1 learners of Dutch performed significantly more poorly than German or Romance language speakers on a Dutch gender agreement task, suggesting that while gender incongruity might pose an obstacle for learners of gendered languages, simply having learned an L1 containing grammatical gender might give them an advantage over speakers whose language does not contain grammatical gender.

Native speakers of non-gendered languages do not have any framework for gender, and therefore may have more difficulty with grasping the concept of grammatical gender when learning a second language. Hawkins and Chan's (1997) "failed functional features hypothesis" suggests that grammatical features not present in

one's first language are especially difficult to acquire in a second language due to this lack of grammatical framework. Without the awareness of grammatical gender and how it is applied, native speakers of non-gendered languages do not have the foundational knowledge on which to apply genders acquired in a new language. Whereas native speakers of gendered languages have preexisting knowledge of gender systems, and the ability to understand how genders might be applied in a new language.

1.3 Implicit vs. Explicit Grammar Instruction

Given these differences, it is possible that the method of language instruction might influence learners' ability not only to acquire grammatical gender, but to acquire it correctly and efficiently. There has been some debate over whether second language instruction should employ explicit or implicit training of grammatical features. Explicit training involves the outright instruction of grammatical features, for example conjugation drills, while implicit training uses a more immersive technique, with the assumption that second language learners will learn the language in a similar way to how babies learn their native languages; i.e., without explicit instruction about grammar rules. Immersion has long been heralded as the most organic and successful way to learn a second language because it employs native-like learning processes, and allegedly leads to more native-like proficiency in a second language.

However, it must be noted that second language acquisition is different from first language acquisition in many ways, not excluding the possibility of the rules from one's first language interfering with rules of a second language. It is for this reason that metalinguistic awareness might be necessary for second language learners when learning novel language rules in a second language. Metalinguistic awareness has, in fact, been

shown to be beneficial in the learning of a second language. Tipitura and Jean (2014) found that second-grade French immersion students performed better at gender assignment tasks when they received explicit instruction about gender and how to use phonemic cues to identify the genders of nouns compared to when they did not receive this explicit instruction. Andringa and Curcic (2015) investigated how linguistic structures that do not exist in one's first language are best learned in a second language. They taught Dutch participants a series of new words with a linguistic feature called differential object marking (DOM). DOM is a form of noun classification, used to differentiate inclusivity, which is somewhat similar to grammatical gender but does not exist in Dutch. Half of the participants were given explicit instruction about DOM , while the other half did not receive any instruction. Results showed that explicit instruction resulted in better learning of both DOM and the new words, relative to implicit instruction. This suggests that for individuals learning novel grammatical features in a second language, explicit instruction to promote metalinguistic awareness of the feature might be helpful in learning.

Presson, MacWhinney and Tokowicz (2011) trained adult native English speakers on grammatical gender in French across three conditions: explicit instruction and immediate feedback, immediate feedback without explicit instruction, and no instruction (implicit category). Results showed that participants who had received feedback with explicit instruction performed significantly better on a gender assignment task than participants in the other groups, suggesting that for English native speakers (and possibly speakers of other non-gendered languages) explicit instruction is beneficial when learning grammatical gender classes in a second language.

The current research in explicit grammatical gender instruction is lacking in solid, consistent findings related to adult second language learners. Moreover, the research on explicit grammar instruction for grammatical gender has focused only on learners whose first language does not contain grammatical gender. Thus, there has not been any conclusive research to determine the best way to teach adults from different linguistic backgrounds grammatical gender in a second language, nor has there been any direct comparison of explicit gender learning across groups of individuals from languages with and without grammatical gender.

1.4 Behavioural Methods of Assessing Gender Learning

There have been a number of studies exploring the cognitive effects of grammatical gender on language comprehension both in first and second languages. Konishi (1993) had monolingual German and Spanish participants describe a series of objects. Half of the objects were classified as feminine in German but masculine in Spanish, and the other half were classified masculine in German, but feminine in Spanish. German and Spanish participants had a tendency to describe the objects that were masculine in their own language as being more *potent* than those words that were feminine in their own language. Potency was measured as how "good" participants considered the concept to be. The authors suggested that this data is representative of the influence of gendered classes on the way that speakers of certain language process language and perceive the world. It also demonstrates the pervasiveness of gender classes in the processing of linguistic material for L1 speakers of gendered languages.

Segel & Boroditsky (2011) examined the work of approximately one million sculptors and artists who have created pieces to represent abstract concepts such as *love*

and *sin* and *justice*. The authors found that artists whose first languages are languages with grammatical gender represented these abstract terms as people of genders that correspond to the gender class of the term 78% of the time. For example, in German the word for *sin* is feminine, German artists typically represent sin as a woman. However, in Russian, sin is a masculine term, and is typically represented by artists as a man. Segel and Boroditsky controlled for changes in grammatical gender by assessing the language associated with each piece as it was at the time the piece was created. This article further demonstrates the way that nuances in language, such as grammatical gender classes influence the way that people perceive and understand the world and the objects around them. Given these effects of gender on cognition and the associations that are made based on gender categorization, one could posit that measuring such associations could function as an implicit method for measuring gender learning.

Phillips and Boroditsky (2003) assessed the way gender learning affected the categorization of nouns. They taught participants a series of feminine and masculine gendered nouns in an artificial language, and asked them to rate how alike these individual nouns were to female and male humans pre- and post- training. It was found that after learning a series of gendered words in an artificial language, English-speaking participants rated feminine nouns in this artificial language as being more similar to female humans, and masculine nouns as being more similar to male humans. Moreover, in a similar experiment, Phillips and Boroditsky found that nouns within the same gender categories were grouped together when participants were asked to randomly assign nouns to groups. A similar study by Beller (2015) asked participants who were speakers of gendered languages to assign female or male voices to inanimate objects. It

was found that participants were significantly more likely to assign female voices to feminine nouns and male voices to masculine nouns.

This body of research pertaining to the categorization of nouns based on grammatical gender suggests that speakers of gendered languages do tend to categorize nouns in their lexicon according to gender. Thus, a gender association task similar to those of Phillips and Boroditsky, or Beller, could be used as a valid and reliable way to assess gender learning in a second language. If participants are accurately learning gender in a second language, then their likeness ratings of nouns with the same gender in that language should increase post-training. However, at the early stages of training it is likely that this increase in likeness rating will be useful only for those whose first language does not contain grammatical gender. Individuals whose first language contains grammatical gender likely already have strong connections between nouns of the same gender in their first language, and an increase in likeness of nouns after a brief training paradigm is unlikely.

1.5 Physiological Methods for Assessing Gender Learning

A notably useful neurophysiological tool used for language research is electroencephalography (EEG) and event related brain potentials (ERP). EEG records electrical activity from the brain, non-invasively, using electrodes placed on the scalp. ERPs refer to the measure of electricity measured at a specific point in time, or in response to a specific event. This technique has exquisite temporal resolution and can help characterize the nature and timing of different operations in language processing. It is well established that for native speakers, when there are certain incongruencies in linguistic stimuli, such as incorrect grammatical gender, there will be an enhanced

positive-going waveform approximately 600 ms after the stimulus is presented; this is known as the P600. The P600 was first demonstrated by Osterhout and Holcomb (1992), and was elicited in native English speakers in response to syntactic anomalies, but not for syntactically normal sentences. The P600 has since been shown to be elicited for native speakers by both visual and auditory stimuli, in response to syntactic and grammatical anomalies including tense, gender, and cases (Gouvea, Phillips, Kazanina & Poeppel, 2010). The P600 is believed to be reflective of processing that involves revising and repairing semantic errors in language. Kaan and Swaab (2003) performed a study wherein they provided English native-speakers with ambiguous sentences requiring revision (in which the sentence is unusual but plausible) and sentences requiring repair (in which the sentence is unusual and non-plausible). They found P600s were elicited in response to these sentences in different scalp locations. Sentences requiring ambiguity resolution elicited frontal P600s, while those requiring repair elicited posterior P600s.

When measuring performance in non-native speakers, a P600 comparable to that obtained by a group of native speakers implies more native-like processing, indicating that the non-native speaker has learned to a degree that is native-like. Loerts, Stowe, and Schmid (2013) completed an ERP study to investigate the P600 as it pertains to gender violations. They presented Dutch native speakers with a series of sentences that contained gender violations. They found consistent P600 effects for gender violations. This indicates that a P600 effect in response to gender violations in a second language would imply native-like proficiency.

Dowens, Guo, Guo, Barber, and Carreiras (2011) presented Mandarin L1s, who

were proficient late learners of Spanish with sentences containing both number and gender violations. Consistent P600 effects were found for both gender and number violations. Because Mandarin does not contain either number or gender markers, this indicates that P600 effects can be present for late second language learners whose first language does not contain the grammatical feature violated in the second language.

Other studies have confirmed a P600 to grammatical violations in L2 learners, which mirror native speakers' cortical response, although this effect may be weaker in lower-proficiency learners. For example, McLaughlin and colleagues (2010) demonstrated that after one year of second language exposure, L2 learners had near-native like P600s for syntactic violations. These ERP effects were even more pronounced, and nearer to native-like, after three years of language training.

A study by Foucart and Frenck-Mestre (2011) used EEG to test gender violations in speakers' native and second languages. They presented German L1 speakers who were late but proficient learners of French, and native French speakers, with a series of determiner-noun pairs containing gender violations. Results showed P600s for violations that were consistent between native and second languages, but not for violations that were inconsistent across the two languages. This suggests that for second language learners learning a language with features present in their first language, features that are inconsistent across the two languages will be more difficult to learn in the L2.

These studies show the possibility of late L2 learners acquiring native-like brain activity in response to specific features in second languages, and demonstrate the utility of EEG and ERPs in determining the degree of learning. However, what they show is native-like processing after an extended amount of exposure and training in the second

language. The results for short term training are less widespread, and it is not well-known whether these native-like results will be seen after only a short period of instruction. Morgan-Short, Sanz, Steinhauer, and Ullman (2010), conducted a study to evaluate learners' after a shorter training paradigm. Participants were instructed to speak and understand an artificial language over the course of three training sessions. Participants were divided into explicit (classroom-like) and implicit (immersion-like) training, and were assessed after the first and third days of training. Results showed weak, but present ERP effects, including P600s and N400s (another language-specific ERP effect; see below) for gender violations in both groups, which varied as a function of proficiency and training group, such that at low levels of proficiency N400s were only elicited in response to gender agreement violations for the implicit group, and for gender agreements in the explicit group. At higher proficiency, however, N400s and P600s were elicited in response to agreement violations in both the implicit and explicit training groups.

Another ERP effect that is used as an indicator of early language learning is the N400 (Kutas & Hillyard, 1980). Like the P600, the N400 is an ERP effect that occurs in response to linguistic stimuli. The N400 is a negative-going brain response occurring approximately 400 milliseconds after presentation, in response to semantically incongruent stimuli. If an N400 is found post-training in a new language in response to semantically incorrect stimuli, then it can be presumed that some learning did occur in the new language. For example, Pu, Holcomb, and Midgley (2016) demonstrated that after only four hours of instruction in Spanish, adult English L1 learners demonstrated strong N400 effects in response to semantically incongruent sentences in spanish. The

N400 has also been demonstrated in response to mismatched gender-noun pairs. For example, Morgan-Short et al. (2010) presented L2 learners of an artificial language with adjective-noun pairs which were either semantically congruent (i.e. the adjective was reflective of the gender of the noun), or that were semantic violations (i.e. the adjective did not indicate the correct gender of the noun). They found that the semantic violations led to N400 effects. Foucart and Frenck-Mestre (2012) showed an N400 effect in response to violations after a grammar learning task in an L2 to be indicative of early learning, whereas a P600 effect in response to the same stimuli were shown to be indicative of proficient L2 learners, or native-speakers. Therefore, the N400, like the P600, can be used to assess the early stages of grammar learning in a second language.

Steinhauer and Dury (2009) describe the way in which ERP responses evolve throughout the second language learning process. Past literature by CPH proponents has suggested that ERPs changed as a function of age of acquisition, where P600s were elicited in response to ambiguous stimuli by individuals who had acquired a language at a younger age, and N400s by those who were late learners of a language. However, Steinhauer and Dury, through the analysis of many ERP studies investigating L2 acquisition, have demonstrated that differences in ERP effects are not related to age of acquisition, but to proficiency, and that ERP effects change over time as L2 learners become more proficient in the second language. The N400 is described as an early indicator of L2 learning, representing a reflection of rote memorization of grammatical information captured by syntactic violations. As a person becomes more proficient in a second language, the ERPs shift from N400s to P600s, which have been shown to be indicative of greater proficiency, as they reflect deeper processing such as revision and

repair. A shift from an N400 to a P600 in response to the same violation might indicate a shift from rote memorization to more automatic, rule-based processing, implying more proficient understanding.

1.6 The current study

The current literature surrounding grammatical gender learning and ERPs in English native-speakers is sparse and inconclusive. The aim of this study was to address the question of whether grammatical gender in a second language could be taught to individuals whose first language does not contain the feature, and what ERP effects would be seen in this early learning stage. A third research question for this study was whether the method of training (implicit vs. explicit) would make a difference in English native speakers' ability to learn grammatical gender.

Finally, this study represents the first stage in a larger project aimed at comparing second language learners with different linguistic backgrounds to determine if having grammatical gender in one's first language would yield different results from not speaking a gendered first language. It is the goal that the findings of this study can be combined with future research using francophone participants in order to make conclusions about these possible differences.

To achieve the goals of this study, I conducted a training study to teach participants a new gender system in a miniature artificial language. I used a between-groups design, training native English speakers on an artificial language containing 44 novel vocabulary words, half of which had a feminine gender and the determiner *das*, and the other half of which had a masculine gender and the determiner *dos.* Using a two-alternative forced choice task, one group of the participants were taught the words

explicitly (with gender rule explanations) and a second group learned implicitly (the existence of the gender system, and gender of each word, were left to be inferred from examples). To assess sensitivity to grammatical gender, and to test the degree to which participants could recognize violations in the new language, participants completed a match-mismatch task both before and after two days of training. On each trial, a line drawing of one of the training items was shown followed by a spoken word in the artificial language. The word-image pairs formed four conditions: (1) match (correct word, correct gender); (2) gender violation (correct word, incorrect gender); (3)semantic violation (incorrect word, correct gender); (4) double violation (incorrect word, incorrect gender). After each spoken word, participants were asked to indicate with a button press whether the word matched the picture in both meaning and gender. Patterns of brain activity were compared pre- and post-training, using non- invasive electrical recordings (electroencephalography, or EEG). Furthermore, participants completed a gender association task (GAT; adapted from Phillips and Boroditsky, 2002) prior to and after completing the two days of training. The purpose of the GAT was to assess the cognitive organization of gendered nouns prior to and after training. It was expected that after training in the gendered language a shift in noun organization would be seen, such that objects with feminine and masculine genders would be grouped with humans of the same genders.

1.7 Hypotheses

First, participants in both categories were expected to have higher accuracy ratings post-training on the match/mismatch task than pre-training for match, semantic mismatch, and double violation conditions, indicating that the nouns in the artificial

language had been learned. The participants in the explicit category were expected to have higher accuracy improvement rates post-training for gender mismatches and double mismatches than the participants in the implicit category if, as predicted, explicit instruction led to better learning of syntactic gender.

Based on past research regarding ERPs and language learning, I hypothesized that semantic and double violations would produce N400s post-training for all conditions, implying early learning of nouns by participants in both training categories. I hypothesized that gender violations would produce either an N400 or a P600 effect. The presence of an N400 in response to gender violations would imply a reliance on rote memorization of determiner-nouns pairs. A P600 is indicative of revision and repair that is seen in more proficient learners. Therefore, a P600 in response to gender violations was considered a possibility for the participants explicit training condition, although the brief training period may not have been sufficient to result in participants' treating the gender violations as "syntactic" as opposed to violations of rote-memorized word-pair associations; only an N400 was expected for the implicit training condition.

Participants in the explicit training category were hypothesized to have significantly higher GAT scores for object-human pairs of the same gender post training compared to before training. Such an increase in likeness ratings would indicate a shift in cognitive organization of the items within the same gender category in the artificial language, and would imply that the genders had been learned, at least subconsciously. Congruently gendered human-object pairs are hypothesized to have a greater increase in GAT likeness ratings compared to incongruent human-object pairs. No differences are expected in other pairs, which were included for balancing purposes only. Finally a

difference in likeness ratings was hypothesized between the explicit and implicit groups, with the explicit group having significantly higher likeness ratings on congruently gendered object-human pairs compared to the implicit condition because the implicit group, having no basis for gender, was not expected to learn gender, whereas the explicit training group was given the framework for gender learner.

CHAPTER 2: METHODS

2.1 Participants

Twenty-three native English speakers participants were recruited through posters and online advertisements. Two participants did not complete the training and are not included in any analyses. The first five participants were treated as pilot data and were entirely excluded from analysis due to a combination of being incompletely saved by the computer program running the experiment, or because they did not contain the proper trigger code data needed to sync stimuli with ERP data. Data from two additional participants were removed for the same reasons. Data from one participants were excluded from GAT analysis due to an apparent response bias (the participant chose the same answer for every trial). Sixteen participants' data were included in accuracy ratings. Of the 16 people whose data was used in analysis, eight were female. The mean age of participants was 29.9 years, (*sd* =12.94 years). Four participants reported limited knowledge (very low proficiency) of other gendered languages. On average participants had completed 18 years of formal education, including kindergarten or primary (*sd* = 2.8 years). Ten participants reported having achieved a bachelor's or an advanced degree. All participants were right-handed according to their responses to the Edinburgh Handedness Inventory (Oldfield, 1971). Participants were systematically assigned to implicit and explicit categories by predetermined participant numbers that corresponded to these categories. The explicit category (n = 7) included 4 female participants, and had a mean age of 27.42 years. The implicit category had a mean age of 32 years. T-tests show no significant differences in age between the two groups. Criteria for participating in this research included right handedness, being an adult native English speaker without fluency in any gendered language, not having any neurocognitive deficits or conditions

that affect attention, and having a hairstyle that allowed for an electroencephalography cap to be applied (i.e. no tightly braided or bound hair.). All participants provided informed consent according to the Declaration of Helsinki; all study procedures were approved by the Dalhousie University Research Ethics Board.

2.2 Materials

2.2.1 Language and Education Survey

A survey was administered to gather information on individual differences including age, sex, handedness, linguistic background, and education. This survey included a modified version of the Edinburgh Handedness Inventory (Oldfield, 1971).

2.2.2 Artificial Language Vocabulary

Forty-six novel words were created for the purpose of this study using an online word generator (wordgenerator.net). All words were designed to contain only phonemes and phoneme combinations present in English, and ranged from one to three syllables. In an initial validation study, a group of English and French native speaker volunteers (n =19) completed a word association task with the artificial words to indicate what they thought of when they read the artificial words, to ensure that none of the words had strong associations with other real words in English or French. All words were then vetted using an online urban dictionary (urbandictionary.com) to check that they did not have colloquial meaning. These words were then assigned meanings to four pictures of human females and four pictures of human males, 12 animals, 24 objects, and two determiners (one feminine, and one masculine). The images of human females and males depicted people in stereotypically female or male roles or occupations (e.g., female humans included a girl, a nun, a teacher, and a witch, and male humans included

a priest, a fireman, a king, and a wizard). To ensure suitability to planned future research with Francophone participants, half of the non-human words were assigned to be the same gender as in French, and the other half were designed to be the opposite gender in French. All human words were designed to have same gender as they do in French, thus all human females were assigned the feminine gender, and all human males were assigned the male gender. This was done for ease of teaching in the implicit group, as well as for testing in the gender association task. A list of the words used in this study can be found in Appendix A.

2.2.2 Gender Association Task

The gender association task (GAT) was adapted from Phillips and Boroditsky (2003). The GAT in the present study asked participants to rate 162 pictures of noun pairs on how similar they seem before and after word-learning using a seven-point likert scale. All of the images were of nouns taught in the artificial language, and the images used in the GAT were the same ones used in the learning paradigm, but differed from those used in the match/mismatch task. Noun pairs were balanced to include an equal number object-object and object-human pairs for both male and female humans. The inclusion of object-object pairs was done to deter participants from guessing the nature of the task in order to avoid participant bias. All pairs were also counterbalanced for gender in the artificial language as well as in French, such that there were an equal number of pairs that were congruently and incongruently gendered in the artificial language as well as in French. Half of the pairs overlapped in gender for French and the artificial language, and half differed between French and the artificial language. An example of the GAT can be seen in Figure 2.1. All images for the learning task were

taken from a database of cartoon images created by Copernicus Studios Inc. (Halifax, NS) for use in research in our lab, or from open source image databases using a Google image search of the web.

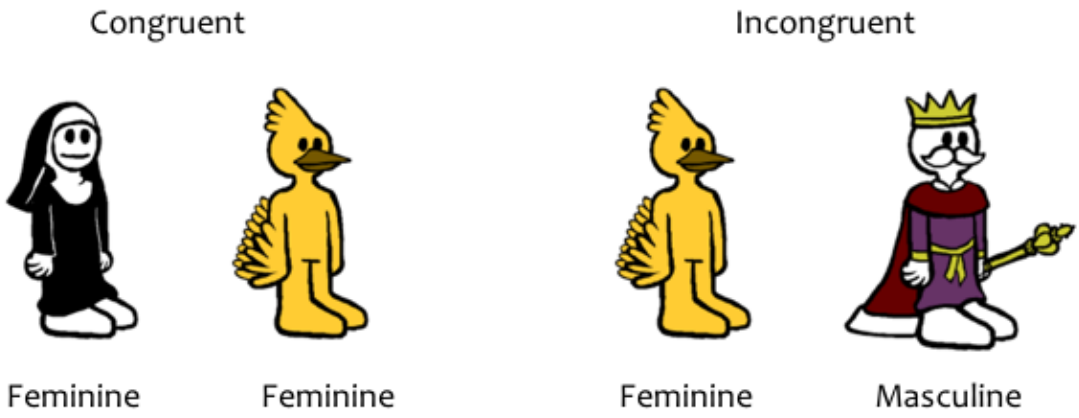


Figure 2.1. Example of congruent and incongruent GAT pairs

2.2.3 Two Alternative Forced Choice Learning Task

Two, two-alternative forced choice tasks were created using PsychoPy (Pierce, 2009) for the purpose of this study; one for explicit and one for implicit instruction of gender and vocabulary in the artificial language. Participants were either prompted to read about gender assignment rules in a new language which corresponded either to an explicit condition or an implicit condition. The instructions for the explicit condition read “Now you will learn some words in a language you have never heard before. Words in this language have "Grammatical Gender." This means that some words are categorized as feminine, and others are categorized as masculine. Feminine words get the determiner "das." Masculine words get the determiner "dos." All female humans have the determiner "das," and all male humans have the determiner "dos."” Implicit instructions did not include the text explaining gender rules. In each trial, two pictures

appeared on the screen, and a word played over the speaker. Participants indicated, using specified keys on a keyboard, whether they believed they heard the word matching the image on the left of the screen, or the word matching the image on the right of the screen. Participants were given feedback, indicating if their choice is correct, and the correct answer remains on the screen. Each of the 44 words in the previously described artificial language (36 objects/animals, and eight humans) were shown three times in random sequence, for a total of 132 trials. An example of a trial is shown in figure 2.2. All cartoon images used can be found in Appendix B.

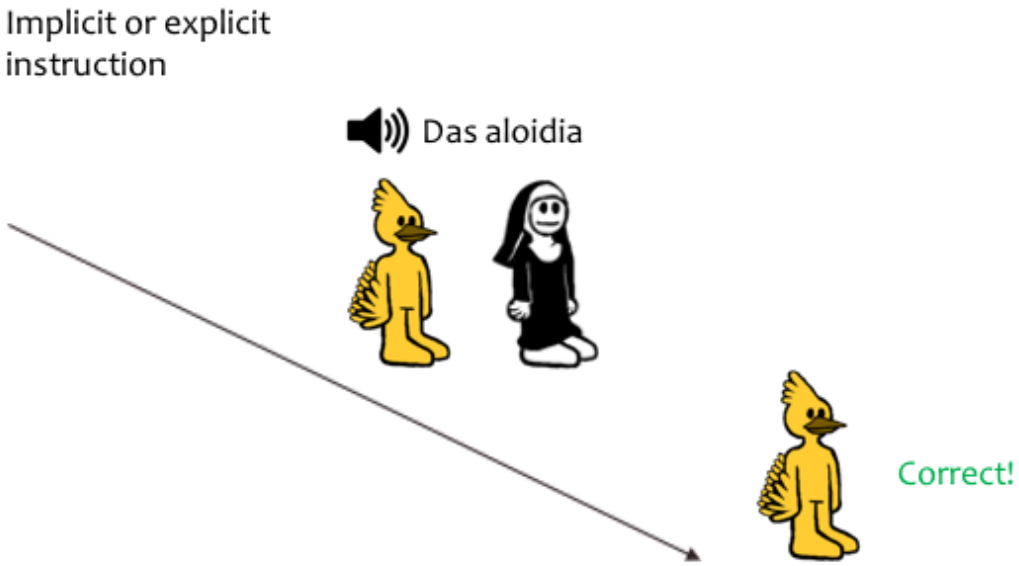


Figure 2.2. Example of a trial of the two alternative forced choice training task

2.2.4 Match/ Mismatch ERP task

A match/mismatch task was created using PsychoPy for use in the EEG portion of this study. An image, different in style from the ones used in the learning task, appeared on the screen and a word was played over the speaker in the artificial language.

All words included a determiner, followed by a noun 600 ms after determiner presentation. Determiners and words were recorded and played separately for ease of ERP analysis; this allowed for time locking to both determiners and nouns in all conditions. After stimulus presentation, participants were told to indicate, using specific keys on a keyboard whether or not they thought the word that they heard over the speaker matched the image they saw on the screen. Explicit instructions about what kind of "match" (i.e., semantic, gender, or both) to look for were not given.

Trials were of four types: (1) match (correct word, correct gender); (2) gender violation (correct word, incorrect gender); (3)semantic violation (incorrect word, correct gender); (4) double violation (incorrect word, incorrect gender). Eaxmples of each match category can be found in Figure 2.3. Each category had 40 trials for a total of 160 trials (although some unintentional variance may have occurred). Line drawings found through Google image search were used for the match mismatch task to ensure that participants had in fact learned the words for the items in the images and not relied on recognition of specific images. All line drawings used in this task can be found in Appendix C.

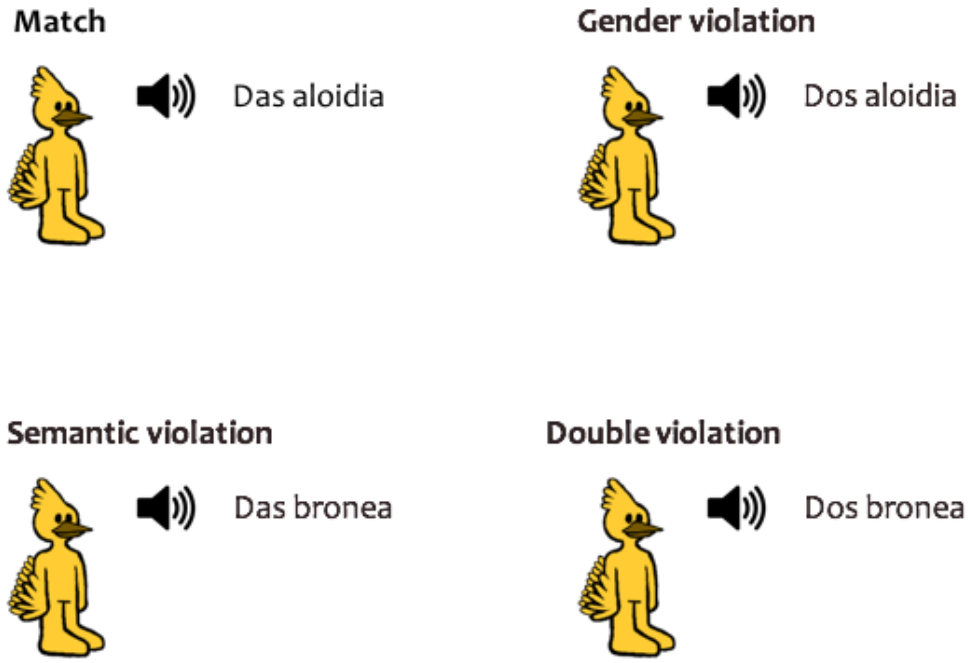


Figure 2.3. Example of each condition in the match/mismatch task. Note that das represents the feminine determiner, and dos the masculine determiner. In this example, the word for bird is feminine.

2.2.5 Naming Task

A naming task was created using PsychoPy for use in the learning assessment portion of this study. The purpose of the learning task was two-fold. First, it acted as a measure of participants' recall of the words taught in the learning task. Second, it acted as a prompt to encourage participants to think about the new words in the artificial language prior to completing the GAT in order to activate knowledge about the gender system in the artificial language. The same 44 images that were used in the learning task appeared in random sequence on the screen for five seconds each. Participants were instructed to say aloud, in the artificial language, the name of each of the images at their presentation. Participants were instructed to say "pass" if they did not recall the name of the image. Because the participants could not have had any knowledge of the artificial

language prior to the learning paradigm, this task was only performed at the end of the experiment, and not as a pre-test.

2.3 Procedure

Participation in the present study required three visits to the lab on three consecutive days. Participants began the first session by reading and signing a consent form, and then completing a computer-based survey to gather information on their handedness, and language and education backgrounds. Next, they completed the GAT, as described above. Then, the EEG cap was applied (see next section), and participants completed the match/mismatch ERP task. Finally, participants completed the first of two alternative forced choice learning task. On the second visit participants completed the second two alternative forced choice learning task. On the third day participants completed the naming task, followed by the GAT, and finally finishing with the match/mismatch ERP task.

2.4 EEG Acquisition and Data Analysis

EEG data were collected using 64 silver-silver chloride active, preamplified electrodes (Acticap; BrainVision, Morrisville, NC), which filled with electrolyte gel (SuperVisc; BrainVision) and which were attached to an elastic cap and connected to a 64-channel amplifier (QuickAmp; Advanced NeuroTechnology, Enschede, Netherlands). The impedance of each electrode was lowered below a threshold of 30 kΩ. Adhesive electrodes were applied above and below the right eye, and on the outside of the right and left eyes to track bipolar eye-movements. Data were digitized at a sampling rate of 512 Hz, on-line low-pass filtered at 138 Hz, and average-referenced via ASALAB software (Advanced Neuro Technology, Enschede, Netherlands).

Individual participants' EEG data were converted from ANT to EEGlab format using EEGlab (Delorme & Makeig, 2004) software in Matlab (Mathworks, Nattick, PA) before being processed for individual component analysis using MNE (Gramfort et al., 2013). Bandpass filters were set at 0.1 to 40 Hz for the EEG data. Epoch windows corresponding to time-locked events were set from 200 ms pre-stimulus onset to 1000 ms after stimulus onset. The time-locked events of interest were the determiners and nouns. Excessively noisy channels and individual trials were removed based on visual inspection generated using MNE. Independent components analysis (ICA) using the FastICA package (Hyvärinen, 1999) was performed to find and correct ocular artifacts, and artifacts were rejected using visual inspection of topographical scalp maps of all of the averaged individual components. Individual components were removed based on topographical distribution and spectral frequencies that indicated eye movement, single noisy channels, and effects that appear to be heavily weighted in only a few trials, and with large variance between those and trials. After ICA, data at the scalp electrodes were re-referenced to an average of the mastoid electrodes. Then, the electrodes that were previously removed were interpolated using data from the surrounding channels.

Mean amplitudes in each time window were calculated for each participant at each electrode for each trial. Analysis focused on the 300 - 500 ms and 700 -900 ms time windows, which were selected based on prior literature concerning the predicted ERP effects (N400 and P600, respectively). The electrodes were divided into nine regions of interest according to the following: left posterior (electrodes labelled according to the International 10-10 System as P7, P5, P3, PO7), central posterior (P1, Pz, P2, PO3, POz, PO4, O1, Oz, O2), right posterior (P8, P6, P4, PO8), left midline (T7,

FC5, FC3, T7, C5, C3, TP7, CP5, CP3), midline central (FC1, C1, CP1, Cz, CPz, FC2, C2, CP2), and right midline (FC4, FC6, FT8, C4, C6, T8, CP4, CP6, TP8), left anterior (AF7, F7, F5, F3), middle anterior (Fp1, AF3, F1, Fz, Fp2, AF4, F2), and right anterior (AF8, F4, F6, F8).

Statistical analysis of EEG data was performed in the R software, version 3.2.2 (R Core Team, 2017). Linear mixed effects (LME) was used to analyze the ERP data. LME is an extension of the general linear model which controls for both fixed and random effects. The dependent variable used in this analysis was mean amplitude of the waveform in the window of interest. The fixed effects were time (pre- and post-training), and match condition (match, semantic mismatch, gender mismatch and double mismatch). The random effects were subject and trial.

The *bam()* function from package *mcgv* (Wood, 2011) was run in R to determine the best fit for the data for each of the 300 - 500 ms and 700 - 900 ms timeframes. Amplitudes were the dependent The fixed effects for this LME model were: time (pre and post training), violation conditions (determiner violation, semantic violation, and double violation), and regions of interest (ROI).

The analysis procedure was as follows. First, the initial model was run, and data points whose residuals were outside 2.5 standard deviations of the mean were identified and removed as outliers. Approximately 1.9% of the data were removed at this step. The normality of the data following outlier removal was verified using q-q plots, histograms, and residual plots. Following outlier removal, the full model with all fixed and random effects described above was fitted to the data. Then, progressively simpler models were fitted by systematically removing fixed or random effects terms singly; for fixed effects

this started by comparing the full 3-way interaction model, Time × Condition × ROI, with a model including only all 2-way interactions). Models were compared using Akaike information criterion (AIC) weights, which estimates the fit of each model compared to other models. The optimal model was determined as the model with the lowest AIC value, representing the most likely model, penalized by the number of parameters to discourage overfitting.

CHAPTER 3: RESULTS

3.1 Learning Assessment

3.1.1 Naming Task

Naming task data were scored manually. Participants were scored on both gender and noun recall. On average, participants recalled 1.4 nouns correctly regardless of gender, 1 gender noun pair correctly, and 1 gender correctly regardless of the nouns.

3.1.2 Match-Mismatch Task: Accuracy Ratings

Accuracy data was assessed using a logistic regression run in R using generalized linear mixed effects modeling (LME) with a binomial distribution to assess whether accuracy increased post-training, and if any such changes were different across match conditions or between explicit and implicit training groups. Fixed effects input into the model included session (pre- or post-training), violation condition, and training condition, and random effects included individual participants, images and sounds. The results of the model indicated a significant three-way interaction between session, violation, and training condition, $\chi^2 = 12.49$, $p = .0059$. Post-hoc tests were computed to identify the cells for which accuracy ratings were above chance (50% for a yes/no task); all p-values reported have been Bonferroni-corrected for the total number of cells in the design (16). Results showed that accuracy was not significantly different from chance in any pre-training conditions. Post-training accuracy was shown to be above chance in the implicit training group for the double violation, $z = 3.75$, $p = .0058$, and semantic violation, $z = 4.21$, $p = .0006$, conditions, and in the explicit training group for double violations, $z = 3.39$, $p = .0214$. Accuracy was shown to be significantly below chance for gender violations post-training in both the implicit ($z = -4.69$, $p < .0001$), and explicit ($z = 4.42$, $p = .0003$) training groups. The model-estimated accuracy rates for each cell are

shown in Figure 3.1.

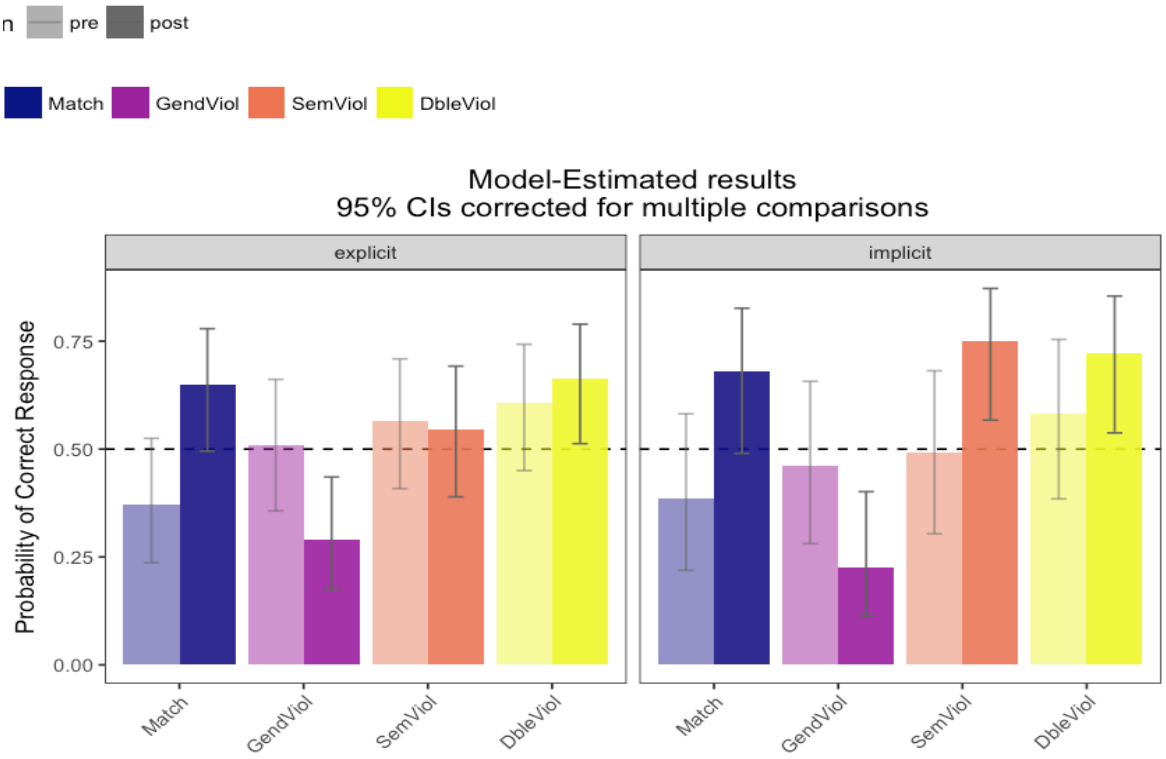


Figure 3.1. Model Estimated results for accuracy scores on match/mismatch trials pre- and post- training for implicit and explicit training groups.
*=p<.05, **= P<.001

Additional post-hoc tests were computed to assess whether accuracy scores improved from pre- to post-training. Results showed a statistically significant increase in accuracy post-training in both groups for match trials, explicit $z = 6.501$, $p < .0001$, implicit, $z = 5.39$, $p <= .0001$, and as well in the implicit group only for semantic violations, $z = 4.84$, $p < .0001$, and double violations, $z = .00559$, $p <= .0001$. Significant decreases in accuracy were seen for for the gender violation condition in both the implicit, $z = -1.47$, $p <= .0001$ and explicit, $z = -1.78$, $p <= .0001$ conditions.

A *d'* analysis was conducted to evaluate whether or not participants had a response bias when completing the match/mismatch task both pre- and post- training. D'

shown in Figure 3.1.

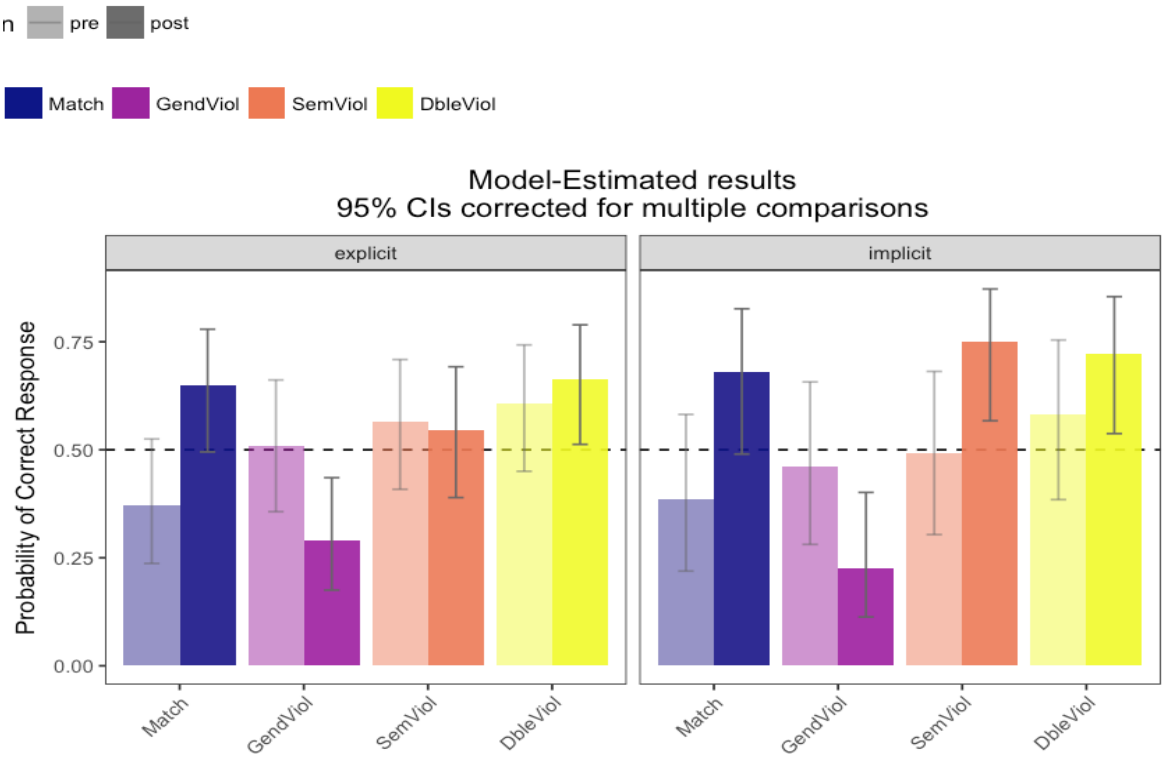


Figure 3.1. Model Estimated results for accuracy scores on match/mismatch trials pre- and post- training for implicit and explicit training groups.
*=p<.05, **= P<.001

Additional post-hoc tests were computed to assess whether accuracy scores improved from pre- to post-training. Results showed a statistically significant increase in accuracy post-training in both groups for match trials, explicit $z = 6.501$, $p < .0001$, implicit, $z = 5.39$, $p <= .0001$, and as well in the implicit group only for semantic violations, $z = 4.84$, $p < .0001$, and double violations, $z = .00559$, $p <= .0001$. Significant decreases in accuracy were seen for for the gender violation condition in both the implicit, $z = -1.47$, $p <= .0001$ and explicit, $z = -1.78$, $p <= .0001$ conditions.

A *d'* analysis was conducted to evaluate whether or not participants had a response bias when completing the match/mismatch task both pre- and post- training. D'

analysis pre-training this showed a bias toward choosing mismatch, $d' = -.318$, $C = .448$, while post- training there was a slight bias toward choosing match, $d' = .383$, $C = -.262$.

3.2 EEG Analysis

Linear Mixed Effects Modelling was conducted for ERP data analysis as described in the previous chapter, in the 300 - 500 ms and 700 - 900 ms timeframes to test for N400 and P600 effects, respectively. The data from two additional participants were removed due to being incomplete because of a technical error (trigger codes marking the onset of critical words were not recorded properly). The resulting groups were unbalanced and small; thus ERP data was analyzed only across collapsed conditions, and not between groups.

3.2.1 300 - 500 ms

3.2.1.1 Nouns

For the 300 - 500 ms time window, the nouns and determiners were assessed individually from the onset of each sound file. The optimal model for explaining the nouns was found to include a 3-way interaction between violation type, session (pre- or post-training), and ROI (the location of the electrodes on the scalp, as described in the previous chapter). A summary of the main effects and interactions of this model can be found in Table 3.1. Planned comparisons between match and mismatch amplitudes at each ROI and for each group describe these interactions; here and in all subsequent ERP analyses, these comparisons were corrected for 54 multiple comparisons using Holm's method (3 mismatch types × 2 sessions (pre/post) × 9 ROIs. N400s were elicited upon presentation of nouns in the double violation and semantic violation conditions in both

the mid-central and mid-posterior ROIs. The model-estimated effect sizes (in μV) of the match - mismatch difference are shown in Figure 3.2. As can be seen, statistically significant differences in waveforms were seen. Significant negativity was elicited post-training in response to nouns in the gender violation categories in the mid-central (t = -3.55, p = .039), left - central (t = -3.85, p = .0354), left-posterior (t = -3.85, p = .0354) and mid-posterior (t = -5.27, p <= .001) ROIs. Significant negativity was also seen in the double violation condition post- training mid-centrally (t = -7.05, p <= .001) and mid-posteriorly (t = -7.16, p <= .001). Finally, negativity was seen in response to nouns in the semantic-violation category mid- centrally (t = -5.82, p <= .001) and mid-posteriorly (t = -6.86, p <= .001).

Table 3.1. Summary of linear mixed effects model for nouns at 300-500ms.

	df	F	p-value
Sentence Type	3	14.80	<.001
Session	1	2.24	.007
ROI	8	0.78	.625
Sentence Type x Session	3	27.74	<.001
Sentence Type x ROI	24	2.04	.002
Session x ROI	8	1.11	.352
Sentence Type x Session x ROI	24	2.37	<.001

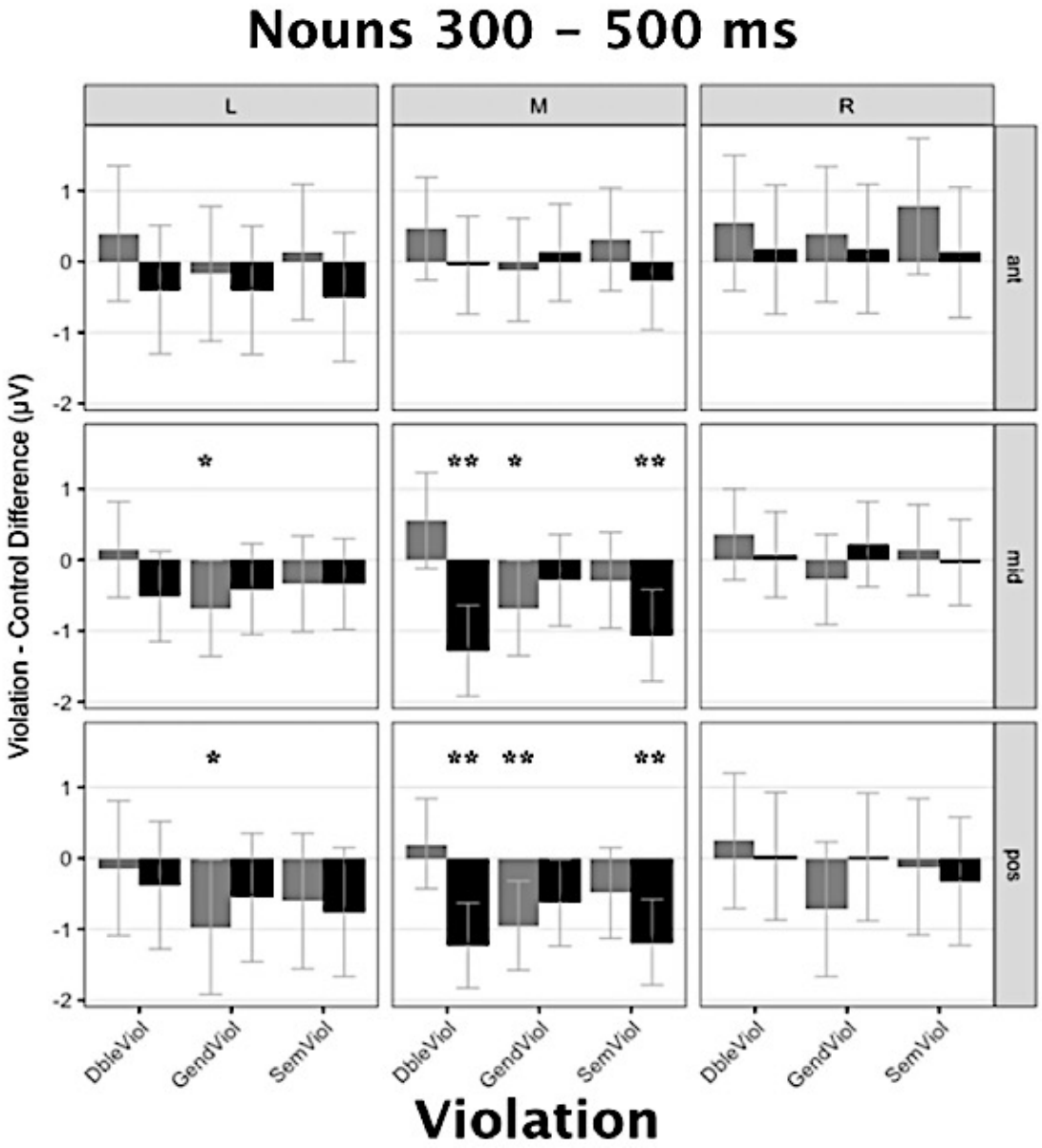


Figure 3.2. *Model-Estimated Values for 300-500ms ERP effects on nouns.*
*p<.05; ** p<.001

3.2.1.2 Determiners

The best model for describing the ERP data from determiners in the 300 - 500 ms time window was also found to be a three-way interaction between violation type, session and ROI. A summary of the main effects and interactions of this model can be found in Table 3.2. Planned comparisons between violation and control amplitudes at

each ROI and for each group describe these interactions. The model-estimated effect sizes of the match - mismatch difference are shown in Figure 3.3. Semantic violations elicited an increase in negativity in response to determiners in the left anterior ROI (t = -4.16, p = .0032), and double violations at the determiner level elicited increased negativity post-training in the left anterior (t = -4.86, p <= .001), left middle (t = -4.71 , p <= .001), and mid anterior ROIs (t = -4.48, p = .0011).

Table 3.2. Summary of linear mixed effects model for determiners at 300-500ms.

	df	F	p-value
Sentence Type	3	3.34	.018
Session	1	0.45	.501
ROI	8	0.66	.730
Sentence Type x Session	3	0.48	.696
Sentence Type x ROI	24	0.67	.883
Session x ROI	8	2.60	.008
Sentence Type x Session x ROI	24	1.23	.201

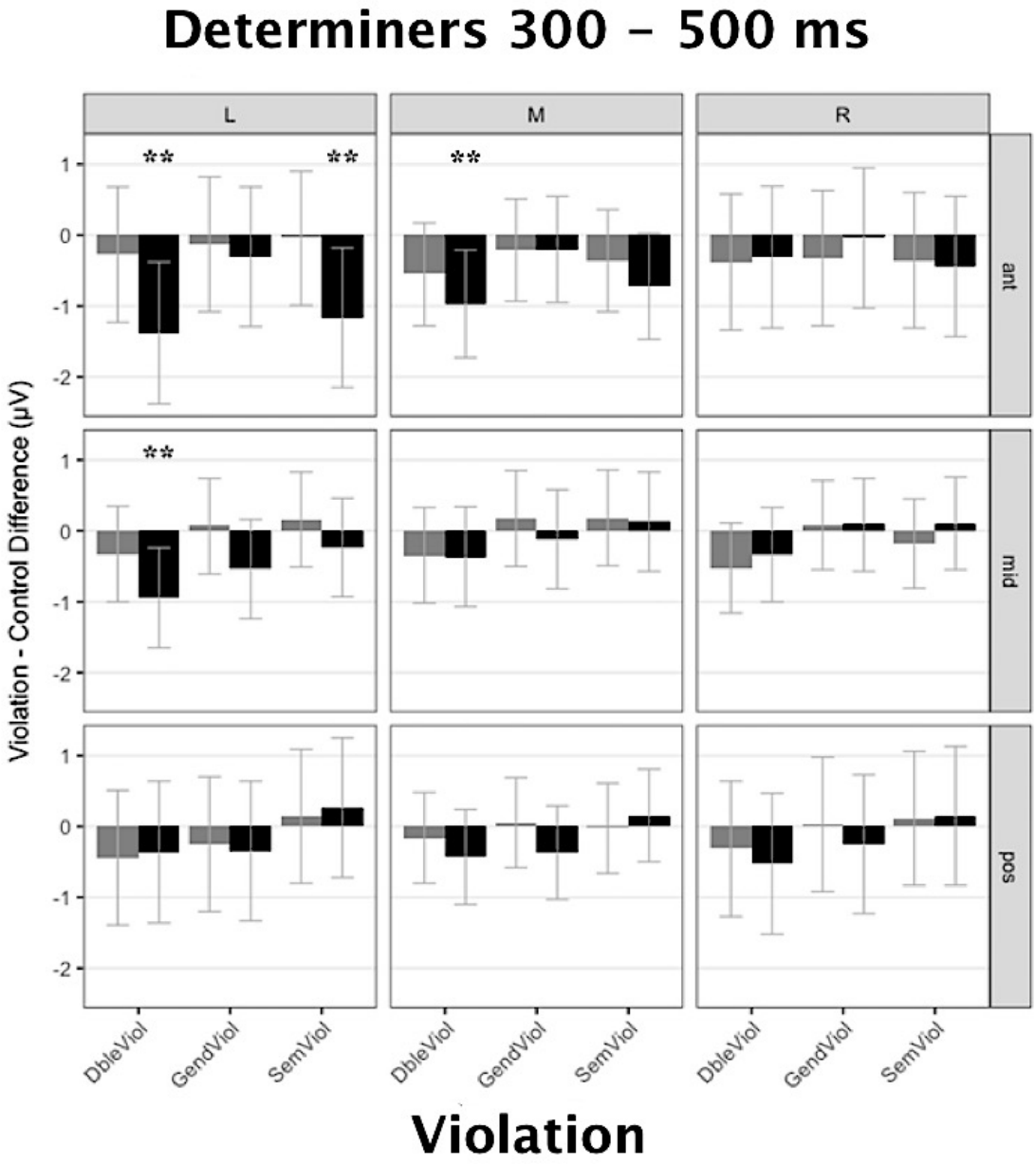


Figure 3.3. *Model-Estimated Values for 300-500ms ERP effects on determiners.*
*p<.05; ** p<.001

3.2.2: 700 - 900 ms

3.2.2.1 Nouns

The best model for describing the nouns was found to be a three-way interaction between sentence-type, session, and ROI. A summary of the main effects and interactions of this model can be found in Table 3.3. Planned comparisons between

match and mismatch amplitudes at each ROI and for each group describe these interactions. The model-estimated effect sizes of the match - mismatch difference are shown in Figure 3.4. As can be seen, a significant difference in negativity was seen pre- and post training in the double violation condition for nouns in the mid-central (t = -3.71, p = .022), and mid-posterior (t = -3.97 , p = .0159) ROIs, and in the semantic violation condition in the mid-posterior ROI (t = -3.54, p = .027), with post-training waveforms showing increased negativity in both cases. Statistically significant differences in waveforms were also seen pre- and post- training in the double violation condition in the mid central (t = -5.61, p <= .001) and mid-posterior (t = -5.69 , p <= .001), with increased negativity seen in the pre-training condition in these cases.

Table 3.3. Summary of linear mixed effects model for nouns at 700-900ms.

	df	F	p-value
Sentence Type	3	12.13	<.001
Session	1	2.47	.116
ROI	8	3.39	<.001
Sentence Type x Session	3	22.99	<.001
Sentence Type x ROI	24	1.44	.076
Session x ROI	8	0.53	.835
Sentence Type x Session x ROI	24	1.99	.003

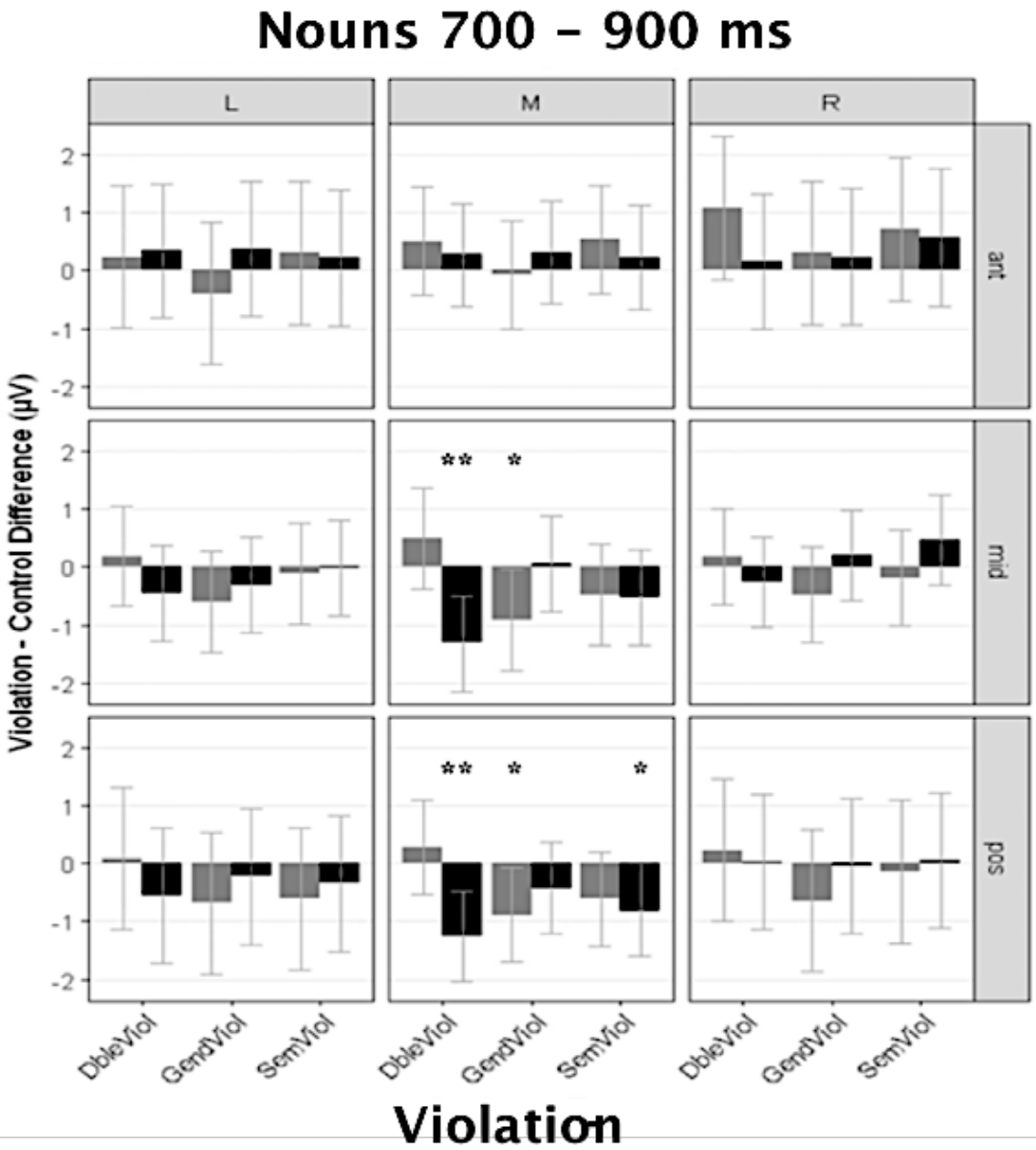


Figure 3.4. *Model-Estimated Values for 700-900ms ERP effects on nouns.*
*p<.05; ** p<.001

3.2.2.2 Determiners

Two models met the criterion for "best model" (lowest AIC value) for the data from determiners in the 700-900 ms time window: the full 3-way interaction model, and the model containing only all 2-way interactions. Because we defined the optimal model as the one having the lowest AIC value and the least number of terms, we took the

model without the 3-way interaction as the best one for this data set. A summary of the main effects and interactions of this model can be found in Table 3.5. Planned comparisons between violation and control amplitudes at each ROI and for each group describe these interactions. No significant effect were found. The model-estimated effect sizes of the match - mismatch difference are shown in Figure 3.4.

Table 3.4. Summary of linear mixed effects model for nouns at 700 - 900 ms.

	df	F	p-value
Sentence Type	3	1.69	.17
Session	1	2.46	.12
ROI	8	1.62	.11
Sentence Type x Session	3	9.57	<.001
Sentence Type x ROI	24	1.56	.04
Session x ROI	8	1.41	.19

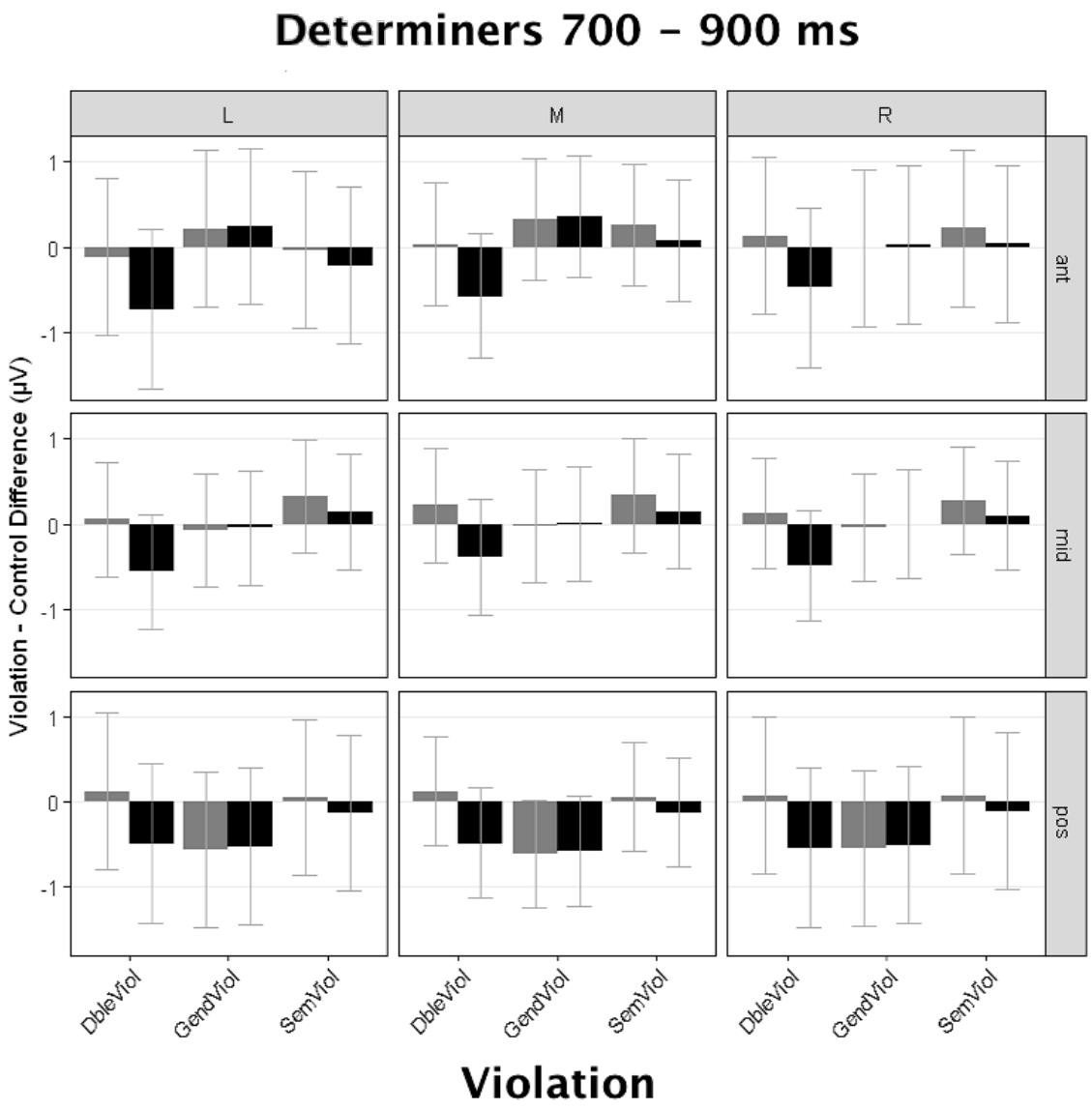


Figure 3.5. *Model-Estimated Values for 700-900ms ERP effects on determiners.* *p<.05; ** p<.0013

Visual inspection of the waveforms pre- and post-training in response to determiners show small but apparent differences for semantic violations, with negativity being slightly higher pre-training. Even smaller differences in waveforms are seen for the gender violation and double violation conditions, with negativity being slightly increased post-training (Figure 3.6). Visual inspection of the waveforms pre- and post-

training in response to nouns shows small differences. Greater negativity is seen pre-training for the gender and semantic violations, while negativity seems to be increased post-training for the double-violation condition (Figure 3.7). The inspection of overall difference in waveforms pre- and post-training shows that there is very little in the way of compelling ERP effects (Figure 3.8).

Visual inspection of the topographical plots for determiners in the 300 - 500 ms time window shows small but evident changes in activity pre- and post-training in the double violation condition, but not for other conditions (Figure 3.9). No marked differences can be seen on topo plots pre- and post- training for nouns in the 300 - 500 ms time window (Figure 3.10). Visual inspection of topographical plots of determiners at the 700 - 900 ms time window show evident differences pre- and post-training in the gender and double violation conditions, and very small differences in the semantic condition (Figure 3.11). Visual inspection of the nouns in the 700 - 900 ms time window shows little evidence of changes pre- and post- training, with the exception of some evident change in the semantic violation condition (Figure 3.12).

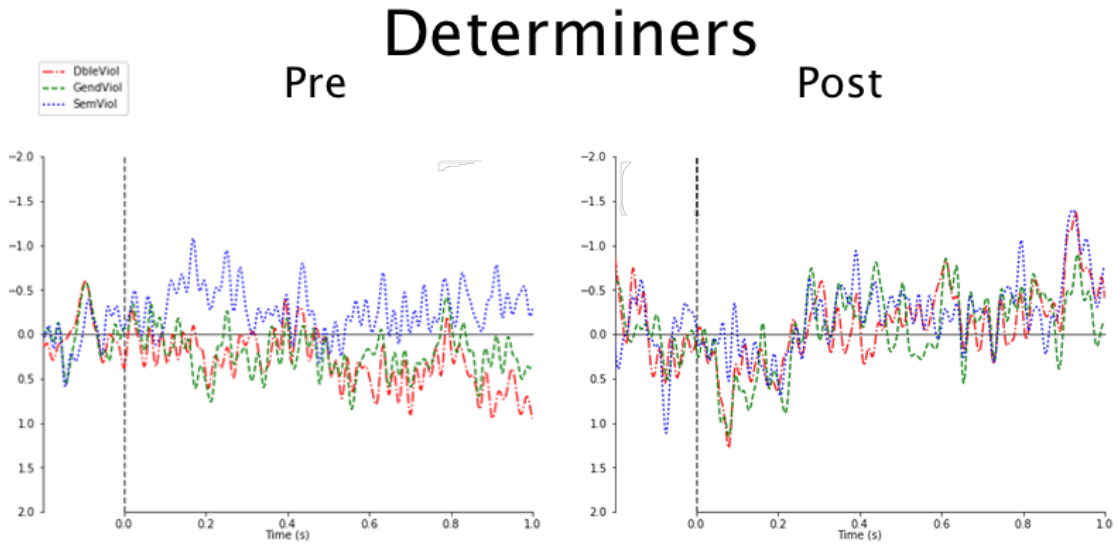


Figure 3.6. *Waveform differences across conditions pre- and post-training in response to determiners.*

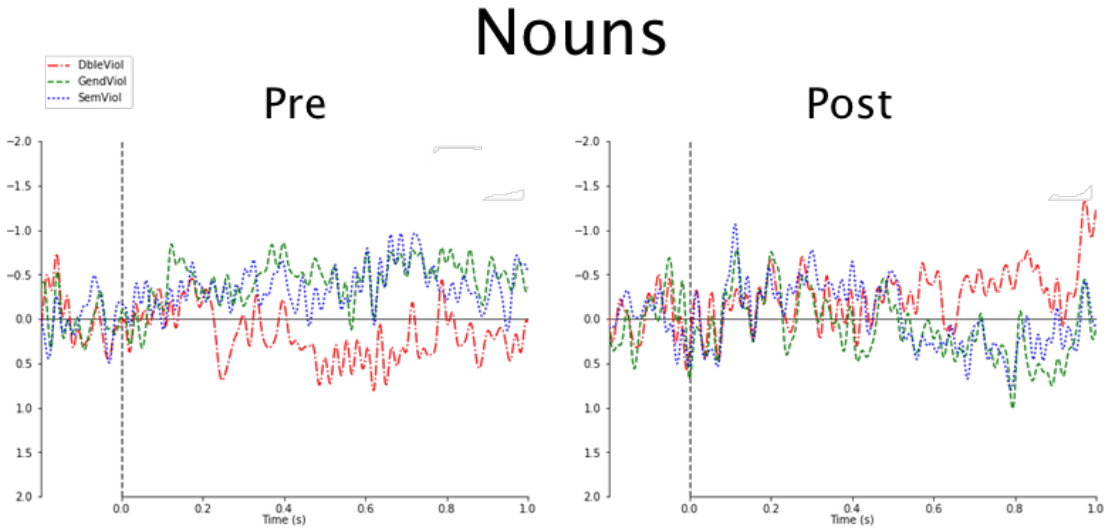


Figure 3.7. Waveform differences across conditions pre- and post-training in response to nouns.

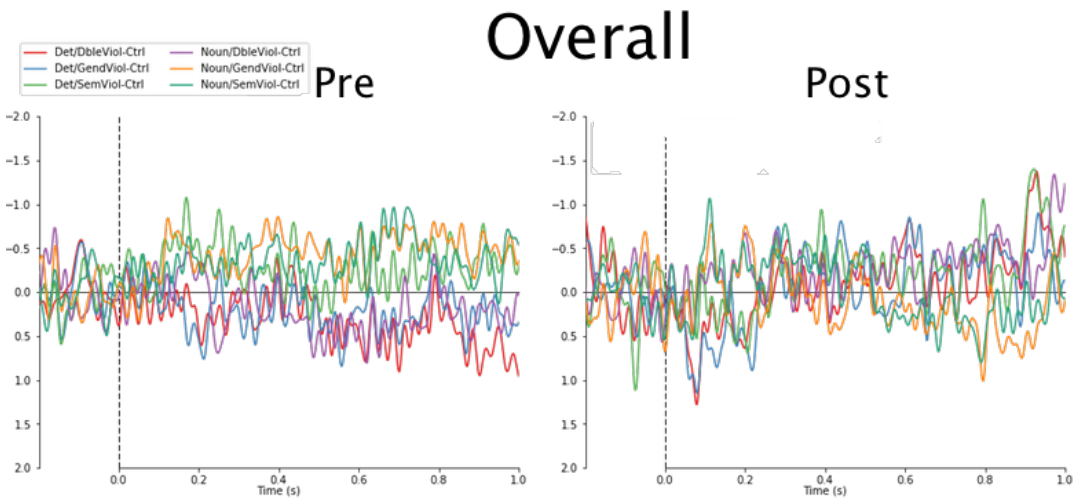


Figure 3.8. *Overall waveform differences across conditions pre- and post-training.*

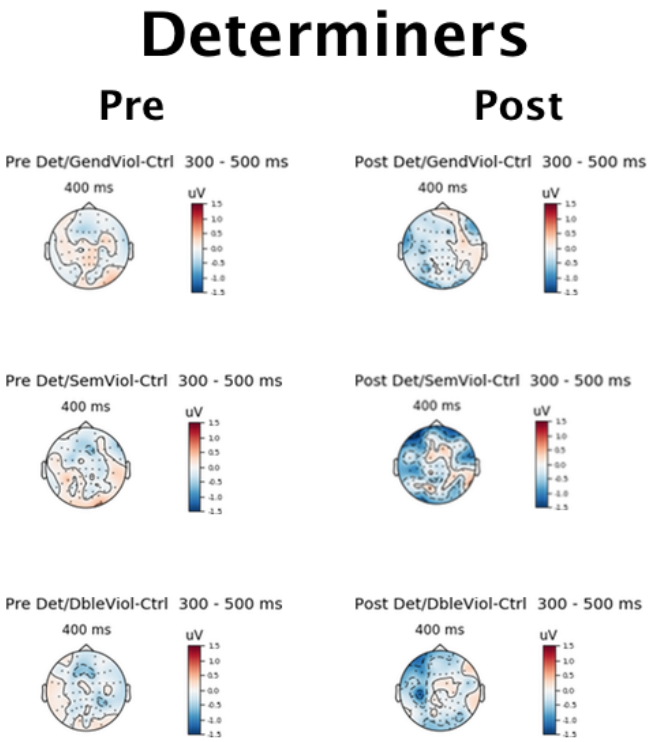


Figure 3.9.
Topographical plots of average ERP scalp distribution at 400 ms for determiners.

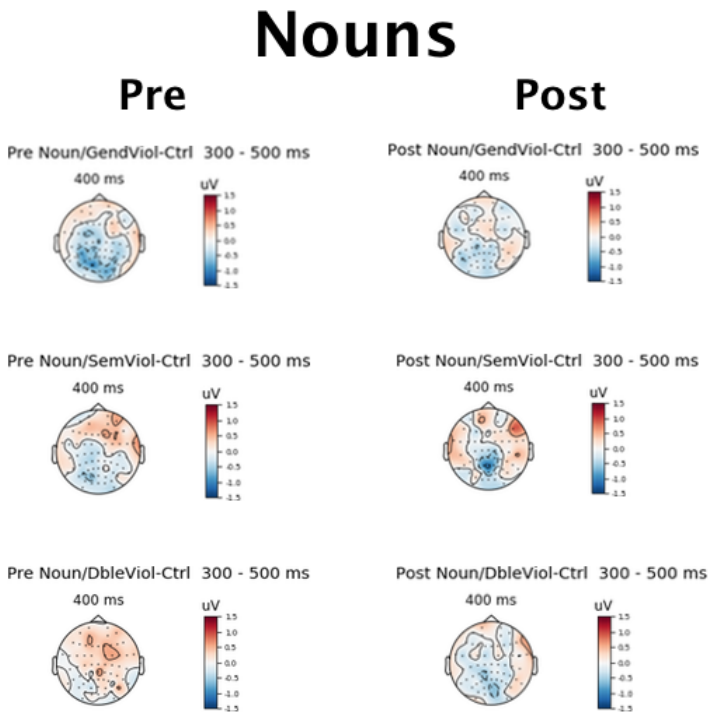


Figure 3.10. *Topographical plots of average ERP scalp distribution at 400 ms for nouns.*

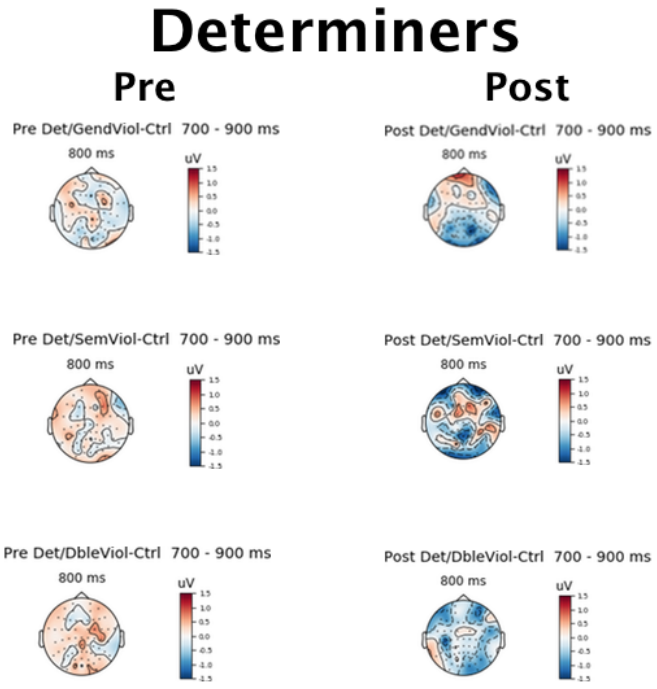


Figure 3.11. *Topographical plots of average ERP scalp distribution at 800 ms for determiners.*

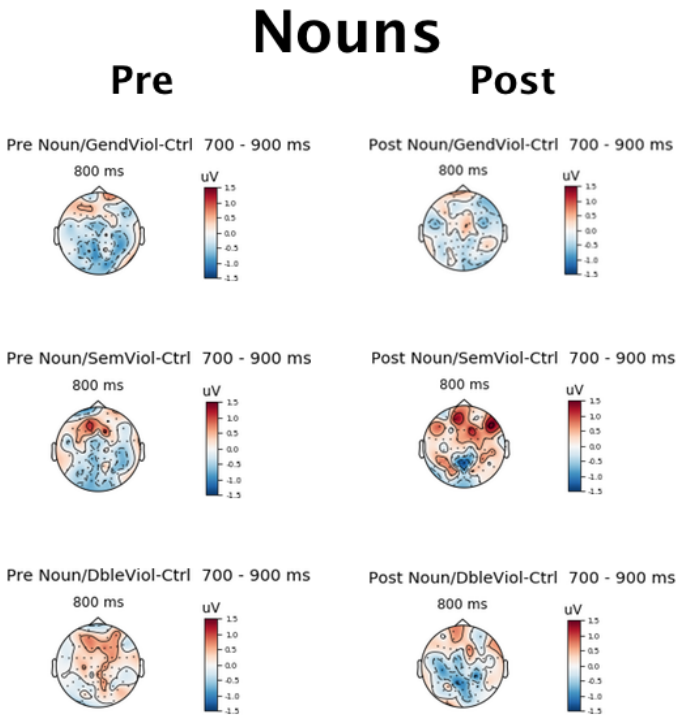


Figure 3.12. *Topographical plots of average ERP scalp distribution at 800 ms for nouns.*

3.3 GAT Analyses

A two-way repeated measures ANOVA was performed to assess the differences between pre-and post training GAT ratings on congruent and incongruent object-human pairs between the explicit and implicit training groups. The three-way interaction model was not statistically significant, indicating no significant interaction between time, training group, and congruency. However, a significant main effect of time was found, $F(1) = 10.14$, $p = .007$, indicating a statistically significant difference between all GAT scores post-training (M = 2.55) and pre-training (M = 2.12). A summary of the main effects and interactions of this model can be found in Table 3.5.

Table 3.5. Summary of ANOVA model for GAT pre- and post- training.

	df	F	p-value
Congruency	1	3.002	.109
Time	1	10.139	.008
Congruency x Training Group	1	.079	.784
Time x Training Group	1	.383	.548
Congruency x Time	1	1.404	.259
Congruency x Time x Training Group	1	.339	.571

CHAPTER 4: DISCUSSION

4.1 Research Questions

Grammatical gender is a challenging feature to learn in a second language, especially if it is not present in one's first language. Some research has indicated that learning language features not present in one's own language requires a level of explicit instruction, while other research suggests that implicit instruction is best for this type of instruction. Moreover, past research investigating the acquisition of grammatical gender in adults L2 learners is incomplete, and does not include a substantial amount of data concerning ERPs, which can be used as a good indicator of how aspects of language are learned. The aim of this study was to address three research questions:

1. Can grammatical gender in a second language be taught to individuals whose first language does not contain the feature?
2. What ERP effects would be seen in this early learning stage?
3. Would the method of training (implicit vs. explicit) make a difference in English native speakers' ability to learn grammatical gender.

It was expected that, after a two-day training paradigm including a two-alternative forced choice learning task divided between implicit and explicit training groups, participants would learn nouns and their genders in such a way that produces ERP effects in response to specific mismatch conditions, and that these responses would differ between training groups. Specifically, it was hypothesised that P600s would be elicited in response to nouns in the semantic violation and double violation conditions post-training, indicating noun learning, and that P600 effects would also be elicited for participants in the explicit training category in response to determiners, but not for implicit learners, indicating that the explicit training was more beneficial in teaching

grammatical gender than the implicit training. N400 effects were expected to be elicited in response to all violations post-training for the explicit group, and for semantic and double violations in the implicit category, indicating early learning of gender and nouns in the explicit category, and early learning of nouns in the implicit category. It was further hypothesized that likeness ratings on the gender association task would increase only for the explicit category post-training, and only for congruently gendered object-human pairs, and not for object-human pairs that were incongruently gendered.

4.1.1 Question 1

Scores on the naming task were very low (~2.5% correct on average). These low scores are consistent with prior research. Past research in the field of memory and learning indicate that a task that relies on recall, such as a naming test, is a much more difficult one than one that relies on recognition, such as the match/mismatch test, or even the learning paradigm that was used in this study (Loftus, 1971). Moreover, achievement on recognition tasks tend to be indicative of early learning (Schmidtke, 2014), whereas achievement on recall tasks tend to be indicative of more advanced learning (Peters, Hulstijn, Sercu, & Lutjeharms, 2009). Given the brevity of the training paradigm, a low recall score is not unexpected and is indicative of low proficiency, but cannot be taken as a complete lack of learning.

Accuracy ratings in the match/mismatch task—a measure of recognition—did show evidence of learning: post-training accuracy was significantly better than chance for both semantic and double violations in the implicit group, and double violations in the explicit group Moreover, accuracy improved significantly from pre- to post-training on match trials in both groups, and additionally for semantic violations in the implicit

group. The significant improvements for match trials must be tempered by the fact that post-training the accuracy scores were not significantly better than chance. However, accuracy was numerically well above chance for both groups (65% and 68% for explicit and implicit groups, respectively), and the *p*-values approached significance (.066 and .086 for explicit and implicit groups, respectively), consistent with some degree of learning of novel word meanings. It is also notable that accuracy was generally higher post-training in the implicit than in the explicit group, and significantly above chance for both semantic and double violations, whereas explicitly-trained subjects showed better-than-chance performance only for double violations. This suggests that the implicit training may have been more successful at early noun training than the explicit group.

In the gender mismatch condition, where the determiner was incorrect but the noun was correct, there was a statistically significant *decrease* in accuracy across both training groups, with performance going from chance levels to below-chance levels. In other words, post-training, participants were significantly more likely than chance to categorize gender mismatches as "matches." At first glance this may seem to contrast what might be expected after such a learning task. However, this pattern of results suggests that in this task participants were focusing on the relationship between the picture and the meaning of the noun, rather than on the gender marked by the determiner. This might imply two things. First, it is possible that the genders of the nouns were not learned, a possibility that seems to be supported by the results of the gender association task. Moreover, given that the training task only used correct gender-word pairings and the feedback that participants received was only with respect to

picture-word associations, it is not unreasonable that participants would focus only on nouns in the match/mismatch task.

4.1.2 Question 2

To answer the second question, three hypotheses were made.

First, it was hypothesized that all semantic violations would elicit N400s in all conditions post- training. Semantic violations elicited negativity in the 300 - 500 ms time window in the mid-central and mid-posterior ROIs. Double violations also elicited negativity post-training in response to nouns in the mid-central, and mid-posterior regions of interest, which is characteristic of an N400, and suggests early learning patterns, in this case, of the nouns in the artificial language.

Second, it was hypothesized that gender violations would produce either an N400 or a P600 effect, and that the elicited effect would differ between training groups; the explicit group was expected to show P600s, and the implicit group to show N400s, indicating that the implicit group was relying on rote memorization of gender-noun pairs whereas the explicit group was using more advanced processing, implying deeper learning. This hypothesis was not tested due to low power, and unbalanced groups. However, results showed an increase in negativity in response to determiners in the left-anterior ROI, and double violations at the determiner level elicited increased negativity post-training in the left-anterior, left-central, and mid-anterior ROIs. However, these effects are not characteristic of an N400. The N400 has a central-parietal scalp distribution, and the distribution elicited here is in the left-anterior region. This is consistent with another component identified as *left-anterior negativity* (LAN). The LAN is an effect elicited between 150 ms and 500 ms, and is usually associated with

morphosyntactic processing (Steinhauer & Dury, 2012). However, prior research has also demonstrated that the LAN can be an indicator of grammatical processing, and is often elicited, followed by a P600, in response to grammatical violations (Steinhauer & Connolly, 2008). The LAN is usually seen in late L2 learners, and can be representative of advanced, native-like proficiency in an L2 (Weber-Fox & Neville, 1996; Steinhauer & Connolly, 2008). Given that the LAN is often associated with late learners, its elicitation to determiners in the gender and double-violations in this study, which had only a two-day training paradigm, is peculiar. In a study investigating ERP effects of grammatical violations between early bilinguals, who had learned an L2 before the age of 11 years, and later learners, Weber-Fox and Neville found LAN effects present only for violations in early learners. In the present study, participants were all older than 18 years, and by no means were they bilingual. Moreover, the behavioural data regarding gender violations is not consistent with deep learning, and suggests rather that no gender learning occurred. Accuracy ratings for gender and double violations did not increase significantly post-training, and the GAT scores post-training did not differ between congruently gendered and incongruently gendered pairs, suggesting that any increase in ratings were not due to gender learning. Therefore, these apparent LAN effects are likely not, in fact, representative of the neurological underpinnings consistent with other LAN studies, and these effects are likely not due to gender learning.

Semantic violations also elicited LAN-like effects for determiners in the left anterior region of interest. Because the determiners in this condition were congruent with the gender of the image on the screen, this effect was unanticipated. Moreover, a similarly counterintuitive result was seen pre-training. For nouns, gender violations pre-

training elicited LAN in a number of scalp locations. This effect was not hypothesized or anticipated, and is somewhat anti-intuitive, given that at this time-point participants had not had any instruction in the artificial language, and would have no way of knowing that these violations existed. These two effects were not supported by behavioural data, and are especially puzzling. These might add evidence to suggest that the LAN-like effects are not actually LANs, and are not representative of any kind of learning.

Overall, P600 effects were not seen, and ERP effects were not as widespread in the 700 - 900 ms time window as in the 300-500 ms, where some increased negativity, including possible N400s and LAN were seen. This is consistent with past research indicating that N400s are more indicative of early learning than P600s. Given that the training paradigm in this study was less than one hour in total, it is quite unlikely that participants had gained a strong grasp on the artificial language at the time of testing. However, results do indicate that although genders were likely not learned, the nouns in the artificial language were learned to a degree that is consistent with early learning. These ERP effects are consistent with the behavioural results of this study. Accuracy data show post-training effects for both double violations and, in the implicit category, semantic violations, indicating noun learning. Moreover, a lack of ERP effects for gender violations is supported by accuracy data that do not any post-training effects for the gender mismatch condition.

4.1.3 Question 3

To answer the third question, it was hypothesized that GAT likeness ratings of congruently gendered object-human pairs would increase more for the explicit condition

compared to the implicit condition, and that the increase seen for the congruent pairs would not be seen for the incongruent pairs. These hypotheses were not supported by the results of this study, as both groups showed a significant increase in GAT likeness ratings post-training compared to pre-training for all object-human pairs. Moreover, the only significant predictor of GAT increase was time, indicating that there was not a significant difference in GAT increase between congruent and incongruent pairs, and that they both had a significant increase post-training compared to pre-training. It must be noted that although significant, the increase pre- and post-training was less than one half point (.48) on a 7-point likert scale, and even post-training the likeness rating was "not alike".

The significant increase in post-training for the incongruent pairs could be due to a familiarity effect. The participants had already been exposed to the same pairs in the pre-test only two days prior to completing the post-test. It is possible that any increase, including the increase for the congruent pairs was due to having already seen the same objects together in the pre-test, and not due to the training paradigm.

4.2 Limitations and Future Directions

At the outset of this study the goal was to assess the difference between implicit and explicit training on grammatical gender learning between francophone and anglophone adult learners of an artificial language. Past research (e.g. Sabourin et al., 2016) suggests that those with grammatical gender in their first language might require different instruction from those without it when learning a new language containing grammatical gender. However, due to a lack of francophones in a predominantly anglophone region willing to participate, it was not possible to recruit enough volunteers

to achieve that goal. In the future, collaboration with researchers in communities where French is a common language would be helpful in recruiting volunteers to address this question.

Another issue with the execution of this study was with the recruitment of participants. The study required three visits to the lab for up to two hours each, on three consecutive days. Many potential volunteers were unable to participate due to timing issues, and several dropped out after the first or second session because they did not feel they had time to come back for the other session(s). Given this difficulty with recruitment, the number of participants used in this study is low (n=16), and did not allow for between groups comparisons of training groups. Moreover, the ERP effects that were seen in this experiment must be interpreted within the context of the low N, and with the understanding that they are subject to low power, and should thus be considered pilot data. The data may be reflective of trends, but cannot be taken as a solid confirmation of any underlying effects to gender processing in adult L2.

Having the training for the study take place online, rather than in the lab, might encourage more people to complete the study, as they would not be required to travel to the lab for at least one of the sessions. Moreover, an online version of the training paradigm might allow for more training days to be included, possibly resulting in better learning of both the nouns and the genders.

Additionally, this experiment might be ameliorated if the design was changed from a between-groups to a within-groups design. As such, participants would learn the first half of the nouns and genders implicitly, and then learn the second half of the nouns explicitly. This would allow for a smaller sample size, as higher power requires fewer

participants in a within-subjects design than in a between-subjects design. Of course, in this design the number of words in each condition would need to be increased, and the training would become twice as long, however, an online training paradigm might . A combination of collaboration with a lab in a French-speaking area with online sessions and a within-subjects design might increase the number of participants in each linguistic group, increase the power, and make it more likely to find differences, if they exist, in training methods for individuals of different linguistic backgrounds when learning grammatical gender in a second language.

4.3 Conclusions

Overall, it was determined that the two-day training paradigm (two-alternative forced choice task) was at least somewhat effective at teaching nouns to the participants in both categories, as N400 results in response to semantic violations were consistent with noun learning. The physiological data was supported by accuracy ratings on the match/mismatch trials, which showed increases in accuracy ratings for semantic mismatches, but not for other mismatch categories. Differences in gender learning between training groups were not assessed in this study, but LAN-like effects were seen post-training in response to determiner violations. Behavioural data does not indicate that these effects are, in fact LANs, given that the accuracy ratings for determiner violations did not increase significantly post-training, and the scores on the gender assessment task did not increase more for congruently gendered pairs than incongruent ones, indicating no gender learning due to training. Moreover, the low N in this study means that all results should be interpreted within the context of pilot data.

Future directions for this study might include a change in design, a focus on francophone participants to determine if transfer effects would result in differences compared to the anglophone participants, and a longer, online, training paradigm.

References

Andringa and Curcic Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. *Perception & Psychophysics*, *58*(7), 992-1004. doi:10.3758/BF03206827

Beller, S., Brattebø, K., Lavik, K., et al. (2015). Culture or language: what drives effects of grammatical gender?. *Cognitive Linguistics*, 26(2), pp. 331-359. Retrieved 3 Jul. 2017, from doi:10.1515/cog-2014-0021

Birdsong, D., & Molis, M. (2001). On the Evidence for Maturational Constraints in Second-Language Acquisition. *Journal of Memory and Language*, *44* (2) , 235-249.

Bobb, S. C., Kroll, J. F., & Jackson, C. N. (2015). Lexical constraints in second language learning: Evidence on grammatical gender in German. *Bilingualism: Language And Cognition*, *18*(3), 502-523. doi:10.1017/S1366728914000534

Boroditsky L, Schmidt L, Phillips W. Sex, syntax and semantics. *Language in mind: Advances in the study of language and thought* [e-book]. Cambridge, MA, US: MIT Press; 2003:61-79.

Brice, A. E., & Brice, R. (2008). Examination of the critical period hypothesis and ultimate attainment among Spanish-English bilinguals and English-speaking monolinguals. *Asia Pacific Journal Of Speech, Language, And Hearing*, *11*(3), 143-160. doi:10.1179/136132808805297188

Chiswick, B.R. & Miller, P.W. (2008). A test of the critical period hypothesis for language learning. Journal of Multilingual and Multicultural Development, 29(1).

Christofides, L. N. & Swidinsky, R. (2008). The Economic Returns to a Second Official Language: English in Quebec and French in the Rest-of-Canada. , *IZA Discussion Paper No. 3551*

De Carli, F., Dessi, B., Mariani, M., Girtler, N., Greco, A., Rodriguez, G., & ... Morelli, M. (2015). Language use affects proficiency in Italian–Spanish bilinguals irrespective of age of second language acquisition. *Bilingualism: Language And Cognition*, *18*(2), 324-339. doi:10.1017/S1366728914000054

Delorme, A. & Makeig, S. (2004). "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics," *Journal of Neuroscience Methods, 134,* 9-21

Dowens, M. G., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish—Evidence from Event Related Potentials. *Neuropsychologia*, *49*(7), 1651-1659. doi:10.1016/j.neuropsychologia.2011.02.034

Flege and Yemi-Komshian (1999) Age Constraints on Second-Language Acquisition Journal of Memory and Language 41, 78–104

Foucart, A., & Frenck-Mestre, C. (2011). 'Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1-L2 syntactic similarity': Erratum. *Bilingualism: Language And Cognition*, *15*(1), 202. doi:10.1017/S1366728911000137

Foucart, A., & Frenck-Mestre, C. (2012). 'Can late l2 learners acquire new grammatical features? Evidence from ERPS and eye-tracking': Corrigendum. *Journal Of Memory And Language*, *67*(1), 238. doi:10.1016/j.jml.2012.02.009

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic

processes underlying the P600. *Language And Cognitive Processes*, *25*(2), 149-188. doi:10.1080/01690960902965951

Gramfort A., Luessi M., Larson E., Engemann D.A., Strohmeier D., Brodbeck C., Goj R., Jas M., Brooks T., Parkkonen L., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci*, *7*,. doi: 10.3389/fnins.2013.00267

Grosjean, F., Dommergues, J., Cornu, E., Guillelmon, D., & Besson, C. (1994). The gender-marking effect in spoken word recognition. *Perception & Psychophysics*, *56*(5), 590-598. doi:10.3758/BF03206954

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition*, *29*(3), 503-511. doi:10.3758/BF03196401

Hawkins, Roger & Cecilia Yuet-hung Chan. (1997). "The partial availability of Universal Grammar in second language acquisition: the `failed functional features hypothesis'". Second Language Research 13, 187-226

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60 - 99.

Kaan, E., & Swaab, T. Y. (2003). Repair, Revision, and Complexity in Syntactic Analysis: An Electrophysiological Differentiation. *Journal of Neuroscience, 15*(1), 98-110.

Kempe, V., Brooks, P. J., & Kharkhurin, A. (2010). Cognitive predictors of generalization of Russian grammatical gender categories. *Language Learning*, *60*(1), 127-153. doi:10.1111/j.1467-9922.2009.00553.x

Konishi T. (1993) The semantics of grammatical gender: a cross-cultural study. Journal of Psycholinguistic Research 22: 519-534.Kurinski, E., & Sera, M. D. (2011). Does learning Spanish grammatical gender change English-speaking adults' categorization of inanimate objects?. *Bilingualism: Language And Cognition*, *14*(2), 203-220. doi:10.1017/S1366728910000179

Kutas, M.; Hillyard, S. A. (1980). "Reading senseless sentences: Brain potentials reflect semantic incongruity". *Science, 207*, 203–208

Lemhöfer, K., Schriefers, H., & Hanique, I. (2010). Native language effects in learning second-language grammatical gender: A training study. *Acta Psychologica*, *135*(2), 150-158. doi:10.1016/j.actpsy.2010.06.001

Lenneberg, E.H. (1967). *Biological Foundations of Language*. Wiley.

Loerts, H., Stowe, L. A., & Schmid, M. S. (2013). Predictability speeds up the re-analysis process: An ERP investigation of gender agreement and cloze probability. *Journal Of Neurolinguistics*, *26*(5), 561-580. doi:10.1016/j.jneuroling.2013.03.003

Loftus, G. R. (1971). Comparison of recognition and recall in a continuous memory task. Journal of Experimental Psychology, 91(2), 220-226. DOI: 10.1037/h0031841

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*(Suppl 2), 123-150. doi:10.1111/j.1467-9922.2010.00604.x

McManus, K., & Marsden, E. (2016). L1 explicit instruction can improve l2 online and

offline performance. *Studies In Second Language Acquisition*, doi:10.1017/S027226311600022X

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. *Language Learning*, *60*(1), 154-193. doi:10.1111/j.1467-9922.2009.00554.x

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal Of Cognitive Neuroscience*, *24*(4), 933-947. doi:10.1162/jocn_a_00119

Osterhout, X. & HOLCOMB, X. (1992). Event-Related Brain Potentials Elicited by Syntactic Anomaly. *Journal of Memory and Language, 31*. 785-806.

Parkkonen, L., & Hamalainen, M. S. (2013). MNE software for processing MEG and EEG data. *Neuroimage, 86(*1*),* 446-460.

Peirce, J. W. (2009) Generating stimuli for neuroscience using PsychoPy. Front. Neuroinform. 2:10. doi:10.3389/neuro.11.010.2008

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German Vocabulary Through Reading: The Effect of Three Enhancement Techniques Compared. *Language Learning*, *59*(1), 113–151. http://doi.org/10.1111/j.1467-9922.2009.00502.x

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. Paper presented at the 25th Annual Conference of the Cognitive Science Society, Boston, MA.

Presson, N., MacWhinney, B., & Tokowicz, N. (2014). Learning grammatical gender: The use of rules by novice learners. *Applied Psycholinguistics*, *35*(4), 709-737. doi:10.1017/S0142716412000550

Pu, H., Holcomb, P. J., & Midgley, K. J. (2016). Neural changes underlying early stages of L2 vocabulary acquisition. *Journal of Neurolinguistics*, *40,* 55-65. doi:10.1016/j.jneuroling.2016.05.002

Sabourin, L. (2001). L1 effects on the processing of grammatical gender in L2. In S. Foster-Cohen, & A. Nizegorodcew (Eds.), EUROSLA YEARBOOK, 2001 (pp. 159-169). (EUROSLA YEARBOOK; Vol. 1). AMSTERDAM ME: John Benjamins Publishers.

Sabourin, L., Stowe, L. A., & De Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*,22(1), 1–29.

Sabourin, L., Stowe, L. A., & de Haan, G. J. (2016). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, *22(1)*. 1-29 doi:

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: evidence from pupillometry. *Frontiers in Psychology*, *5*, 137. http://doi.org/10.3389/fpsyg.2014.00137

Segel, E. & Boroditsky, L. (2011). Grammar in art. *Front. Psychology* 1:244. doi: 10.3389/fpsyg.2010.00244

Snow, C. E. & Hoefnagel-Höhle, M. (1978). The Critical Period for Language Acquisition:Evidence from Second Language Learning. Child Development, 49, 4 (Dec., 1978), 1114-1128

Statistics Canada. (2013). The evolution of English–French bilingualism in Canada from 1961 to 2011. *Retrieved from*

Steinhauer, K. (2014). Event-related potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics*, *35*(4), 393-417. doi:10.1093/applin/amu028

Steinhauer, K., & Connolly, J. F. (2008). Event-Related Potentials in the Study of Language. *Handbook of the Neuroscience of Language*, 91-104

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and language* , *120* (2), 135–62.

Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: evidence from event-related brain potentials. *Second Language Research*, *25* (1), 13–41.

Tremblay, A. (2013). LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. R package version 2.0. *Comprehensive R Archive Network* .

Weber-Fox , C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8* (3), 231-256, doi>10.1162/jocn.1996.8.3.231

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B), 73* (1), 3-36.

Appendix A

List of Auditory Stimuli in English, French, and the Artificial Language

English	**French**	**AL (IPA pronunciation)**	**French Gender**	**AL Gender**
Bird	Oiseau	Aloidia (əloɪdiʌ)	M	F
Spoon	Cuillère	Banafy (bᵊnæfi)	F	M
Turtle	Tortue	Boff (bɑf)	F	M
Tree	Arbre	Bronea (broniʌ)	M	M
Skunk	Moufette	Calan (kᵊlæn)	F	F
Fish	Poisson	Chack (tʃæk)	M	F
Orange	Orange	Claster (klæstr)	F	F
Store	Magasin	Clon (klɑn)	M	M
Raccoon	Raton laveur	Cocoji (kokodʒi)	M	F
Fork	Fourchette	Cug (kʌg)	F	M
Sled	Traineau	Dook (duk)	M	F
Ghost	Fantôme	Eap (ip^h)	M	M
Woman	Femme	Rup (rʌp)	F	F
Nun	Religieuse	Wald (wɑld)	F	F
Teacher	Enseignante	Sazu (sæzu)	F	F
Witch	Sorcière	Erf (ərf)	F	F
Fence	Clôture	Foxclore (fɑksklor)	F	F
Monkey	Singe	Grage (grɛɪdʒ)	M	F
Sheep	Mouton	Jeddy (dʒɛdi)	M	F
Book	Livre	Kall (kɑl)	M	M

House	Maison	Kaloolon (kʌlulɑn)	F	M
Squirrel	Écureuil	Kem (kɛm)	M	M
Horse	Cheval	Kevaro (kɛɪvɑro)	M	F
Garbage can	Poubelle	Lozo (lozo)	F	F
Cup	Tasse	Lum (lʌm)	F	M
King	Roi	Apamos (æpəmos)	M	M
Priest	Prêtre	Dobb (dob)	M	M
Fireman	Pompier	Fout (fʌut)	M	M
Wizard	Magicien	Mutron (mutʃrɑn)	M	M
Truck	Camion	Nade (nɛɪdv)	M	F
Ladder	Échelle	Neb (nɛb)	F	F
Lemon	Citron	Nuluzi (nuluzi)	M	M
Chair	Chaise	Oaroon (orun)	F	M
Bear	Ours	Powl (pʌuwˤl)	M	M
Branch	Branche	Praw (prɑ)	F	F
Frog	Grenouille	Roplixoo (roplɪksu)	F	M
Can	Canette	Rultat (rultæʔ)	F	F
Apple	Pomme	Sab (sæb)	F	M
Bathtub	Bain	Sertave (sˤrtɛɪv)	M	F
Blackboard	Tableau	Slace (slɛɪs)	M	M
Cake	Gâteau	Soan (son)	M	M
Church	Église	Viss (vɪs)	F	M
Goat	Chèvre	Vodrine (vodʒrin)	F	F

Road	Rue	Yolle (yol)		F	F

Appendix B

Images for Learning Task

BIRD

SPOON

TURTLE

TREE

SKUNK

FISH

ORANGE

STORE

RACCOON

FORK

SLED

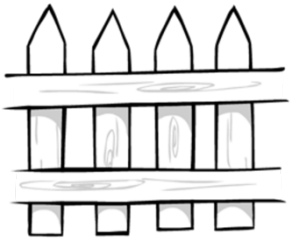
GHOST

WOMAN

NUN

TEACHER

WITCH



FENCE

MONKEY

SHEEP

BOOK

HOUSE

SQUIRREL

HORSE

GARBAGE CAN

CUP

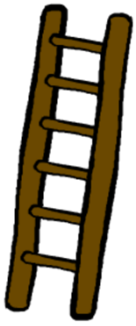
KING

PRIEST

FIREMAN

WIZARD

TRUCK

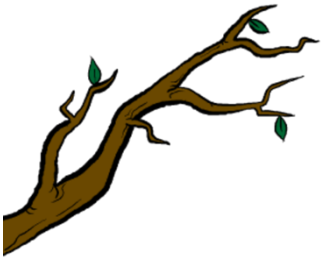


LADDER

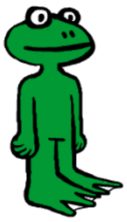
LEMON

CHAIR

BEAR



BRANCH

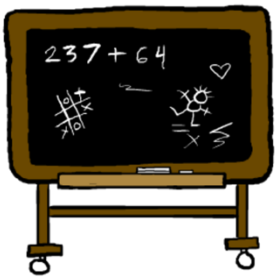


FROG

CAN

APPLE

BATHTUB



BLACKBOARD

CAKE

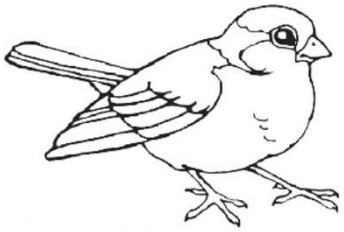
CHURCH

GOAT

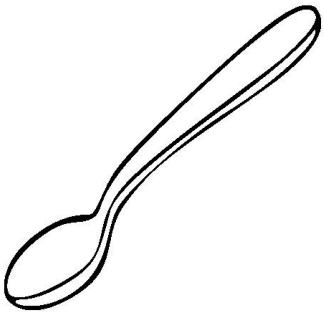
ROAD

Appendix C

Line Drawings



BIRD

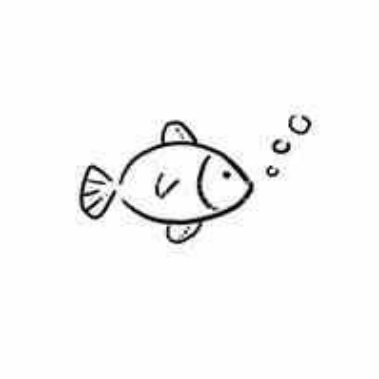


SPOON

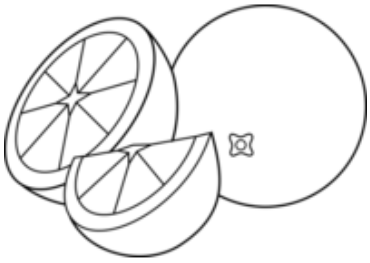
TURTLE

TREE

SKUNK



FISH

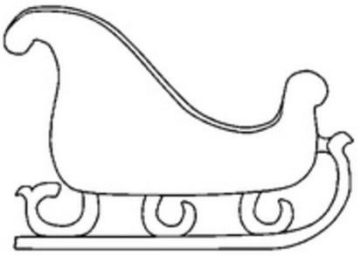


ORANGE

STORE

RACCOON

FORK



SLED

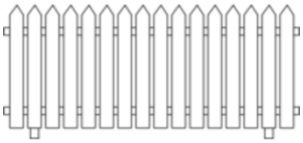
GHOST

GIRL

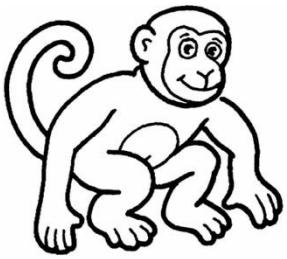
NUN

TEACHER

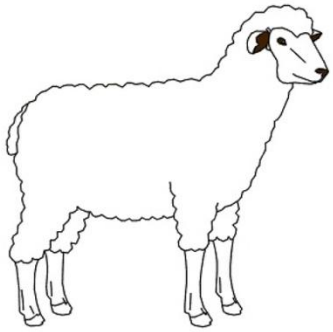
WITCH



FENCE



MONKEY

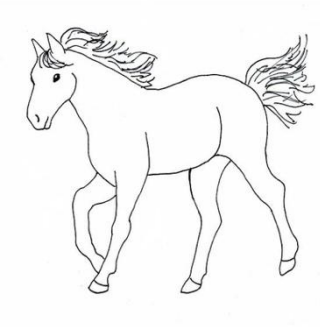


SHEEP

BOOK

HOUSE

SQUIRREL



HORSE

GARBAGE CAN

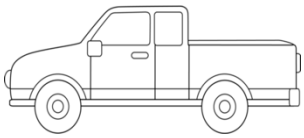
CUP

KING

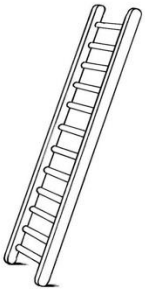
PRIEST

FIREMAN

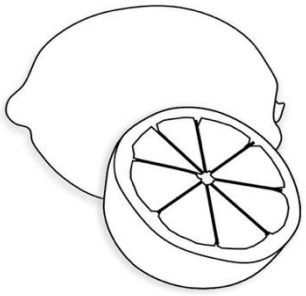
WIZARD



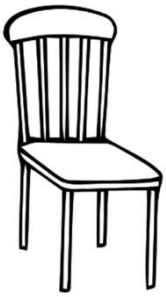
TRUCK



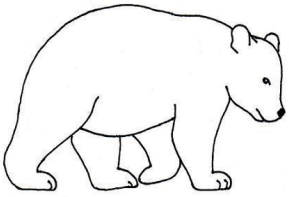
LADDER



LEMON

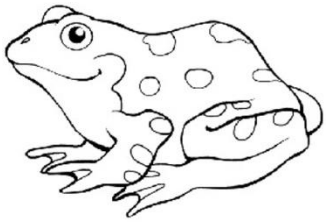


CHAIR

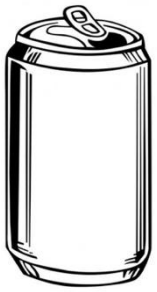


BEAR

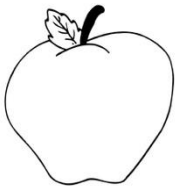
BRANCH



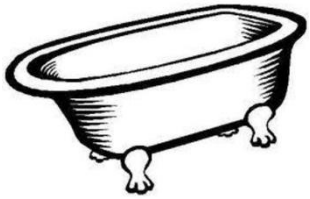
FROG



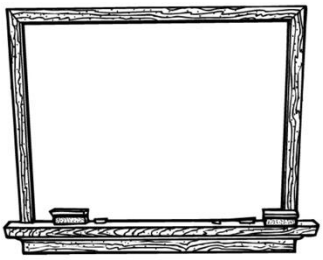
CAN



APPLE



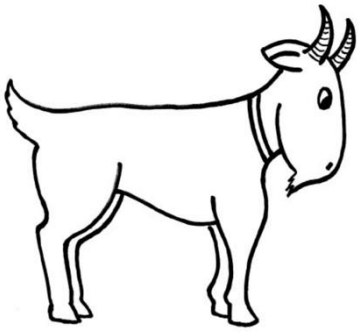
BATHTUB



BLACKBOARD

CAKE

CHURCH



GOAT

ROAD

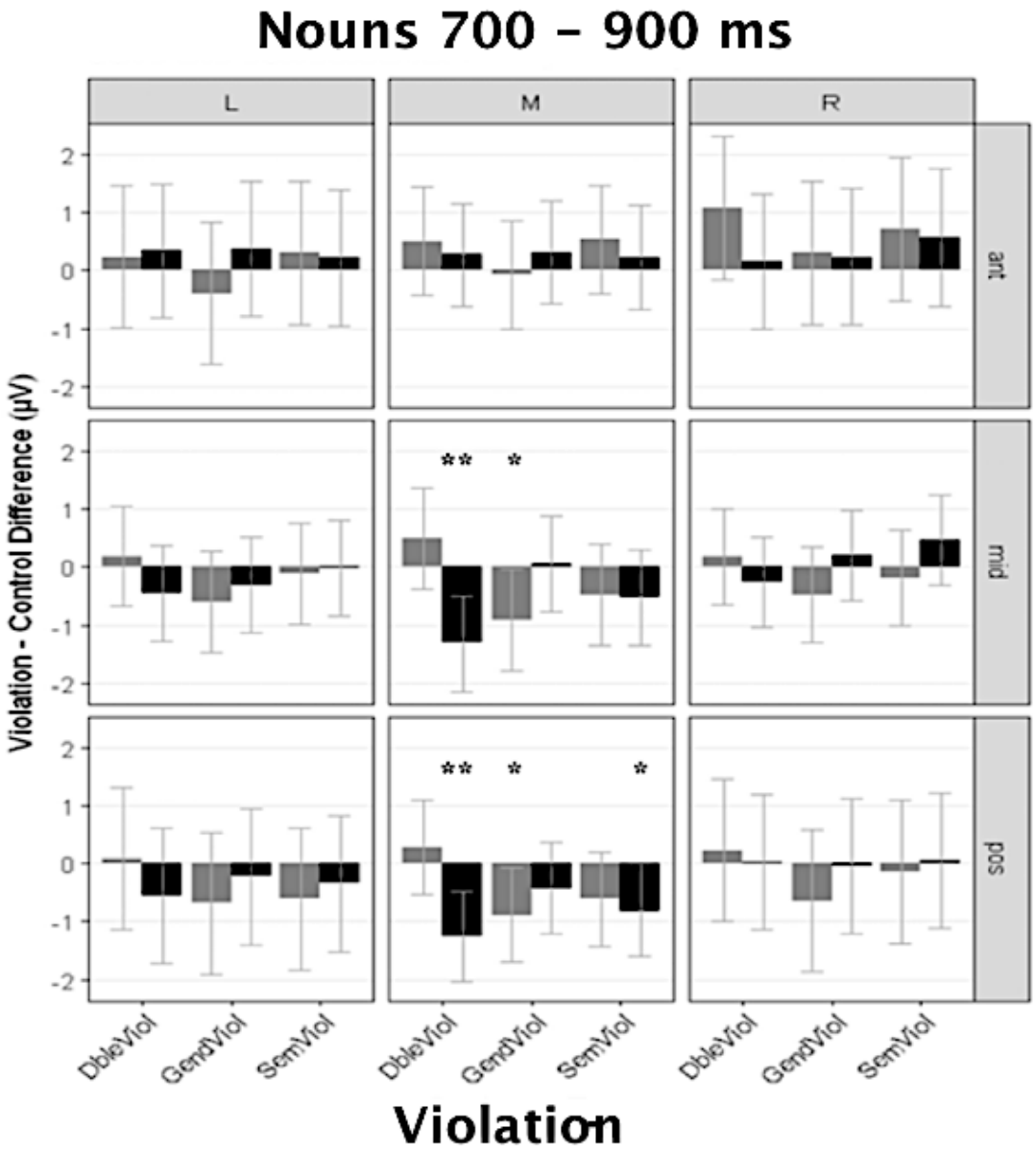


Figure 3.4. *Model-Estimated Values for 700-900ms ERP effects on nouns.*
*p<.05; ** p<.001

3.2.2.2 Determiners

Two models met the criterion for "best model" (lowest AIC value) for the data from determiners in the 700-900 ms time window: the full 3-way interaction model, and the model containing only all 2-way interactions. Because we defined the optimal model as the one having the lowest AIC value and the least number of terms, we took the

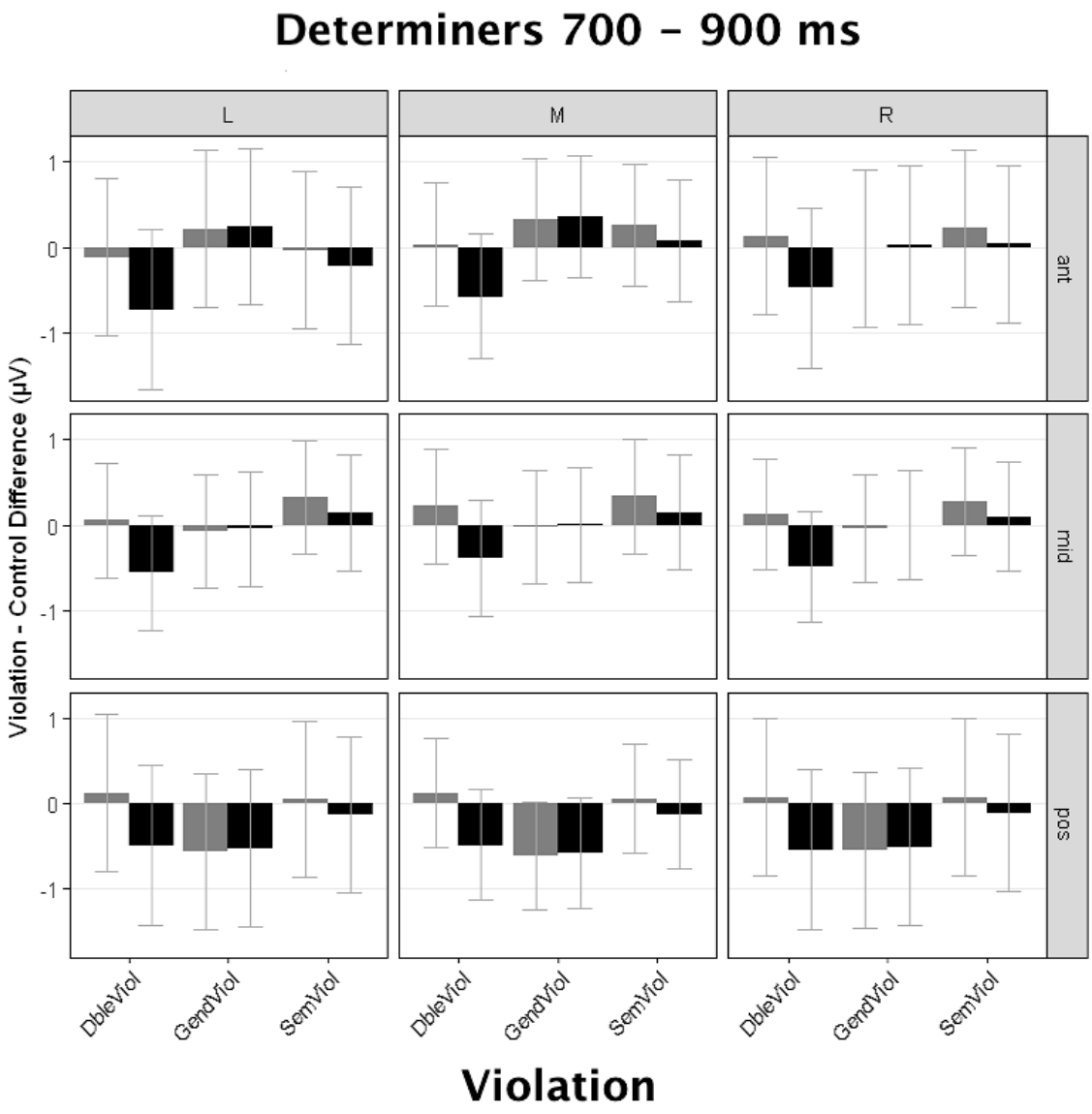


Figure 3.5. *Model-Estimated Values for 700-900ms ERP effects on determiners.* *p<.05; ** p<.0013

Visual inspection of the waveforms pre- and post-training in response to determiners show small but apparent differences for semantic violations, with negativity being slightly higher pre-training. Even smaller differences in waveforms are seen for the gender violation and double violation conditions, with negativity being slightly increased post-training (Figure 3.6). Visual inspection of the waveforms pre- and post-

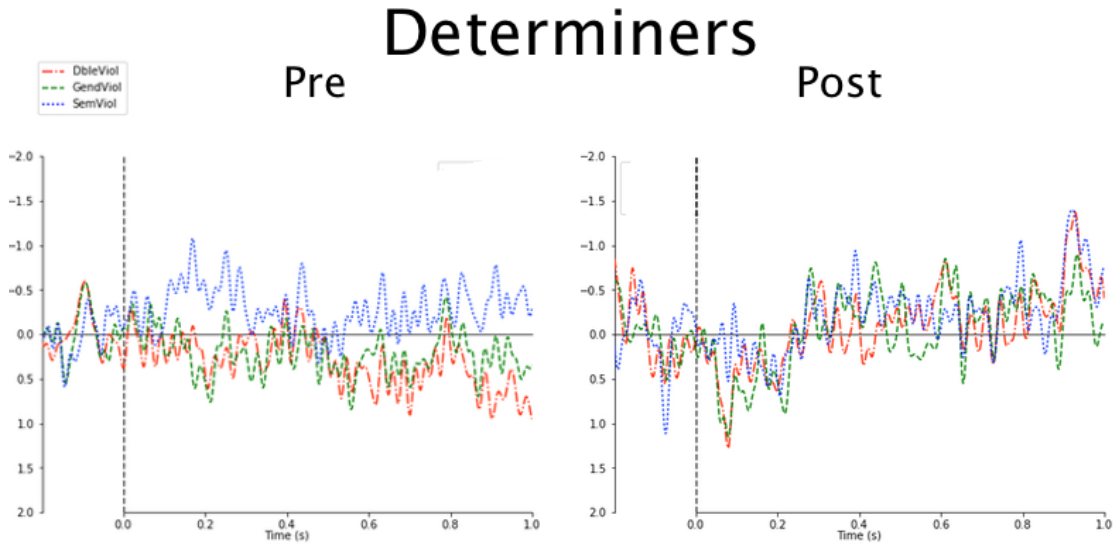


Figure 3.6. *Waveform differences across conditions pre- and post-training in response to determiners.*

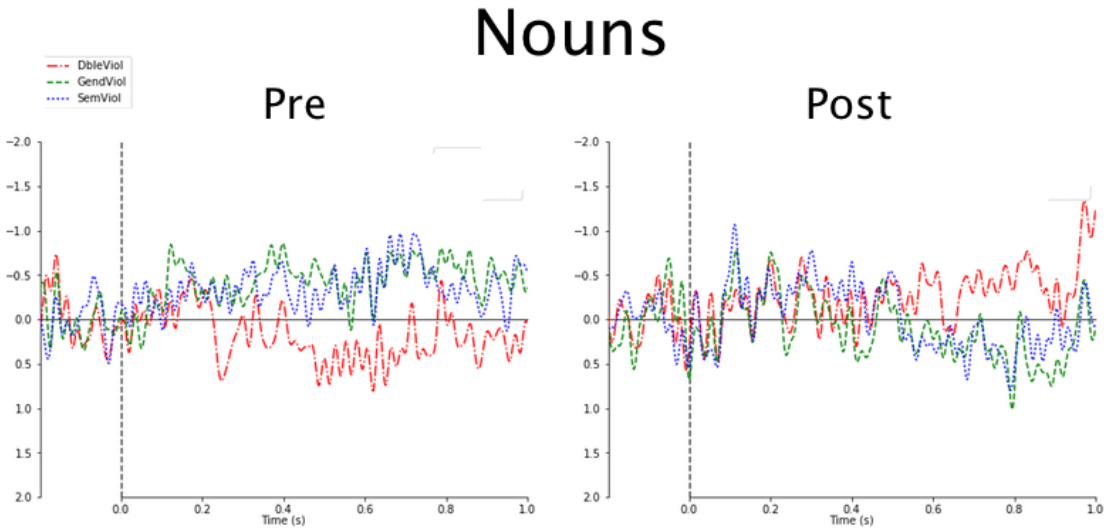


Figure 3.7. Waveform differences across conditions pre- and post-training in response to nouns.

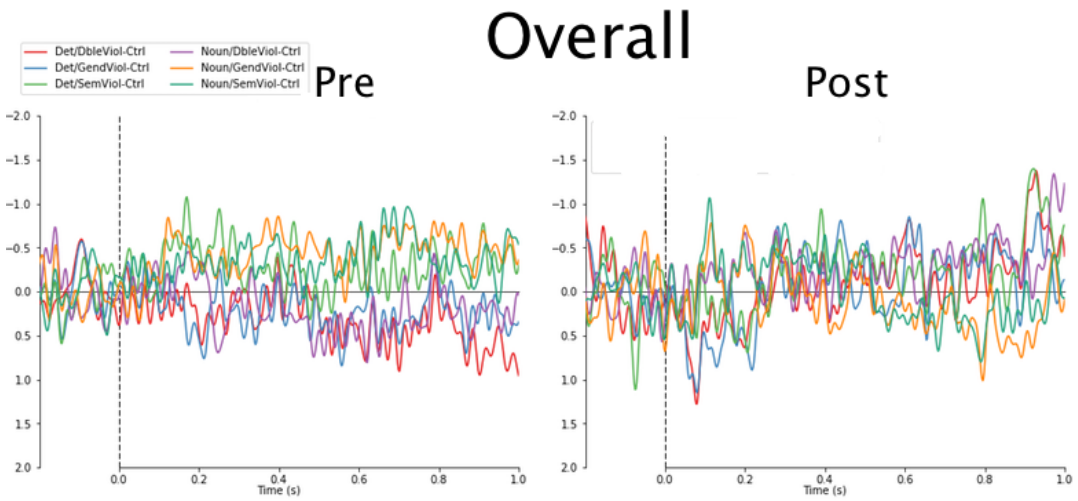


Figure 3.8. *Overall waveform differences across conditions pre- and post-training.*

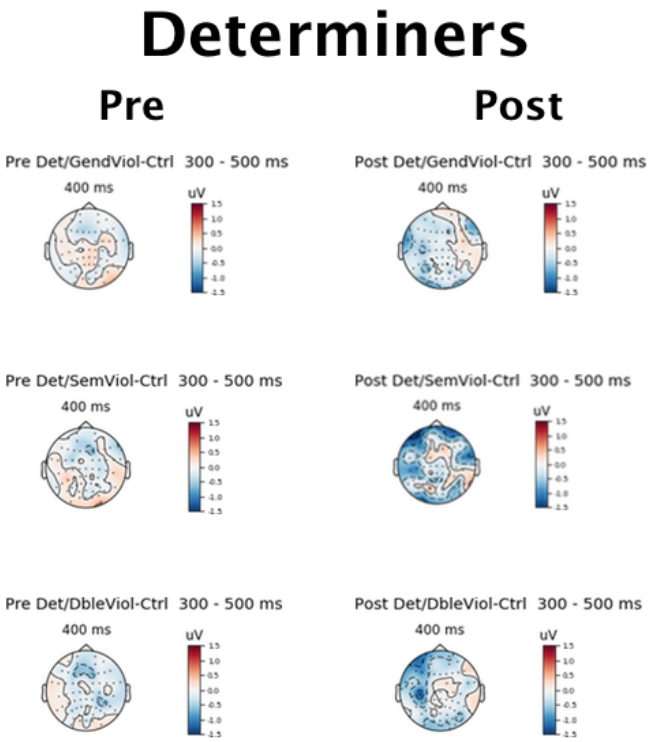


Figure 3.9.
Topographical plots of average ERP scalp distribution at 400 ms for determiners.

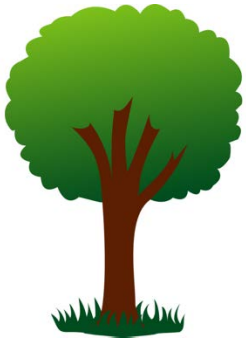
Appendix B

Images for Learning Task

BIRD

SPOON

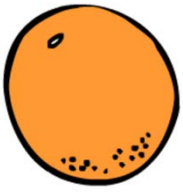
TURTLE



TREE

SKUNK

FISH

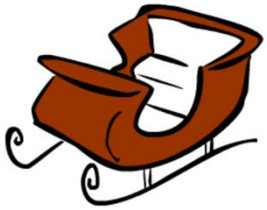


ORANGE

STORE

RACCOON

FORK



SLED

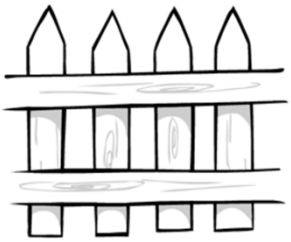
GHOST

WOMAN

NUN

TEACHER

WITCH



FENCE

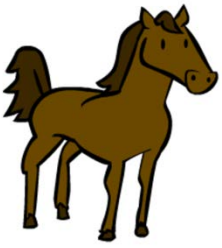
MONKEY

SHEEP

BOOK

HOUSE

SQUIRREL



HORSE

GARBAGE CAN

CUP

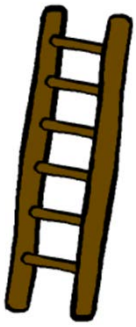
KING

PRIEST

FIREMAN

WIZARD

TRUCK

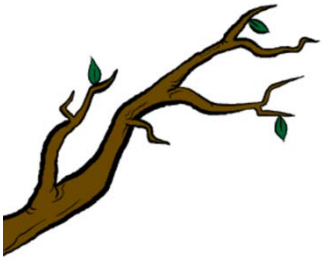


LADDER

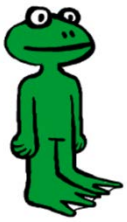
LEMON

CHAIR

BEAR



BRANCH

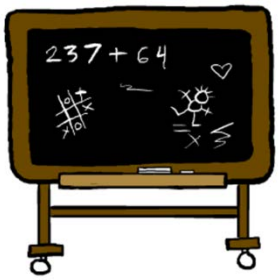


FROG

CAN

APPLE

BATHTUB

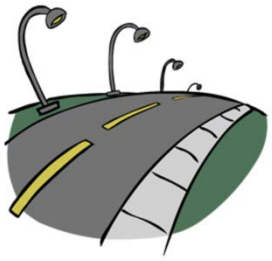


BLACKBOARD

CAKE

CHURCH

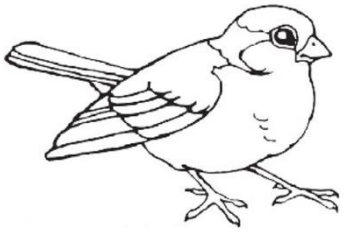
GOAT



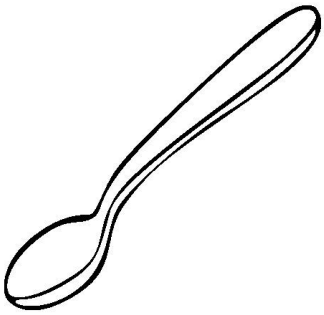
ROAD

Appendix C

Line Drawings



BIRD

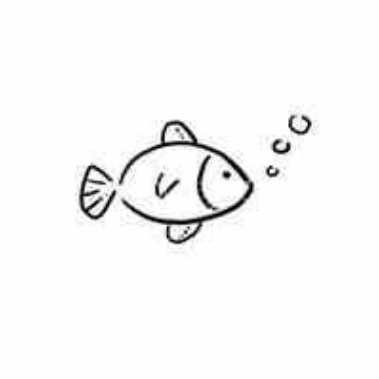


SPOON

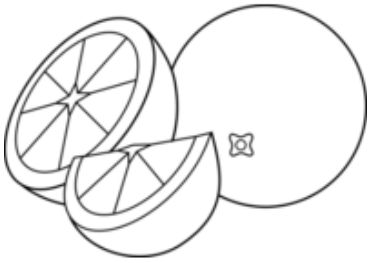
TURTLE

TREE

SKUNK



FISH

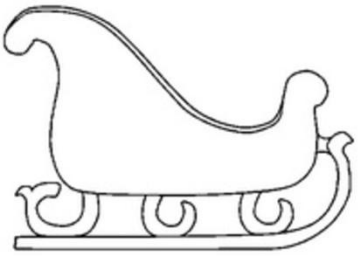


ORANGE

STORE

RACCOON

FORK



SLED

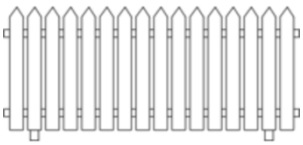
GHOST

GIRL

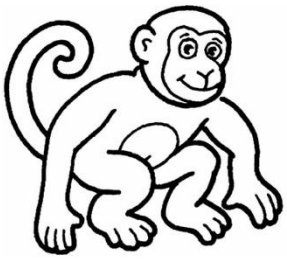
NUN

TEACHER

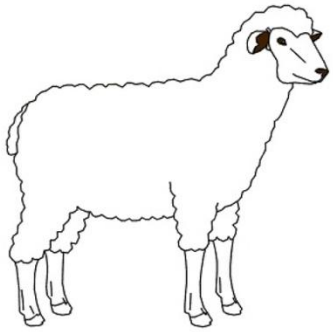
WITCH



FENCE



MONKEY

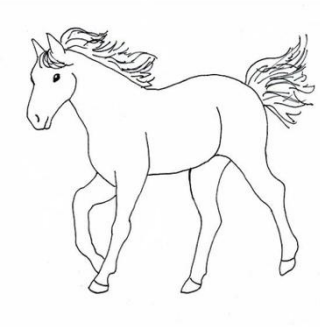


SHEEP

BOOK

HOUSE

SQUIRREL



HORSE

GARBAGE CAN

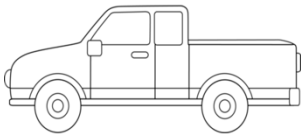
CUP

KING

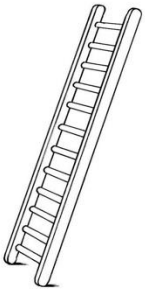
PRIEST

FIREMAN

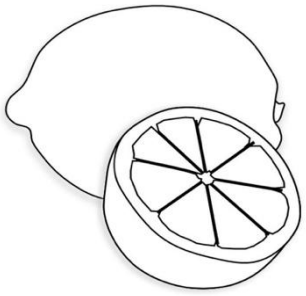
WIZARD



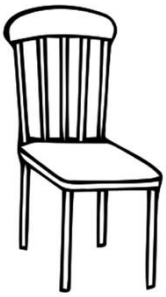
TRUCK



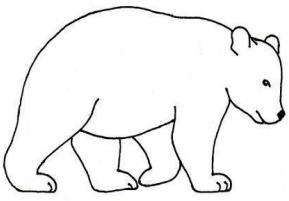
LADDER



LEMON

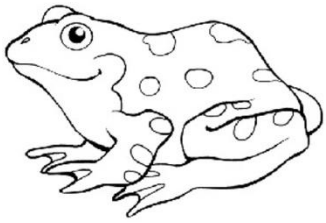


CHAIR



BEAR

BRANCH



FROG