WHY WAIT? PSYCHOLINGUISTIC INVESTIGATIONS OF THE ROLES OF LEARNING CONDITION AND GENDER STABILITY IN L2 GENDER-BASED ANTICIPATION

BY

KAILEN THOMAS WILLIAM SHANTZ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
with a concentration in Second Language Acquisition & Teacher Education
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Assistant Professor Darren Tanner, Chair
Associate Professor Duane Watson, Vanderbilt University
Professor Kara Federmeier
Professor Silvina Montrul

ABSTRACT

It is well documented that grammatical gender poses a pervasive problem for adult second language learners. Whereas native speakers can use prenominal grammatical gender marking to anticipate upcoming nouns in sentences, L2 learners often show a reduced or absent ability to use gender in this manner (Grüter, Lew-Williams, & Fernald, 2012; Hopp, 2013, 2016). The Lexical Gender Learning Hypothesis (LGLH) proposes a chain of causality to account for this finding: 1) Differences in the conditions under which children and adults learn a language lead to weaker links between nouns and their gender representations for adult L2 learners; 2) These weaker links lead to greater variability in gender assignment; 3) This increased variability in gender assignment reduces the extent to which adult L2 learners use gender predictively. Across three experiments, this dissertation provides the first direct test of the LGLH. Results do not find evidence for the claim that learning context affects the stability of gender assignments nor the ability to use gender as an anticipatory cue. The data do, however, support the hypothesis that gender assignment variability modulates the anticipatory use of gender marking. These findings indicate that L2 knowledge plays an important role in online L2 processing, and that failure to adequately account for this knowledge may lead to an underestimation of L2 performance.

ACKNOWLEDGEMENTS

There are many people who've helped me get here, and while I can't thank all of them by name, I want to start by acknowledging this fact. If I've been successful, it's only because I've enjoyed the support and friendship of many great people over the past six years of graduate school. So, thank you to the friends, family, instructors, mentors and all others who I've been privileged to know along the way.

Thanks in particular to my partner, Phil. Your enthusiasm for science and thirst for knowledge have been a constant inspiration to me and I am both a better person and scientist because of you.

I must also thank Amanda Kim, Anqi Hu (Jojo), Kari Schwink, and Bhasvera (Shayne) Chammavanijakul for their hard work in helping me to collect and annotate a large portion of the data presented in my dissertation. You were indispensable, and I was lucky to have such dedicated people working with me on this project.

Thanks, as well, to Michelle Sims, both for your friendship and for generously lending me your time and expertise in preparing my acoustic stimuli. I would also like to thank the two anonymous individuals who patiently allowed me to record them saying the weird sentences that eventually became my experimental stimuli.

I also want to thank Dr. Holger Hopp for helping to organize my research trip to Germany and for providing me with lab space and support during this trip. Thanks, also, to Dr. Carrie Jackson and Dr. Janet Van Hell for their assistance in organizing my visit to Penn State for data collection.

Thank you to my committee members, Dr. Duane Watson, Dr. Kara Federmeier and Dr. Silvina Montrul. Your feedback at the various stages of this dissertation has been invaluable in shaping both my research and my thinking, and your encouragement and support throughout the years I’ve known each of you has meant a lot.

Above all others, I am incredibly grateful to my advisor, Dr. Darren Tanner. My decision to come to UIUC was very difficult to make, but your unwavering support and mentorship have helped me to not simply “get through” grad school, but to exceed all expectations I had for myself and for my time here. Beyond teaching me how to do quality science, you’ve also modeled what exemplary mentorship is. I wouldn’t be who I am without your guidance.

Finally, thank you to the National Science Foundation for funding my research (NSF BCS-1627808), to the German Academic Exchange Service for helping to fund my research trip to Germany, and to the Social Sciences and Humanities Research Council of Canada and the University of Illinois for financial support during my degree.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.. 1

1.1. GRAMMATICAL GENDER... 2

1.1.1. L1 processing of grammatical gender.. 4

1.1.2. L2 processing of grammatical gender.. 8

1.2. PREDICTIVE LANGUAGE PROCESSING 13

1.2.1. Evidence for graded prediction 14

1.2.2. Prediction versus facilitated integration 17

1.2.3. Factors modulating prediction 22

1.3. THE LEXICAL GENDER LEARNING HYPOTHESIS... 28

1.4. RESEARCH QUESTIONS 30

1.5. DESCRIPTION OF EXPERIMENTS 31

1.5.1. Experiments 1 and 2 31

1.5.2. Experiment 3.. 32

CHAPTER 2: LEARNING CONTEXT AND GENDER-BASED ANTICIPATORY EYE-MOVEMENTS IN AN ARTIFICIAL GRAMMAR.. 34

2.1. EXPERIMENT 1.. 35

2.1.1. Participants 35

2.1.2. Materials... 36

2.1.3. Procedure 37

2.1.3.1. Training Task 37

2.1.3.2. Visual world eye tracking task.. 38

2.1.3.3. Gender assignment tasks... 40

2.1.4. Data preprocessing and analysis 41

2.1.5. Hypotheses.. 49

2.1.6. Results... 50

2.1.6.1. Gender assignment performance 50

2.1.6.2. Eye tracking results 51

2.1.7. Experiment 1 summary and discussion 61

2.2. EXPERIMENT 2 67

2.2.1. Participants 67

2.2.2. Materials 67

2.2.3. Procedure 67

2.2.4. Data preprocessing and analysis 68

2.2.5. Hypotheses 68

2.2.6. Results 69

2.2.6.1. Gender assignment performance 69

2.2.6.2. Eye tracking results 70

2.2.7. Experiment 2 summary 78

2.3. DISCUSSION OF EXPERIMENTS 1 AND 2 79

CHAPTER 3: NEURAL INDICES OF GENDER-BASED ANTICIPATION IN L1 AND L2 GERMAN 84

3.1. PARTICIPANTS 85

3.2. MATERIALS 86

3.3. PROCEDURE 88

3.4. DATA ACQUISITION AND ANALYSIS 91

3.5. HYPOTHESES 97

3.6. RESULTS 99

3.6.1. Lateralized picture monitoring task 99

3.6.1.1. Behavioral results 99

3.6.1.2. ERP results 102

3.6.2. Gender decision task 105

3.7. DISCUSSION OF EXPERIMENT 3 108

3.7.1. Lateralized picture monitoring task 108

3.7.2. Gender decision task 113

CHAPTER 4: GENERAL DISCUSSION... 115

CHAPTER 5: CONCLUSION.. 120

CHAPTER 6: REFERENCES.. 121

APPENDIX A. ITEMS FOR EXPERIMENTS 1 AND 2.. 137

APPENDIX B. ITEMS FOR EXPERIMENT 3.. 138

CHAPTER 1: INTRODUCTION

Of the twelve extant languages with the highest enrollment numbers in US institutions of higher education, eight have grammatical gender (Goldberg, Looney, & Lusin, 2015). Accounts of grammatical gender have proposed that it aids language use by facilitating reference tracking (Heath, 1975; Zubin & Köpcke, 1986) or by making nouns more predictable in context (Bates, Elman, & Li, 1994; Dye, Milin, Futrell, & Ramscar, 2017). Evidence consistent with the latter proposal has accumulated over recent decades for native speakers of languages with gender. Prenominal gender marking has been shown to facilitate processing of nouns for native speakers in auditory gating and lexical decision (Grosjean, Dommergues, Cornu, Guillelmon, & Besson, 1994), shadowing (Guillelmon & Grosjean, 2001), visual search (Dahan, Swingley, Tanenhaus, & Magnuson, 2000) and visual world paradigms (e.g. Lew-Williams & Fernald, 2007, 2010). Despite the processing benefits conferred by gender marking, adult second language (L2) learners do not reliably make use of gender marking as an anticipatory cue, even at high levels of proficiency. The source of this reduced use of gender marking, moreover, remains unclear. Whereas earlier proposals based in linguistic theory suggested that deficits in grammatical representation may be to blame (Franceschina, 2005; Hawkins & Franceschina, 2004), more recent evidence has linked L2 deficits to variable accuracy in gender assignment (i.e. linking nouns to their associated gender features), resulting in reduced and variable use of grammatical gender for anticipatory processing (Hopp, 2013, 2016). This and related findings have led to the Lexical Gender Learning Hypothesis (LGLH; Grüter et al., 2012; Hopp, 2013, 2016), which posits a chain of causality where key differences in the learning conditions for children compared to adults lead to weaker links between nouns and their corresponding gender representations (cf.

Gollan, Montoya, Cera, & Sandoval, 2008). These weaker links result in greater variability in gender assignment and consequently reduce or eliminate the use of gender marking as an anticipatory cue to an upcoming noun. Though there is evidence for various aspects of this hypothesis (Arnon & Ramscar, 2012; Grüter et al., 2012; Hopp, 2013, 2016; Montrul, Davidson, De la Fuente, & Foote, 2014; Siegelman & Arnon, 2015), no study has yet systematically manipulated learning conditions in order to examine how learning context impacts both gender assignment and anticipatory processing. In addition, no existing studies have obtained multiple measures of gender assignments to individual nouns; consequently, the role of gender assignment variability in previous studies cannot be reliably assessed. The goal of this dissertation is therefore to directly test the claims of the LGLH in order to arrive at a more comprehensive understanding of what makes using grammatical gender predictively so challenging for adult L2 learners.

1.1. Grammatical gender

Grammatical gender is a linguistic phenomenon present in many of the world's languages. It refers to a property of nouns whereby they can be grouped into classes such that the words expressing agreement with nouns exhibit mutually exclusive sets of properties across these groupings (Corbett, 1991; Hockett, 1958). That is, words that agree with a noun of one class will take a set of forms that are not taken by words agreeing with nouns of a different noun class. For example, determiners and attributive adjectives in German must agree in gender with the noun they are associated with. In the nominative case, a definite determiner takes the form *der* if it agrees with a singular masculine noun, *die* for singular feminine nouns, and *das* for singular neuter nouns. None of these forms may occur with nouns of other gender classes for definite, singular, nominative nouns, and hence the mutual exclusivity of agreement patterning which allows different gender classes to be identified.

To account for gender as a linguistic phenomenon, two prevailing accounts have been proposed in the functionalist tradition. The first account proposes that grammatical gender facilitates the tracking of referents in discourse by providing an additional cue to help resolve referential ambiguity (e.g. Heath, 1975; Zubin & Köpcke, 1986). In other words, when there are multiple possible referents of an anaphoric expression, grammatical gender agreement can help to disambiguate between possible referents when they belong to different gender classes. This possibility finds support from a set of self-paced reading studies which found that anaphoric expressions beginning with a gender-marked pronoun are read faster when the gender marking disambiguates between two possible antecedents relative to when the antecedents have the same grammatical gender (Carreiras, Garnham, & Oakhill, 1993; Garnham, Oakhill, Ehrlich, & Carreiras, 1995).

The second account proposes that grammatical gender facilitates comprehension by making nouns more predictable (Bates, Devescovi, Hernandez, & Pizzamiglio, 1996; Bates et al., 1994; Dye et al., 2017; Grosjean et al., 1994). That is, when a noun is preceded by gender marking, it allows a comprehender to form an expectation for a noun belonging to a specific gender class rather than any noun in their language. By limiting the scope of possible nouns to a subset of the lexicon, gender marking thus makes an impending noun less surprising relative to when a noun is not preceded by any gender marking. Evidence consistent with this latter proposal has accumulated over recent decades for native speakers of languages with gender. Prenominal gender marking has been shown to facilitate processing of nouns for native speakers in auditory gating and lexical decision (Grosjean et al., 1994), shadowing (Bates et al., 1996;

Guillelmon & Grosjean, 2001), visual search (Dahan et al., 2000), and visual world paradigms (Dussias et al., 2013; Grüter et al., 2012; Hopp, 2013, 2016; Hopp & Lemmerth, 2016; Huettig & Brouwer, 2015; Huettig & Janse, 2016; Lew-Williams & Fernald, 2007, 2010; Loerts, Wieling, & Schmid, 2013; Lundquist, Rodina, Sekerina, & Westergaard, 2016). Recent corpus work, moreover, has shown that German nouns have lower entropy (i.e. they are less surprising) following informative gender-marked determiners relative to when they are preceded by uninformative determiners (Dye et al., 2017). This lends further support to the idea that the facilitation observed for nouns that are preceded by gender marking reflects predictive processing. This and other aspects of grammatical gender processing will be taken up in more detail in the following sections.

1.1.1. L1 processing of grammatical gender

The dominant account of how grammatical gender is represented in the mental lexicon holds that language learners form abstract gender nodes for each of the gender classes in their language, and that all nouns belonging to a given gender class are connected to the gender node for that class (Levelt, Roelofs, & Meyer, 1999; see Schriefers & Jescheniak, 1999 for a review of the evidence). While there is general agreement that gender information is likely represented in an abstract form, there is disagreement in the literature about whether gender information about a noun can be accessed via form-level information in native production (compare Caramazza & Miozzo, 1997 and Levelt et al., 1999). There is, however, an abundance of evidence from comprehension tasks showing that access to grammatical gender is affected by the probabilistic relationship between a noun's form and the gender classes of a language (Bates, Devescovi, Pizzamiglio, D'amico, & Hernandez, 1995; Gollan & Frost, 2001; Holmes & Dejean De La

Bâtie, 1999; Holmes & Segui, 2004; Montrul et al., 2014; Peereman, Dufour, & Burt, 2009; Spalek, Franck, Schriefers, & Frauenfelder, 2008; Taft & Meunier, 1998). In particular, native speakers tend to be faster and more accurate at making gender decisions to nouns whose form is highly predictive of the noun's gender class (e.g. feminine Spanish nouns ending in -*a*) relative to nouns whose form is not strongly predictive of any gender class (e.g. feminine Spanish nouns ending in -*l*) or nouns whose form is strongly predictive of a gender class different from that of the noun (e.g. feminine Spanish nouns ending in -*o*). In light of such findings, Gollan and Frost (2001) proposed that there are two mechanisms by which grammatical gender information is accessed: a direct route through which abstract gender representations are accessed via lemma-level representations, and a form-based route in which gender information is computed based on probabilistic relationships between a noun's form and the gender classes of a language.

Regardless of the mechanisms by which gender information is retrieved, an abundance of research has shown that L1 speakers of languages with grammatical gender are highly sensitive to the relationship between nouns and gender-marked determiners and adjectives in agreement relations with them. In self-paced reading, L1 Spanish speakers show increased reading times on post-nominal adjectives whose gender is incongruent with a preceding noun (Foote, 2011; Sagarra & Herschensohn, 2010). Results from eye tracking further find that effects of gender congruence manifest most robustly on measures of late processing, but not on measures of early processing (Deutsch & Bentin, 2001; Keating, 2009). This suggests that the disruption to reading experienced by native-speakers when encountering a grammatical gender violation may reflect reanalysis or revision of a sentence. Consistent with this, event-related potential (ERP) studies find that the P600, which has been hypothesized to reflect processes of extended analysis, reanalysis, revision, or integration difficulty (Brouwer, Fitz, & Hoeks, 2012; Friederici, 2002;

Kaan, Harris, Gibson, & Holcomb, 2000; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Osterhout, Holcomb, & Swinney, 1994), is reliably elicited in native speakers upon apprehension of a grammatical gender violation in sentences (Alemán Bañón, Fiorentino, & Gabriele, 2012; Alemán Bañón & Rothman, 2016; Barber & Carreiras, 2005; Davidson & Indefrey, 2009; Demestre & García-Albea, 2007; Gunter, Friederici, & Schriefers, 2000; Hagoort & Brown, 1999; Hagoort, 2003; Lemhöfer, Schriefers, & Indefrey, 2014; Molinaro, Vespignani, & Job, 2008; Nevins, Dillon, Malhotra, & Phillips, 2007; O'Rourke & Van Petten, 2011; Sabourin & Stowe, 2008; Wicha, Moreno, & Kutas, 2004). When gender agreement violations occur in isolated word pairs, however, native speakers exhibit N400 effects rather than P600 effects (Barber & Carreiras, 2005; Barber & Carreiras, 2003).

Native speakers' sensitivity to gender dependencies has also been shown to extend to predictive processing. In highly constraining sentences, native speakers of languages with grammatical gender can anticipate not only a specific noun, but have also been shown to anticipate the grammatical gender of an expected noun (Foucart, Martin, Moreno, & Costa, 2014; Foucart, Ruiz-Tada, & Costa, 2015; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003; Wicha et al., 2004; Wicha, Moreno, & Kutas, 2003). There is also strong evidence that native speakers use prenominal gender marking to anticipate upcoming nouns. Native speakers identify nouns faster when they are preceded by gender-marked versus -unmarked articles (Grosjean et al., 1994). They are also faster to repeat an auditorily presented noun when it is preceded by a congruent gender-marked article or adjective relative to when it is preceded by a gender-neutral item (Bates et al., 1996; Guillelmon & Grosjean, 2001). In looking-while-listening and eye tracking, native speakers orient faster to a noun when it is preceded in an utterance by an informative gender-marked determiner relative to when gender marking on the determiner is uninformative about which noun will be uttered (Dussias et al., 2013; Grüter et al., 2012; Hopp, 2013, 2016; Huettig & Brouwer, 2015; Huettig & Janse, 2016; Lew-Williams & Fernald, 2007, 2010; Loerts et al., 2013; Lundquist et al., 2016). Recent work has further shown that native speakers also use gender marking on pre-nominal adjectives to anticipate upcoming nouns (Hopp & Lemmerth, 2016). In addition, native speakers have been found to use gender marking to rule out gender-incongruent phonological competitors during comprehension (Dahan et al., 2000).

Summarizing, native speakers are highly sensitive to agreement relationships involving grammatical gender, and use grammatical gender marking to aid processing when it serves as a reliable cue to an upcoming word. A likely mechanisms by which pre-nominal gender marking facilitates the processing of impending nouns is through pre-activation of nouns that belong to the gender class denoted by a prenominal gender-marking element, which in turn facilitates lexical access upon apprehension of the noun (Bates et al., 1996, 1994; Grosjean et al., 1994; Guillelmon & Grosjean, 2001). It has been argued that such a mechanism would be of limited utility given the large number of nouns that belong to gender classes in a given language (Bölte & Connine, 2004), and that this would be a computationally expensive mechanism (Friederici & Jacobsen, 1999). However, these arguments were based on considerations of the predictive value of isolated sequences of a gender-marked element and a noun (e.g. determiner + noun), which ignores the fact that comprehenders usually encounter nouns in communicative contexts in which top-down information from the discourse, sentences and event knowledge may constraint the number of plausible nouns at any given time. Thus, these arguments likely underestimate the utility of gender marking as an anticipatory cue in real-world communication.

1.1.2. L2 processing of grammatical gender

Like native speakers, late L2 learners of languages with grammatical gender show sensitivity to agreement relationships. In offline comprehension tasks probing knowledge of grammatical gender, advanced L2 learners regularly show ceiling performance, demonstrating knowledge both of gender assignment and of gender agreement (Alarcón, 2011; Grüter et al., 2012; McCarthy, 2008; Montrul, Foote, & Perpiñán, 2008; White, Valenzuela, Kozlowska–Macgregor, & Leung, 2004). In online studies, L2 speakers show native-like increases in reading times on post-nominal adjectives that do not agree in gender with their preceding noun (Foote, 2011; Keating, 2009; Sagarra & Herschensohn, 2010). ERP studies have further found that L2 learners can show qualitatively native-like P600 responses to gender agreement violations (Alemán Bañon, Fiorentino, & Gabriele, 2014; Alemán Bañon, Miller, & Rothman, 2017; Dowens, Guo, Guo, Barber, & Carreiras, 2011; Dowens, Vergara, Barber, & Carreiras, 2010; Foucart & Frenck-Mestre, 2011, 2012; Gabriele, Fiorentino, & Alemán Bañón, 2013; Meulman, Wieling, Sprenger, Stowe, & Schmid, 2015; Tokowicz & MacWhinney, 2005). Together, these findings indicate that late L2 learners can successfully acquire appropriate representations for grammatical gender, and can use this information online to rapidly compute gender agreement. This further suggests that the difficulty L2 learners experience with grammatical gender does not likely stem solely from a representational deficit for gender features, as earlier hypotheses had suggested (Franceschina, 2005; Hawkins & Franceschina, 2004; Hawkins & Chan, 1997; Tsimpli & Dimitrakopoulou, 2007).

Recent evidence, in fact, indicates that the difficulty late L2 learners experience with grammatical gender may partly reflect variability in the stability of gender representations for individual nouns (i.e. the strength of the association between a noun and its corresponding gender node). In an ERP study with late L1-German learners of Dutch, Lemhöfer et al. (2014) found no P600 effects for objective gender violations, but did find P600 effects when trials were re-sorted based on participants' subjective gender assignments as determined in an offline gender assignment task (see Lewis, Lemhöfer, Schoffelen, & Schriefers, 2016 for similar findings in the modulation of neural oscillations). This indicates that item-specific knowledge of nouns' grammatical genders is a key factor that determines whether late L2 learners may exhibit qualitatively native-like processing of gender agreement. The role of item-specific knowledge in grammatical gender processing is further underscored by recent findings from Alemán Bañon et al. (2017). These authors examined whether gender assignment accuracy predicts the magnitude of P600 responses to gender agreement violations in L1-English L2 learners of Spanish. The P600s in their analyses were computed only for trials on which participants had correctly assigned the appropriate gender in the gender assignment task. A regression analysis on these data found no effect of gender assignment performance on P600 magnitudes. In other words, gender assignment performance did not account for any variance in P600 magnitudes after item-specific knowledge was accounted for by restricting trials to those for which the gender of a noun was known.

In addition to knowledge of a noun's grammatical gender, other factors have also been found to modulate the processing of grammatical gender agreement in an L2. Gabriele et al. (2013) and Morgan-Short, Sanz, Steinhauer, and Ullman (2010) both observed an absence of P600 effects to gender agreement violations at lower proficiency levels, with P600 effects emerging as proficiency increased. There is also evidence that cross-language similarity influences L2 gender processing. Some research indicates that L2 learners are more likely to show P600 responses to gender agreement violations if the agreement rules are similar in their

L1 to their L2 (Foucart & Frenck-Mestre, 2011) or if their L1 is typologically closer to their L2 (Sabourin & Stowe, 2008). Finally, sensitivity to gender agreement violations has been shown to diminish as the distance between agreeing elements increases (Foote, 2011; Gabriele et al., 2013; Keating, 2009), similar to findings in native speakers (Gabriele et al., 2013; O'Rourke & van Petten, 2011).

Though late L2 learners can show native-like patterns of processing with grammatical gender, they have also been shown to differ from native speakers in some ways. Production data find that late L2 learners make persistent gender assignment errors in speech, even at high levels of proficiency (e.g. Bruhn de Garavito & White, 2002; Dewaele & Véronique, 2001; Grüter et al., 2012; White et al., 2004). In addition, whereas studies more consistently report P600 responses to determiner–noun violations (but see Meulman, Stowe, Sprenger, Bresser, & Schmid, 2014; Sabourin & Stowe, 2008), some fail to find any P600 response to adjective–noun agreement violations in late L2 learners (Foucart & Frenck-Mestre, 2011, 2012; Morgan-Short et al., 2010), sometimes observing N400-like negativities instead. It is unclear, however, to what extent L2 proficiency might play a role in these latter findings, as a number of studies have reported N400 effects to morphosyntactic violations in the grand-averaged ERPs for lower proficiency learners, sometimes also finding an absence of P600 effects (McLaughlin et al., 2010; Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, & Molinaro, 2006; Tanner, McLaughlin, Herschensohn, & Osterhout, 2013).

Some research further indicates that late L2 learners are unable to or inefficient at using grammatical gender marking as a cue to anticipate and facilitate the processing of upcoming nouns. Guillelmon & Grosjean (2001) found that, unlike native French speakers, late L1-English L2-French bilinguals were no faster at repeating nouns in an auditory shadowing task when they followed an informative gender-marked determiner compared to when they followed an uninformative gender-neutral determiner. Similarly, looking-while-listening studies have found that late L2 learners of Spanish do not orient faster to target objects in a visual display when gender marking is informative compared to when it is uninformative about which noun will be named (Grüter et al., 2012; Lew-Williams & Fernald, 2010). In a set of more recent eye tracking studies, Hopp (2013, 2016), however, reports evidence that late L2 learners' ability to use gender-marked determiners predictively depends on a learner's speed of lexical access and the accuracy with which they correctly assign grammatical gender to nouns. Participants with high lexical access speed and greater accuracy in gender assignment showed evidence of anticipatory eye movements when trials were sorted based on the gender that participants assigned to the experimental stimuli. These results accord well with the findings of Lemhöfer et al. (2014) in suggesting that at least part of the difficulty L2 learners experience with grammatical gender can be traced to the stability of their gender representations for individual nouns. Dussias et al. (2013), moreover, report evidence that the predictive use of pre-nominal gender marking is mediated by proficiency and is facilitated by the presence of gender marking in the L1. Finally, Hopp and Lemmerth (2016) report evidence that the predictive use of grammatical gender in an L2 may depend on L1–L2 congruency in how grammatical gender is marked morphosyntactically, and in whether a noun has the same or different grammatical gender across languages.

To summarize, the research on L2 acquisition and processing of grammatical gender has shown that late L2 learners can acquire appropriate grammatical gender representations, and that they can use this information online to compute gender agreement and anticipate upcoming nouns. Nevertheless, L2 learners make persistent errors in online gender assignment, show variable sensitivity to gender agreement, and often exhibit a reduced or absent ability to use gender marking as an anticipatory cue.

In attempting to account for the difficulty that late L2 learners experience in learning and using grammatical gender, two types of theoretical accounts of L2 inflection have been appealed to: accounts positing a representational deficit (e.g. Franceschina, 2005; Hawkins & Franceschina, 2004; Hawkins & Chan, 1997; Tsimpli & Dimitrakopoulou, 2007) and accounts positing online retrieval or mapping difficulties (Lardiere, 1998; Prévost & White, 2000). The representational deficit accounts claim that abstract grammatical gender features cannot be acquired in a late-learned L2 unless the L1 instantiates grammatical gender. As mentioned earlier, however, the abundance of research showing online sensitivity to grammatical gender agreement in late L2 learners whose L1 does not possess grammatical gender indicates that these learners can, in fact, represent grammatical gender features in their L2.[1]

Accounts that posit online retrieval or mapping difficulties assume that grammatical gender features can be learned and represented, but that difficulty ensues when an abstract morphosyntactic feature must be mapped on to an overt morphophonological form under time pressure, particularly for speech (Lardiere, 1998; Prévost & White, 2000). Such accounts, however, have faced challenges. Grüter et al. (2012) found that L2 production errors with

[1] It is worth noting that representational deficit accounts can explain high accuracy in offline tasks and online sensitivity to gender agreement by assuming that L2ers learn grammatical gender using probabilistic cues (e.g. Hawkins & Franceschina, 2004), and that L2ers rely on these types of cues to assign and compute grammatical gender. There is, indeed, evidence that probabilistic cues impact online gender processing in L2 users (Alarcón, 2011; Bordag, Opitz, & Pechmann, 2006; Bordag & Pechmann, 2007; Taraban & Kempe, 1999). As discussed in section 1.1.1, however, native speakers also show sensitivity to probabilistic cues in gender assignment tasks, in addition to which more recent work has found that native speakers show rapid online sensitivity to probabilistic cues to gender when processing gender agreement (Caffarra & Barber, 2015; Caffarra, Janssen, & Barber, 2014; Caffarra, Siyanova-Chanturia, Pesciarelli, Vespignani, & Cacciari, 2015). Thus, though L2 users appear to rely more on probabilistic cues than native speakers, this cannot be taken to imply that L2ers represent and access grammatical gender in a fundamentally different way than native speakers.

grammatical gender are mostly assignment errors, not agreement errors. This suggests that L2 difficulty with grammatical gender reflects deficits at the lexical level rather than deficits in morphosyntax. Alemán Bañon et al. (2017), moreover, found no evidence for the use of a default gender in production, contrary to the predictions of computational accounts (Prévost & White, 2000; White, 2011; White et al., 2004). Both of these studies further found that late L2 learners do not show an advantage for comprehension over production, in contrast to a key prediction of the computational accounts under discussion.

A more recent account of L2 gender learning and processing, the Lexical Gender Learning Hypothesis (LGLH: Grüter et al., 2012; Hopp, 2013, 2016) makes specific claims about the origins of L2 difficulty with grammatical gender and about the use of gender marking as an anticipatory cue. In order to help situate the LGLH in the broader context of the literature, predictive language processing will first be discussed in greater detail. After this, we will return to the LGLH to describe its core claims, and the evidence in favor of this hypothesis.

1.2. Predictive language processing

There is strong evidence that human language comprehension employs predictive mechanisms (for reviews, see Dell & Chang, 2014; Huettig, 2015; Kaan, 2014; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2013; Van Petten & Luka, 2012). Evidence indicates that these mechanisms anticipate across multiple levels of granularity, ranging from fine-grained information about wordform to course-grained discourse-level information (Brothers, Swaab, & Traxler, 2015; DeLong, Urbach, & Kutas, 2005; Dikker, Rabagliati, Farmer, & Pylkkanen, 2010; Dikker, Rabagliati, & Pylkkänen, 2009; Farmer, Christiansen, & Monaghan, 2006; Federmeier & Kutas, 1999; Federmeier, Kutas, & Schul, 2010; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Fine, Jaeger, Farmer, & Qian, 2013; Kutas & Hillyard, 1984; Laszlo & Federmeier, 2009; Levy, 2008; Lew-Williams & Fernald, 2007; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2004, 2003; Wlotko & Federmeier, 2012). Native speaking adults, moreover, have been shown to generate expectations during comprehension on the basis of a variety of linguistic cues. These include verb semantics (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003), case (Kamide, Scheepers, & Altmann, 2003), verb tense (Altmann & Kamide, 2007), number (Lukyanenko & Fisher, 2016; but see Riordan, Dye, & Jones, 2015), phonological information (Shantz & Tanner, 2017), and grammatical gender (Dahan et al., 2000; Lew-Williams & Fernald, 2007), as well as message-level information such as sentential constraint (Federmeier & Kutas, 1999; Federmeier et al., 2007), discourse context (Brothers et al., 2015; Otten & van Berkum, 2008; van Berkum et al., 2005), disfluencies (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Bosker, Quené, Sanders, & de Jong, 2014) and prosody (Weber, Grice, & Crocker, 2006).

1.2.1. Evidence for graded prediction

While early accounts of prediction viewed it as an all-or-nothing process that incurred costs when an incorrect linguistic item was predicted, more recent accounts postulate a graded process that allows for multiple possible linguistic items to be simultaneously pre-activated without necessarily incurring costs (for discussions, see DeLong, Troyer, & Kutas, 2014; Kuperberg & Jaeger, 2016; Kutas, DeLong, & Smith, 2011). This shift is owed in part to a steadily growing body of evidence showing graded effects of a word's probability in context on how that word is processed. For example, early reading measures such as skipping rate, first fixation duration and gaze duration are reliably influenced by how predictable a word is in context (Ehrlich & Rayner, 1981; Kliegl, Nuthmann, & Engbert, 2006; Luke & Christianson, 2016; Rayner, Slattery, Drieghe, & Liversedge, 2011; Smith & Levy, 2013; see Staub, 2015 for a review). As words become more predictable in context, readers are less likely to fixate them, and, when fixated, early fixation durations are shorter.

In the ERP literature, the N400 has proven a useful index of anticipatory processing at the level of semantic features, offering additional evidence for graded effects of expectancy. The N400 is a negative-going ERP component related to semantic processing that peaks around 400 ms in healthy, young adults (see Kutas & Federmeier, 2011 for a review). When an experimental manipulation reduces the amplitude of an N400, this is generally taken to reflect facilitated semantic processing of some sort, though the precise functional significance of the N400 is still under debate (for a discussion, see Kutas & Federmeier, 2011). Consistent with the eye tracking literature, N400 amplitudes for young adults show an inverse correlation with the expectancy of a word (DeLong et al., 2005; Federmeier et al., 2007; Kutas & Hillyard, 1984; Nieuwland et al., 2018; Wlotko & Federmeier, 2012b).

Another reason for the shift stems from the paucity of evidence for a clear processing cost when a prediction is disconfirmed. For instance, while N400 amplitudes reliably decrease the more expected a word is, the size of these expectancy effects does not appear to depend on the degree of sentence constraint (Federmeier et al., 2007; Kutas & Hillyard, 1984). This suggests that larger N400 amplitudes to unexpected words do not index any processing costs when specific lexical predictions are not borne out (though see Brothers et al., 2015, who find larger N400 amplitudes to unpredicted versus predicted words that are matched on cloze probability. Note, however, that these authors do not interpret their N400 effects as indexing a misprediction cost). Similarly, Luke & Christianson (2016) showed that readers do not spend more time reading words that have more expected alternatives relative to words that are the most expected words in a cloze task. A number of ERP studies manipulating expectancy have, however, observed a frontal positivity following the N400 that is elicited when highly constraining sentences are continued by a word with low cloze probability rather than the strongly expected continuation (DeLong, Quante, & Kutas, 2014; Federmeier et al., 2007; Kutas, 1993; Thornhill & Van Petten, 2012; see Van Petten & Luka, 2012 for a review). It has been hypothesized that this late frontal positivity (LFP) may reflect the costs of a disconfirmed lexical prediction. Consistent with this, Delong, Urbach, Groppe, & Kutas (2011) report that the amplitude of the LFP is sensitive to cloze probability, with lower cloze unexpected words eliciting greater positivities. Brothers et al. (2015) provide stronger evidence that the LFP is specifically tied to disconfirmed lexical predictions rather than predictability as measured by cloze probability. In their study, native English speakers read short discourse contexts that constrained toward two completions that were approximately equally likely. Participants were explicitly instructed to try to predict the final word, and were asked to indicate whether their predictions were correct. Trials were then sorted based on whether participants correctly predicted the final word. Despite both having cloze probabilities of approximately 50%, results showed that unpredicted words elicited a greater post-N400 positivity compared to correctly predicted words; this positivity was largest over frontal and left hemisphere sites. Finally, Rommers, Dickson, Norton, Wlotko, & Federmeier (2016) recently showed further evidence for a possible cost of failed prediction in a reanalysis of the electroencephalogram (EEG) data from Federmeier et al. (2007). They used a time-frequency analysis to examine changes in EEG activity over time as a function of sentential constraint and a word’s expectancy. This type of analysis allows researchers to quantify how much activity is present in the EEG at different frequencies (measured in Hz) and to ascertain how activity in different frequency bands changes over time in response to experimental manipulations or different types of stimuli. Rommers et al. (2016) found that unexpected words elicited increased theta band activity (4-7 Hz) in highly constraining sentences compared to expected words. In contrast, weakly constraining sentences elicited no difference in theta band activity.

At present, it is unclear whether the EEG effects described above truly reflect a “cost” to mispredicting. DeLong, Troyer, et al. (2014) suggest a number of processes that the LFP may reflect, such as conflict monitoring, attentional switching or the updating of a learning mechanism. Moreover, Kutas et al. (2011) call into question whether “cost” is even an appropriate term for processes associated with mispredicting. In particular, they highlight the fact that short term “costs” reflecting error-based learning would confer long-term benefits by making a comprehender’s language system more accurate at generating future expectations; this idea is elaborated on by Kuperberg and Jaeger (2016), who argue for a probabilistic model of prediction that serves to reduce Bayesian surprisal during comprehension. Regardless of whether the described effects reflect costs, their restriction to highly constraining contexts in which specific lexical predictions can be generated make these data difficult to fully reconcile with an all-or-nothing manner of predicting, which should predict “costs” to arise regardless of constraint. Moreover, the sensitivity of the LFP to expectancy makes it perfectly consistent with graded accounts of prediction.

1.2.2. Prediction versus facilitated integration

To this point, the evidence for prediction that has been discussed, though suggestive, is largely ambiguous between an expectation-based account, in which information about upcoming material is predictively pre-activated, and between integrative accounts in which predictable words are more easily integrated into a syntactic frame, discourse model or situation model, but only after the bottom-up activation of a word's features has occurred (see Federmeier, 2007; Kutas et al., 2011 for discussions). The reason for this ambiguity is that the described effects of expectancy on behavioral and neurophysiological measures are observed only after a word has been encountered that confirms or disconfirms a prediction (but for arguments in favor of expectation-based accounts over integrative accounts on the basis of some of these data, see Federmeier, 2007; Kutas et al., 2011; Staub, 2015). Both neurophysiological and behavioral research has, however, provided compelling evidence for predictive language processing that shows effects prior to a predicted word, and which is thus not readily accounted for by facilitated integration.

The earliest such evidence came from eye tracking studies employing the visual world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). When used to study comprehension, the visual world paradigm (VWP) involves participants viewing scenes, objects or words either on a computer screen or in a real-world display while they listen to words or sentences. Eye movements are tracked with millisecond precision, providing a continuous measure of what individuals are visually attending to in real time (for a review of the VWP, see Huettig, Rommers, & Meyer, 2011). In a seminal study using the VWP, Altmann & Kamide (1999) had native English speakers view simple scenes depicting a single person and a set of objects. For example, one scene depicted a boy in a room with a toy car, a ball, a toy train and a cake. While viewing the scenes, participants heard sentences like (1), in which the verb semantics plausibly select only one item from the scene as the object of the sentence (i.e. the cake instead of a toy train, ball or toy car), or like (2), in which all items in a scene are appropriate objects of the sentence.

(1) The boy will eat the cake.

(2) The boy will move the cake.

Their results found that in sentences like (1), participants began to look at the target objects (e.g. cake) before the acoustic onset of the target words, whereas looks to target words for sentences like (2) did not begin until after the onset of the target words. This was taken as evidence that listeners can use verb semantics to anticipate upcoming verb arguments.

ERP studies have also yielded compelling evidence for anticipation prior to information that unambiguously confirms or disconfirms a lexical prediction. Wicha et al. (2004) showed that when reading constraining sentences in which a specific noun has a high cloze probability, native Spanish speakers predict a noun's grammatical gender along with its semantic features. ERPs at the pre-nominal determiners whose grammatical gender was incongruent with the gender of a predicted noun elicited a late positivity relative to determiners that matched the expected noun in grammatical gender. Subsequent research has replicated the main finding of this study (Foucart, Martin, Moreno, & Costa, 2014; Foucart, Ruiz-Tada, & Costa, 2015), though these along with two earlier studies by Wicha and colleagues (Wicha, Bates, et al., 2003; Wicha, Moreno, et al., 2003) showed enhanced negativities to unexpected pre-nominal articles, whereas Wicha et al. (2004) found only an enhanced positivity (see also Van Berkum et al., 2005, who found a positivity to pre-nominal gender-marked adjectives whose grammatical gender is incongruent with an expected noun).

Using a similar paradigm in English, DeLong et al. (2005) capitalized on the phonological alternation for English indefinite articles (*a* and *an*) and had native English speakers read constraining sentences in which either vowel-initial or consonant-initial nouns were highly expected. Results showed that pre-nominal determiners whose phonological form was inconsistent with the expected noun (e.g. *an* when *kite* is highly expected) elicited an N400-like effect compared to determiners that were consistent with highly expected nouns (e.g. *a* when *kite* is highly expected; but see Ito, Martin, & Nieuwland, 2016a and Nieuwland et al., 2018 for recent failures to replicate; see also DeLong, Urbach, & Kutas, 2017a,b for comments and criticisms). This paradigm was also recently used to show that native speakers of Polish can use context to anticipate the animacy of an upcoming noun without necessarily committing to an expectation for a specific noun (Szewczyk & Schriefers, 2013). Because these ERP studies all observe effects on determiners or adjectives that occur prior to an expected noun, and prior to or in the absence of any overt agreement violations, they provide strong evidence for anticipatory processes that cannot readily be explained by an integration account. These findings further indicate that the anticipation of nouns in constraining contexts is not limited to the anticipation of their semantic properties, but includes the grammatical and phonological properties of these nouns as well.

Finally, a handful of studies have reported differences in evoked neural activity to more versus less constraining contexts that precede an expected word. Using time-frequency analysis, Rommers et al. (2016) found that strongly constraining sentence contexts elicited less activity in the alpha range (8-12 Hz) than weakly constraining sentences prior to a critical word (see also Piai, Roelofs, & Maris, 2014; Piai, Roelofs, Rommers, & Maris, 2015 for similar findings). Chou, Huang, Lee, & Lee (2014) and Qian & Garnsey (2016) reported sustained frontal negativities elicited by classifiers that engender more uncertainty about an upcoming noun relative to classifiers that reduce uncertainty to a greater degree. In other words, the less informative, or constraining, a cue is about an upcoming noun, the more negativity it elicits in contexts where a noun is expected to occur. Fruchter, Linzen, Westerlund, & Marantz (2015) recorded native English speakers' neural activity using magnetoencephalography (MEG) while participants made lexical decisions on the second item in pairs of sequentially presented letter strings. Critical stimuli were adjective–noun pairs that differed in the extent to which the adjective was predictive of a specific noun (e.g. compare *stainless steel* to *beautiful scenery*). Predictability was measured by the transitional probability between an adjective and a noun. Their primary analysis focused on neural activity in the left middle temporal gyrus (MTG), which has been tied to lexical access (e.g. Indefrey & Levelt, 2004). The time window for this analysis began after the lexical access window for the adjective and ended before the onset of the critical noun. When adjective frequency was controlled for, the results in this time window showed that activity in the left MTG was modulated by the frequency of the expected noun, but only for adjectives that were highly predictive of a specific noun.

To summarize, there is a wealth of evidence to strongly suggest that human language comprehension employs predictive mechanisms. While some of this evidence is ambiguous between prediction versus integration, a number of experiments have shown compelling evidence for prediction prior to an expected word. Importantly, those ERP studies reviewed in this section that have shown effects of expectancy prior to an anticipated noun have also shown the same effects of expectancy at the noun as the ERP studies reviewed in section 1.2.1 This adds weight to the argument that the N400 and LFP effects observed to expected versus unexpected nouns in those studies reflect predictive processes. Given the strength of the evidence for predictive processing, some researchers have suggested that we stop asking *whether* humans can use top-down information to generate expectations during comprehension, and instead turn to examining *when* this does and does not occur (see Huettig, 2015; Kuperberg & Jaeger, 2016).

1.2.3. Factors modulating prediction

An emerging body of evidence has shown that the extent to which individuals anticipate during comprehension is modulated by a host of experience- and cognition-related factors. In native-speaking populations, children and older adults with low verbal fluency or vocabulary size show reduced or absent evidence for prediction (Federmeier et al., 2010; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Mani & Huettig, 2012). Similarly, children and young adults with lower reading skills exhibit less robust (Mani & Huettig, 2014) or delayed prediction effects (Huettig & Brouwer, 2015), in addition to which low-literate adults do not show evidence for prediction (Mishra, Singh, Pandey, & Huettig, 2012). Age is also an important factor in linguistic prediction, with older adults being less likely to anticipate compared to younger adults (DeLong, Groppe, Urbach, & Kutas, 2012; Federmeier & Kutas, 2005; Federmeier et al., 2010, 2002; Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006). There is, moreover, some evidence that quicker processing speed (Hopp, 2013; Huettig & Janse, 2016) and greater working memory (Huettig & Janse, 2016) lead to greater anticipatory effects. Research also indicates that predictive effects may be delayed or reduced by increased cognitive load (Ito, Corley, & Pickering, 2018) and in bilinguals relative to monolinguals (Dijkgraaf, Hartsuiker, & Duyck, 2016).

Recent work has further shown that a native speaker’s ability or propensity to predict depends on factors that are external to an individual. Sentence reading studies have shown that faster presentation rates lead to reduced or absent prediction effects (Ito, Corley, Pickering, Martin, & Nieuwland, 2016; Wlotko & Federmeier, 2015). Indexical properties of a speaker have also been shown to modulate prediction; Romero-Rivas, Martin, & Costa (2016) found that native Spanish speakers do not show predictive effects in highly constraining sentences spoken by an individual with a foreign accent, whereas predictive effects are found when the same sentences are spoken by a native speaker (see also Bosker et al., 2014). Shantz & Tanner (2017) examined how the reliability of cue-outcome mapping shapes the use of anticipatory cues during the comprehension of low-constraint sentences. They showed that native English speakers reliably use the English *a*/*an* alternation as a cue with which to prepare an upcoming noun-contingent response. Predictive response preparation, however, was only found for participants with a consistent cue-outcome mapping (see Hopp, 2016 for similar findings).

There is also evidence that the use of grammatical gender as a predictive cue may be affected by distributional and dynamic properties of a language itself. In Dutch, Loerts, Wieling, & Schmid (2013) found evidence for the use of the definite determiner *de* (common gender) as a predictive cue to an upcoming noun, but not for *het* (neuter definite determiner). They suggest that this may reflect the fact that *de* only occurs with nouns that have common gender, whereas *het* can occur with neuter nouns as well as common gender nouns when used in the diminutive. Thus *het* does not necessarily provide a reliable cue about the identity of an impending noun. Lundquist, Rodina, Sekerina, & Westergaard (2016) examined the predictive use of grammatical gender marking in speakers of two Norwegian dialects, Tromsø and Sortland. Norwegian historically has a three gender system, however the feminine gender is currently being lost and merging with masculine gender. Speakers of the Tromsø dialect still maintain a distinction between feminine and masculine genders, whereas the Sortland dialect has a relatively stable two-gender system consisting of the neuter and common genders. Using the visual world paradigm, Lundquist et al. (2016) found that the use of gender-marked determiners as anticipatory cues depends not just on which dialect an individual speaks, but also on individual differences in whether a speaker of the Tromsø dialect retains the feminine form in spoken production.

Whereas native speakers show predictive effects in a variety of contexts, second language learners often show a reduced or absent ability to predict in these same contexts (Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Grüter, Lew-Williams, & Fernald, 2012; Hopp, 2013, 2016; Ito, Martin, & Nieuwland, 2016b; Lew-Williams & Fernald, 2010; Martin et al., 2013; Van Bergen & Flecken, 2017; see Kaan, 2014 for a review). This has led to the broad hypothesis that second language users have a reduced ability to generate expectations (RAGE; Grüter, Rohde, & Schafer, 2016). Given the relatively nascent status of research on L2 predictive processing, it is not yet clear why L2 users seem to rely less on predictive mechanisms than native speakers, nor is it clear what the conditions are under which RAGE is and is not observed. The existing research has, however, identified various factors that appear to play a role in L2 predictive processing.

Dussias et al. (2013), for example, found that L2 proficiency mediates the predictive use of grammatical gender; high-proficiency L1-English L2 learners of Spanish showed anticipatory eye movements in a visual word task, whereas lower proficiency L1-English learners did not (see, also, Hopp & Lemmerth, 2016). Their study also found evidence that the ability to use grammatical gender as an anticipatory cue in an L2 is enhanced by having an L1 with grammatical gender. Despite being approximately matched in proficiency with the lower proficiency L1-English learners, participants whose L1 was Italian showed anticipatory eye movements, though only for the feminine nouns. Van Bergen & Flecken (2017) report further evidence that L1-L2 similarity plays a role in L2 predictive processing. They examined whether L1 and L2 speakers can generate linguistic expectations using placement verb semantics in Dutch, which distinguishes between placing an object in a standing position (*zetten*) and placing an object in a lying position (*leggen*). The researchers compared eye movements for native Dutch speakers, and L2 speakers of Dutch whose L1s were either German, English or French. L1-German participants, whose L1 makes the same semantic distinction on placement verbs as Dutch, showed anticipatory eye movements that were not statistically distinguishable from the native Dutch speakers. In contrast, the L1-English and L1-French participants, whose L1s do not make a standing versus lying distinction on placement verbs, showed no evidence of anticipatory eye movements.

Like native speakers, processing speed has been shown to influence prediction in L2 speakers. Hopp (2013) found that both native and non-native speakers of German showed greater predictive effects when they also had faster speeds of lexical access. Relatedly, Ito, Martin, et al. (2016a) found expectation-based N400 effects in L2 speakers of English when using a stimulus-onset asynchrony (SOA) of 700 ms, but not at a shorter SOA of 500 ms (but see DeLong, Urbach, & Kutas, 2017 for criticisms). Native speakers, in contrast, showed N400 effects at both SOAs. The authors suggest that this finding may reflect slower processing in an L2 compared to an L1. Finally, Hopp (2013) also reported that L2 participants who made fewer gender assignment errors in a production task showed larger gender-based predictive effects than participants who made more mistakes (also see, Hopp, 2016). Importantly, trials were analyzed based on the subjective gender assignments of the participants rather than objective gender, so it is unlikely that this effect entirely reflects proficiency or accuracy, per se. That said, it is unclear to what extent this effect of gender assignment accuracy is independent of proficiency, as the participants who were more accurate also scored higher on a separate proficiency test, in addition to which they had greater lengths of exposure and residency.

In light of the research discussed in this section, it is important to consider what it actually means for second language speakers to have a reduced ability to generate expectations. When L2 performance is compared to age- and education-matched peers, the data are certainly consistent with this hypothesis. However, the research reviewed here has also shown that the extent to which native speakers predict, if at all, is at least partly dependent on factors that are directly relevant to performance in an L2. For example, L2 speakers are likely to have lower reading skills in their L2 compared to native speakers (e.g. Favreau & Segalowitz, 1983). If reading skill influences L2 predictive processing the same way it influences L1 predictive processing, this may lead to less robust predictions that are comparable to those of L1 speakers matched in reading skills. Similarly, L2 speakers are also likely to have lower production skills compared to native speakers: they may possess smaller vocabularies, have lower verbal fluency, and exhibit slower speeds of lexical access (see Bialystok, Craik, Green, & Gollan, 2009 for a review). Language production has been implicated as a core component in predictive processing (Dell & Chang, 2014; Federmeier, 2007; Macdonald, 2013; Pickering & Garrod, 2013). Thus, to the extent that prediction in language comprehension relies on production, RAGE in L2 populations may partially reflect deficits in productive skills similar to the effects of vocabulary size and verbal fluency observed in L1 populations (Federmeier et al., 2010, 2002; Mani & Huettig, 2012). Finally, there is strong reason to believe that comprehending in a second language is more cognitively demanding than comprehending in one’s L1. L2 processing is likely to be less automatic than L1 processing (see Segalowitz & Hulstijn, 2005), in addition to which bilingual word recognition requires the suppression of both L1 and L2 competitors (e.g.

Cutler, Weber, & Otake, 2006; Dijkstra, Grainger, & Van Heuven, 1999; Dijkstra & Van Heuven, 2002; Weber & Cutler, 2004). In consideration of recent findings that increased cognitive load reduces prediction (Ito et al., 2018), a possible consequence is that if L2 users must devote greater cognitive resources to comprehending in their L2, they may not have enough resources left to dedicate to generating robust expectations.

It is further possible that the observed RAGE in L2 populations may partly reflect strategies that are adopted by individuals comprehending in their L2. Kuperberg and Jaeger (2016) suggest that the degree to which linguistic information is predictively pre-activated, if at all, may depend on the utility of predicting at a given time. If, for example, increased cognitive load makes it difficult to rapidly generate expectations, the cost of predicting may outweigh the benefits, in which case predicting may not be sufficiently useful. Similarly, if an L2 speaker is not certain of a noun's grammatical gender, then pre-nominal gender marking may have limited utility as a predictive cue. In other words, though Dye et al. (2017) show that by reducing the entropy, or surprisal, of an upcoming noun, pre-nominal grammatical gender marking should always be a useful cue, uncertainty about a noun's grammatical gender may reduce the predictive value of gender marking. Consistent with this, Hopp (2016) showed that native German speakers stop using grammatical gender as a predictive cue in contexts where the mapping between a noun and grammatical gender was variable (much like in L2 production), thereby reducing its predictive value. This latter finding suggests that the reduced or absent use of grammatical gender as an anticipatory cue observed in late L2 populations may reflect a mapping problem between nouns and their grammatical gender representations, rather than necessarily reflecting a reduced ability to predict more generally. In any case, given the possibility that gender marking can make nouns more predictable, facilitating lexical access and perhaps integration processes, L2 users clearly stand to profit from learning to use grammatical gender information predictively. An understanding of their reduced use of gender marking as an anticipatory cue therefore has the potential not only to inform L2 theory, but may also have pedagogical implications for L2 instruction. The reason for this difficulty, however, remains unclear.

1.3. The Lexical Gender Learning Hypothesis

The Lexical Gender Learning Hypothesis (LGLH; Grüter et al., 2012; Hopp, 2013, 2016) has been proposed to account for the observed deficits in gender-based anticipation seen in late L2 populations. It makes three distinct claims about the underlying cause of L2 speakers' reduced ability to use grammatical gender information predictively:

1) Differences in the conditions under which children and adults learn a language lead to the formation of weaker links between nouns and their corresponding gender representations for adults learning a second language (cf. Gollan et al., 2008). More specifically, Grüter et al. (2012, pp. 209-210) suggest that the tendency of children learning a gendered L1 to mis-parse determiner + noun sequences as a single chunk (cf. Carroll, 1989) ultimately leads to a strong association between nouns and their respective gender-marked determiners, and consequently a strong association between nouns and their appropriate gender nodes when these chunks are reanalyzed. Grüter et al. further argue that L2 learners are more likely to rely on multiple cues to word learning as a result of having knowledge from their L1 and of being more cognitively mature, and are therefore less likely to develop a strong association between nouns and their abstract gender nodes due to lower overall reliance on distributional cues (i.e. determiner + noun contingencies).

2) Weaker links between nouns and their gender representations result in greater variability in gender assignment.

3) These unstable gender representations lead to a reduction in or absence of the use of gender marking as an anticipatory cue to upcoming nouns.

There is indirect evidence to support various aspects of these claims. Montrul et al. (2014), for example, had proficiency-matched heritage and late L2 learners of Spanish perform a shadowing task similar to those used by Bates et al. (1996) and Guillelmon and Grosjean (2001). Participants had to repeat as quickly as possible the nouns in three-word sequences (determiner + adjective + noun) in which the gender of a noun either agreed or disagreed with the gender-marking on the determiner and adjective. Native speakers and heritage speakers both showed a slow-down in naming latencies when pre-nominal gender marking did not agree with the noun that was to be repeated. Late L2 learners, in contrast, showed no differences in naming latencies, similar to the findings of Guillelmon and Grosjean (2001). Because the initial exposure to Spanish of heritage speakers took place in early childhood, these findings are consistent with the possibility that differences in learning conditions may impact the predictive use of grammatical gender.

Arnon and Ramscar (2012) report computational and experimental evidence consistent with the first claim of the LGLH. They showed that learning noun labels in isolation (much as in vocabulary lists encountered by adults in the classroom) before encountering these noun labels in sentential contexts with their appropriate determiners (more like in naturalistic settings as in L1 acquisition) will result in poorer learning of nouns' genders compared to when the order is reversed. Siegelman and Arnon (2015) expanded on this finding to further show that initial exposure to unsegmented input (auditory phrases with no pauses between words) makes learners more likely to treat a noun and its determiner as a unit compared to when initial exposure is to segmented input with pauses between words. Together, these latter results indicate that input which does not initially separate a noun from its determiner, or which maximises the likelihood of a noun and a determiner being initially processed as a chunk will lead to stronger links between the two (see, also, Arnon & Christiansen, 2017, who explicitly argue that chunking behavior is a key factor in explaining L1-L2 differences). As previously discussed, Hopp (2013, 2016) has shown that gender assignment accuracy modulates the use of gender marking as an anticipatory cue. These results are consistent with the third claim of the LGLH, though it should be noted that these studies only used single measures of gender assignment. Consequently the stability of these representations across tasks and modalities cannot be adequately assessed.

To date, no study has yet systematically manipulated learning condition in order to examine how learning context impacts both gender assignment and anticipatory processing. This dissertation will address this current gap in the literature by directly testing the claims of the LGLH.

1.4. Research questions

The research questions to be addressed in this dissertation are as follows.

1) Does learning context influence adult L2 learners' stability of grammatical gender representations and ability to use grammatical gender marking to anticipate upcoming nouns?
2) Does variability in gender assignment performance modulate the extent to which adult L2 learners anticipate upcoming nouns?

In addressing these questions, this dissertation will provide the first direct test of the LGLH. Because prior research has shown that having an L1 with grammatical gender may facilitate the acquisition and processing of grammatical gender in an L2, this dissertation further asks:

3) Does L1 influence adult L2 learners' stability of grammatical gender representations and ability to use grammatical gender marking to anticipate upcoming nouns?

Finally, Grüter et al. (2012) specifically posit that weaker links between nouns and their respective gender features should result in slower retrieval of gender information, and that gender assignment errors and less effective use of gender marking as a predictive cue follow as a consequence of this delayed retrieval (pp. 210). Shantz and Tanner (2016), however, found that for highly familiar nouns, late L2 learners of German show no delay in their retrieval of grammatical gender information relative to native speakers. Because nouns in their study were highly trained, gender representations were likely stable for all items. It is therefore possible that gender retrieval might nonetheless be delayed for nouns with less-stable gender representations. Thus, this dissertation also asks:

4) Does the speed of retrieval for gender information predict gender assignment performance?

1.5. Description of experiments

1.5.1. Experiments 1 and 2

Experiments 1 and 2 were designed to address research questions 1, 2 and 3. These experiments employ an artificial grammar learning task modelled after work by Arnon and colleagues (2012, 2015). Native speaker of English (Exp. 1) and native speakers of German (Exp. 2) learned an artificial grammar either under a condition that encourages chunking behavior, or under a condition in which chunking is less likely. This was followed up by a visual world eye tracking task and by a battery of gender assignment tasks. It has been argued that the tendency to chunk contiguous words, such as determiner + noun sequences, plays a major role in explaining L1-L2 difference both in general (Arnon & Christiansen, 2017) and with respect to grammatical gender in particular (Carroll, 1989; Grüter et al., 2012). To the extent that this is true, the artificial grammar learning task thus directly manipulates learning context in a way that parallels differences in L1 vs L2 acquisition. If such differences in learning are responsible for the pervasive L2 difficulties with grammatical gender that have been reported in the literature, the group learning under a more L1-like learning condition should exhibit better learning of grammatical gender. Moreover, by sorting trials in the visual world task based on how stable the gender assignments were for the nouns on each trial, these experiments will further examine the extent to which learning context and gender assignment variability impact the ability of L2 learners to use grammatical gender as an anticipatory cue, thus addressing research questions 1 and 2. Finally, research question 3 is addressed by including a group of participants whose L1 does not employ grammatical gender, the native English speakers in Exp. 1, and a group whose L1 does employ grammatical gender, the native German speakers in Exp. 2. If having grammatical gender in one's L1 truly does facilitate learning gender and learning to use gender as an anticipatory cue, the native German speakers in Exp. 2 should exhibit better gender learning and greater anticipatory effects when compared to the participants in Exp. 1.

1.5.2. Experiment 3

To date, the existing research examining the use of gender marking to anticipate upcoming nouns has been restricted to behavioral measures. This research has found that L2 populations have a reduced ability to use gender predictively. Anticipation can further be modulated by an individuals' accuracy in gender assignment, which, in line with the LGLH, suggests that the stability of gender representations may impact the predictive use of gender marking. However, behavioral measures have been shown – at least in some cases – to underestimate performance in an L2, while neural measures can be more sensitive to subtle processing differences (McLaughlin, Osterhout, & Kim, 2004; Tokowicz & MacWhinney, 2005). Moreover, the use of single measures of gender assignment in previous studies has made it impossible to reliably assess the role that gender assignment variability plays in the anticipatory use of gender marking during comprehension. To address these issues, Experiment 3 combines ERPs with a cued lateralized picture monitoring task using native and non-native speakers of German. Using multiple measures of gender assignment, this experiment will address research question 2 in a natural language context, and thus will be able to determine the extent to which any effects found in Experiments 1 and 2 generalize beyond an artificial language learning context. As in Experiments 1 and 2, trials will be sorted based on the stability of gender assignments for the nouns on each trial to assess the extent to which gender assignment variability modulates the anticipatory use of grammatical gender. In addition, the gender assignment data will be used to assess research question 4, which will be accomplished by examining whether gender assignment latencies are predicted by gender assignment variability.

CHAPTER 2: LEARNING CONTEXT AND GENDER-BASED ANTICIPATORY EYE-MOVEMENTS IN AN ARTIFICIAL GRAMMAR

The LGLH proposes that differences in the conditions under which adults and children learn a language influence if and to what extent speakers of a language use grammatical gender predictively. Though there is evidence consistent with this possibility (Grüter et al., 2012; Guillelmon & Grosjean, 2001; Hopp, 2013, 2016; Montrul et al., 2014), it is indirect and correlational; no study has yet directly manipulated learning condition in any systematic fashion between experimental groups to test this hypothesis. Arnon and colleagues (2012, 2015) showed that learning condition impacts gender assignment accuracy, but they did not test whether this also affects gender assignment variability or anticipatory processing. Grüter et al. (2012) showed anticipatory effects for gender-marked determiners paired with novel nouns in a training paradigm, but did not manipulate learning condition. Similarly, Montrul et al. (2014) found that heritage learners of Spanish, like native speakers, are slower at repeating nouns in a shadowing task when they are preceded by gender-incongruent determiners, whereas late L2 learners showed no such effect.

Moreover, the two purportedly causal mechanisms in the LGLH (learning condition and stability of gender representation) are in fact conceptually independent, such that they may have additive or interactive effects on the ability to use gender as an anticipatory cue. The design of Experiments 1 and 2 will be able to disentangle the relative contribution of these mechanisms in a way that has not yet been tested. Furthermore, prior research has suggested that the processing of grammatical gender in an L2 is influenced by whether or not grammatical gender is instantiated in the L1 (Dussias et al., 2013; Sabourin, Stowe, & de Haan, 2006). While there is evidence that this extends to anticipatory processing in an L2 with transparent gender marking

(Dussias et al., 2013), no study has examined whether the same is true for an L2 with opaque gender marking. These experiments will therefore combine training in an artificial language with the visual world paradigm to investigate whether learning condition influences the consistency with which adult L2 learners assign grammatical gender, the ability of adult L2 learners to use grammatical gender predictively, and whether either of these outcomes is influenced by grammatical gender in the L1.

2.1. Experiment 1

2.1.1. Participants

Participants in Experiment 1 were 64 native speakers of American English (43 Female; mean age = 20.27 years; range: 18-25), all of whom reported normal or corrected-to-normal vision and no known hearing impairments. Four participants reported minimal exposure to a language other than English prior to the age of 5, all of which have grammatical gender (Polish, German, Yiddish, Bulgarian and Italian). None of these participants considered themselves proficient in the languages other than English that they were exposed to in early childhood. These four participants were equally distributed across the two learning conditions. 57 participants reported having spent time learning a language with grammatical gender. The remaining 7 participants reported no experience learning a language other than English. There was no substantive imbalance in how these 7 participants were distributed across learning conditions (3 in one condition, 4 in the other).

2.1.2. Materials

The artificial grammar for this study is largely the same as that used by Arnon & Ramscar (2012) and Siegelman & Arnon (2015). It consists of 24 disyllabic noun labels for familiar concrete objects (see Appendix A), two articles (*sem* and *bol*) and a carrier phrase (*os ferpel en*). Nouns were divided equally into two gender classes so that each noun only occurred with one article; assignment to gender classes was counterbalanced across two experimental lists. Care was taken so that repeated syllable structures in the stimuli (e.g. the *‘ot*’ ending in *etkot*, and *fersot*) were present in both gender classes so that these could not be used as cues to a noun’s gender. A full sentence in this grammar always began with the carrier phrase, followed by an article and then a noun.

The auditory stimuli were created by having a female, native speaker of English produce every possible combination of the carrier phrase with an article and a noun. These were recorded in a sound-attenuating booth with a condenser microphone at 16-bit with a 44,100 Hz sampling rate. A single representative token was extracted from these recordings for the carrier phrase, for each noun, and for each determiner. This ensured that the acoustic properties of the carrier phrase would be identical for all sentences, and that the acoustic properties of the articles were identical for all nouns in a given gender class. To minimize co-articulatory cues, nouns were extracted following the determiner *sem* and determiners were extracted preceding the /h/-initial noun, *hekloo*. All of the extracted sound files were intensity-scaled to 64dB. 36.6 ms of mid-vowel glottal pulses were added to the token of *bol* so that it was approximately equal in duration to the token of *sem*. Two versions of each possible sentence were then created by concatenating these tokens. One version included a 250 ms period of silence between each token (carrier phrase

+ 250 ms + article + 250 ms + noun), and the other version included no pauses between tokens (carrier phrase + article + noun).

Stimuli also included 24 colored images depicting inanimate, high frequency, concrete objects. These were taken from Rossion and Pourtois' (2004) colorized version of the Snodgrass and Vanderwart (1980) dataset. Each depicted object was assigned a noun label from the artificial grammar. Across gender classes, nouns were matched for their age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), concreteness (Brysbaert, Warriner, & Kuperman, 2014), and frequency in English (Brysbaert & New, 2009). Because the same stimuli were used in Experiment 2 as well, which uses native German speakers, nouns were also matched across gender classes for their frequency in German (Brysbaert et al., 2011). Within a gender class, the nouns were further matched for their mean frequencies in English and German. Moreover, each gender class in the artificial grammar contains an equal number of items from each gender class in German (i.e. each gender class in the artificial grammar has 4 nouns with neuter gender in German, 4 with masculine gender, and 4 with feminine gender). This was done to prevent target items or their grammatical gender from being predictable on the basis of a noun's German gender. Finally, stimuli also included a colored image of a man pointing (see Arnon & Ramscar, 2012; Siegelman & Arnon, 2015).

2.1.3. Procedure

2.1.3.1. Training Task

An experiment session began with a training phase consisting of two blocks of 120 trials each. In the noun-label block, participants saw an object presented on a computer screen accompanied by an auditory presentation of the noun label in isolation (e.g. *etkot*). In the sequence block, participants saw an object presented visually on a computer screen along with the image of the man pointing at the object. This was accompanied by an auditory presentation of the carrier phrase plus an article and the noun label for the object (e.g. *os ferpel en sem etkot*). For training, the sentences containing no pauses were used so as to minimize cues to word boundaries (see Siegelman & Arnon, 2015). In both blocks, participants were asked to repeat the sounds they heard in order to enhance learning (cf. Hopman & MacDonald, 2018). Half of the participants began with the noun-label block followed by the sequence block (noun first condition), and the other half of participants began with the sequence block followed by the noun-label block (sequence first condition). Objects and their accompanying auditory stimuli were presented in a randomized order; objects were not repeated within a block until all 24 objects had been presented in an iteration of the training stimuli.

2.1.3.2. Visual world eye tracking task

Training was followed immediately by the visual world eye tracking task. Each trial consisted of two objects from the training stimuli displayed visually on the computer monitor: one to the left of a central fixation cross, and one to the right. Half of the trials were gender-mismatch trials in which the objects displayed belonged to different gender classes, making pre-nominal gender marking an informative cue to the identity of the upcoming noun. The other half of the trials were gender-match trials in which the objects displayed belonged to the same gender class (i.e. uninformative trials). Each object occurred equally often as a target item and as a distractor in informative and uninformative trials. Moreover, each object occurred equally often on the left and right side of the screen in each condition as a target and as a distractor.

In addition, for each condition, each target and distractor occurred equally often in a pair where the German genders of the items were matched or mismatched. To illustrate, a target object that had masculine gender in German might be paired with one noun with masculine gender in the informative condition, and one noun with feminine gender. In the uninformative condition, that same target noun would then be paired with another noun with masculine gender in German, and a noun with neuter gender. When that masculine noun served as a distractor, it would then be paired with one neuter noun in the informative condition, and one feminine noun in the uninformative condition, in addition to one masculine noun in each of these conditions. Target nouns were therefore not predictable on the basis of grammatical gender in German. In total, there were 96 trials (informative/uninformative x target/distractor x 24 items).

Each pairing of a target word with a distractor was, moreover, unique and not repeated within a list. Across lists these pairings were counterbalanced so that the item occurring as a target in one list would serve as the distractor in another list and vice versa. Presentation order and the left/right position of items within pairs were also counterbalanced across lists.

Each trial began with a central fixation cross which participants were required to click on to start the trial. As soon as the fixation cross was clicked, the target and distractor images appeared to the left and right of the fixation cross. Presentation of the auditory sentence began concurrently with the appearance of the images. Sentences were presented over headphones at a comfortable listening volume. Participants were instructed to listen to each sentence and to click on the image that was described in the sentence. No instructions were given about where to look during a trial. The sentences used for this task were those that included the 250 ms pauses between sound files to give participants more time to predict. The duration of the carrier phrase was 780.2 ms, and the duration of each article was 235.4 ms for *bol* and 236.1 ms for *sem* (difference = 0.7 ms). Target nouns thus onset at approximately 1516 ms. Trials ended 2 seconds after a participant clicked on an image.

Eye movements were recorded from each participant's right eye using a desk-mounted Eyelink 1000+ eye tracker (SR Research) with a 500 Hz sampling rate. A 9-point calibration was used at the beginning of the experiment; a drift check was performed at the beginning of each trial. Participants were seated at an approximately 100 cm viewing distance from the computer monitor.

2.1.3.3. Gender assignment tasks

After the visual world task, participants completed three gender assignment tasks to assess their learning, and to use to re-sort trials on the basis of gender assignments. Following Arnon and Ramscar (2012), these were a forced-choice picture matching task, which each participant completed twice, and a sentence production task. In the picture matching task, participants were shown an image while they heard two full sentences in the artificial grammar. On half of the trials, one of these sentences contained an incorrect determiner; on the other half of trials, one of the sentences contained an incorrect noun. Participants were asked to indicate which sentence best matched the picture they were shown. After making their judgment, participants were asked to rate their confidence in their judgment on a scale from 1-4.[2] This task was completed once immediately following the visual word task, and again with a different list after the production task.[3] In the production task, participants were shown an image and asked to name the depicted

[2] Note that confidence ratings are not available for 31 of the 64 native English speakers, as this component to the forced-choice task was added partway through data collection.

[3] Note that due to a programming error, on the second iteration of the forced-choice task some of the items in the article condition were repeated in place of some of the trials intended to test noun knowledge. For the data analyses,

object using a full sentence in the artificial grammar. These sentences were recorded on a hand-held recording device.

Note, with respect to the production data, that a large number of participants ended up using either no article in their productions, or using only one, despite above-chance performance in the forced-choice task. Consequently, using the production data would likely lead to an underestimation of gender knowledge for the participants who did not produce any articles, and for the participants who used only one article this would lead to an overestimation of their knowledge for the gender class of the article they produced, and an underestimation of their knowledge for the other gender class. It was therefore decided not to include the production data in any analyses.

2.1.4. Data preprocessing and analysis

In order to assess if and to what extent participants use pre-nominal gender marking as a predictive cue, the eye tracking data were analyzed by measuring the latency of the first gaze shift to target images on trials where the target image was not already being fixated at the onset of the determiner (e.g. Lew-Williams & Fernald, 2007) and the probability of fixating the target noun (e.g. Altmann & Kamide, 1999). First fixation probabilities were measured in a time window that extended from 200 ms after the onset of the article until 200 ms after the onset of the noun (total duration = 485 ms). This corresponds roughly to the period of time in which information from the determiner is the only source of linguistic information that could mediate

only the first repetition of an item testing article knowledge was used. There were thus still only 24 trials per participant testing article knowledge (one for each item), but there were fewer than 24 trials testing noun knowledge from the second forced-choice task.

eye movements to the target. The latency of the first gaze shift was measured in a time window starting at 200 ms after the acoustic onset of the determiner and extending until the end of the trial. This choice of time windows differs from prior studies (e.g. Grüter et al., 2012), which have typically used a constrained time window with a total duration less than one second. The larger time window for this experiment was chosen based on two major considerations: first, whereas previous studies have typically analyzed their data with ANOVAs, which can be strongly influenced by extreme data-points, it has become relatively straightforward to fit models with distributions more suited to skewed data, such as latency data, by using, for example, log-normal or exponentially modified Gaussian distributions (see, e.g. Baayen & Milin, 2010; Whelan, 2008). Second, because participants had no previous experience with the artificial grammar before training, it was unclear how well first fixation latencies from previous studies could be used to guide the selection of a time window, nor how the choice of time window might impact the results. It was therefore thought better to include all of the data to avoid a potentially problematic increase in researcher degrees of freedom, and to avoid the possibility that time window selection might influence the results. Finally, for trials on which participants correctly selected the target noun, response times for clicking on the target image were also analyzed.

Data from the gender assignment tasks were used to classify trials in the eye tracking task as stable or unstable. Trials were classified as stable if a participant had assigned the correct gender to the target noun and the distractor noun on both iterations of the forced-choice task. Otherwise trials were classified as unstable.

Data were analyzed using Bayesian mixed effects regression modeling with the *brms* package in R (Bürkner, 2017). The choice of Bayesian models was motivated by a number of practical considerations. First, following the recommendations of Barr, Levy, Scheepers, and

Tily (2013) and Schielzeth and Forstmeier (2009), I used the maximal random effects structure for all models. Given that there are strong theoretical motivations for each parameter in each model (see below for model structures), it was important to include a maximal random effects structure so as to avoid violating conditional independence (see Barr et al., 2013 for a discussion). In practice, mixed effects models with complex random effects structures often fail to converge when using standard approaches such as those implemented by *lme4* (Bates, Mächler, Bolker, & Walker, 2015), especially with binary dependent variables (Eager & Roy, 2017). This often leads researchers to adopt ad hoc model simplification techniques to achieve convergence, which implicitly assumes that the failure to converge was due to low or zero variance in the random effects structure (see Eager & Roy, 2017 for a discussion). This approach can be problematic, however, because it increases researcher degrees of freedom in deciding how to simplify the random effects structure, and because recent simulation work has shown that the failure of a model to converge cannot be taken as diagnostic of an unsupported random effects structure (Eager & Roy, 2017). In other words, failures to converge are often due to the estimation algorithm rather than due to the random effects structure not being supported by the data. Consequently, when researchers fit sub-maximal random effects structures due to convergence failures, there may nonetheless be important random variance in the data that is not being modelled, and which may therefore increase the chances of Type I errors (cf. Barr et al., 2013; Schielzeth & Forstmeier, 2009). In contrast to these standard models, Bayesian models have a much greater likelihood of converging on the maximal random effects structure (Eager & Roy, 2017; Kimball, Shantz, Eager, & Roy, in press; Sorensen & Vasishth, 2015).

A further motivation for adopting a Bayesian approach is that the nature of the analyses I conduct for these experiments makes it highly likely that there will be a large degree of imbalance across cells in the model structure. The simulation work done by Eager and Roy (2017) found that imbalance greatly decreases the likelihood of convergence for models fit with *lme4*. This was found to be especially problematic for logistic models with complex random effects structures, which failed to converge on 82% of the simulations. In contrast, the Bayesian models failed to converge only 3% of the time for complex linear models and less than 0.001% of the time for complex logistic models. Their simulations further found that Bayesian models performed much better than *lme4* in providing accurate parameter estimates for logistic models.

Because my data are likely to be highly imbalanced, this also means that there is a strong possibility that the data for my binary dependent variables will contain quasi-separation, which occurs when a binary dependent variable can be perfectly classified by one level of a predictor, but not by all levels of that predictor. For example, if participants never fixated the target on unstable trials, but did fixate the target on some proportion of stable trials, then the data would show quasi-separation. That is, for the predictor of stability, we could perfectly predict the probability of fixating the target on unstable trials (i.e. 0), but not on stable trials. This kind of quasi-separation is unlikely to occur in the fixed effects parameters given the large amount of data, but it is likely for quasi-separation to be present in the random effects structure for at least one if not multiple participants. When present, quasi-separation makes parameters difficult to estimate using maximum likelihood estimation, which is used by *lme4*, and increase the likelihood of non-convergence. Bayesian approaches, in contrast, can improve the likelihood of convergence and provide more reasonable parameter estimates by placing reasonable constraints on the models with the use of weakly informative priors (see Kimball et al., in press for more on quasi-separation and for a demonstration of how Bayesian models can be used to deal with quasi-separation in linguistic data).

Given that my models have somewhat complex random effects structure, contain imbalanced data, and very likely contain quasi-separation, I decided to use Bayesian models from the outset in order to maximize the likelihood of achieving convergence on the full parameter structures, to obtain more reasonable parameter estimates, and to avoid the issues introduced by using model simplification techniques in the face of non-convergence.

All models were fit using the default, weakly informative priors in *brms*. Briefly, the prior specification places constraints on the model by specifying where a researcher believes a true parameter value is most likely to lie, and how strongly that belief is held. For example, if a researcher has a very strong reason for believing that a parameter will have a specific effect size, the researcher can place a more informative prior on that parameter that makes that value more likely, and can further constrain the prior to assume the distribution of parameter estimates will most likely be tightly clustered around that value. Without an extensive literature on which to base informative priors, I opted for weakly informative priors to avoid biasing the analyses in favor of any particular hypotheses. By using weakly informative priors, I assume that the most likely value for each parameter is zero, and that small effects are more likely than large effects. Because these priors are weak, however, they do not prevent the model from estimating non-zero or large effects if the data support such estimates. For more on weakly informative priors, see Gelman (2006), Gelman, Jakulin, Pittau, and Su (2008).

For the model fit to the gender assignment data, the dependent variable was gender assignment stability, which encoded whether or not a participant had assigned the correct grammatical gender to a noun on both iterations of the forced-choice task (1 = stable, 0 = unstable). These data were fit to a logistic model with a Bernoulli distribution. Thus, higher model coefficients reflect an increased likelihood of having a stable gender assignment. Fixed effects parameters were learning group (noun first or sequence first) and the intercept. Random effects included a random intercept for participant and target noun as well as a by item slope for learning group.

Fixation probability data were also fit to logistic mixed effects models using a Bernoulli distribution. Fixation latency data and the response time data were modeled using exponentially modified Gaussian distributions (cf. Baayen & Milin, 2010; Whelan, 2008). Fixed effects for all models fit to the eye tracking data included the intercept, the three-way interaction between learning group, stability (stable or unstable) and condition (informative or uninformative), and all subordinate interactions and main effects. Random effects included by participant random slopes for the interaction between stability and condition as well as the main effects of each, and by item random slopes for the interaction between learning group, stability and condition, as well as all subordinate interactions and main effects.

For all models, categorical predictors were sum coded. Thus, the intercept values estimated in each model represent the overall model estimated means. Each model was run with 4 chains. Each chain consisted of a total of 2000 iterations, 1000 of which were warm-up sampling. For the logistic models, priors for the fixed effects and standard deviations were set to a Student-*t* distribution with a center at 0, a scale of 10 and 3 degrees of freedom. The correlation parameters in all models had an *LKJ*(2) prior. For the models fit to the latency data, priors for the intercepts were set to Student-*t* distributions with 3 degrees of freedom, centers located at the median latency value for each dataset, and scales equal to the standard deviation of each dataset. Priors for the fixed effects parameters were set to a gamma distribution with a shape parameter of 1 and a scale parameter of 0.1. Finally, in the models for the latency data, the variance parameter had a prior with a Student-*t* distribution with 3 degrees of freedom, a center of 0, and a scale equal to the standard deviation of each dataset.

Convergence of the models was assessed by ensuring that there were no divergent transitions post-warmup, by examining the traceplots for good mixing, and by ensuring that the Gelman-Rubin *Rhat* statistic was below the recommended 1.1 (Gelman & Rubin, 1992). The initial model fit to the reaction time data for the native English speakers gave a warning about divergent transitions. Increasing the maximum treedepth above 10 and re-running the model eliminated the problem. No other models gave warnings about divergent transitions. All traceplots showed good mixing, and all *Rhat* values were below 1.1. Thus, all models were deemed to have converged.

The output of Bayesian mixed models is similar to that of frequentist models, however the interpretation differs in a number of important ways. Both types of models provide coefficient estimates, however whereas frequentist models provide a single coefficient estimate for each parameter, Bayesian models provide a distribution of parameter values that provide a rich source of information about each parameter. Generally, this distribution of values is used to calculate a measure of central tendency (typically the mean), and a 95% credible interval. The mean corresponds roughly to the parameter values estimated by frequentist models. The 95% credible interval is similar to a 95% confidence interval, however its interpretation is different in that 95% credible intervals provide the range of parameter estimates in which we can be 95% certain that the true parameter value lies, given the data. Finally, in contrast to frequentist statistics, Bayesian models do not provide p-values. Inferences can be drawn, instead, by examining the posterior distribution of parameter estimates in order to determine how likely a certain hypothesis is given the data. Generally, if a 95% credible interval does not contain zero, we can conclude strong evidence that the parameter value differs from zero (i.e. we can conclude that there is an effect of that parameter on the data). However, even if a 95% credible interval does contain zero, that does not necessarily mean that we should conclude the absence of an effect. Rather, we can use the posterior estimates to calculate the probability that a parameter estimate has an effect in the direction estimated by the model. For example, if a parameter has a positive value, we can calculate the proportion of posterior estimates that are greater than zero (i.e. positive). This yields the probability that a parameter estimate has the sign (positive or negative) estimated by the measure of central tendency, that is, P(sign). If this probability is large, it is reasonable to conclude that there is some evidence for an effect, even if the evidence is not strong. Similarly, when computing pairwise comparisons, if we have specific predictions about the direction of difference, we can use the posterior estimates to calculate the probability that a difference has the predicted sign. For example, the LGLH predicts that first fixation latencies should be earlier on informative than uninformative trials, and that this difference should be larger for stable compared to unstable trials. That is, if we compare the differences for informative and uninformative trials with the subtraction Informative – Uninformative, we would expect this difference to be more negative for stable trials than for unstable trials. Thus, when computing differences we can calculate the posterior probability that Informative Stable - Uninformative Stable < Informative Unstable - Uninformative Unstable.

In the model output below, I report the mean values for each parameter, the standard deviation of each estimate, 95% credible intervals and the P(sign) – i.e. the probability that the parameter value has the sign estimated by the mean. For group differences, the posterior probability is computed for the direction of the effect predicted by the LGLH. For ease of exposition, I will refer to the evidence for effects in the following ways: if the 95% credible interval does not contain zero, this will be considered strong evidence for an effect. If the 95% credible interval contains zero, but the posterior probability for an effect is greater than 90%, I will call this moderate evidence for an effect. Though it is true that any probability greater than 50% indicates stronger evidence for an effect having the sign estimated by the model than the alternative sign, this ignores the fact that null effects are also possible. Conceptually, we can divide the probability space into three intervals: probabilities that most strongly favor a negative effect, probabilities that most strongly favor a positive effect, and probabilities that most strongly favor a null effect. If we do this, we apportion ~33% probability to each interval. Thus, if the posterior probability for an effect is greater than 66% but less than 90%, I will consider this weak evidence for an effect. Posterior probabilities between 33% and 66% will be considered evidence for a null effect. Finally, posterior probabilities less than 33% will be considered evidence for the opposite effect predicted by the LGLH.

2.1.5. Hypotheses

The LGLH implicates both the stability of gender representations and learning condition in the ability to use gender marking as an anticipatory cue. Based on prior work (Arnon & Ramscar, 2012), I expect that participants in the sequence first condition will have stable representations for more nouns than participants in the noun first condition (i.e. be more consistent at assigning the correct grammatical gender to nouns in the gender assignment tasks). If stability of gender representations is the primary driving factor for the ability to use gender marking predictively, I should observe approximately equal anticipatory effects on gender-stable trials, regardless of learning condition, and little to no evidence of anticipation for gender-unstable trials, with approximately equal magnitudes for both learning groups. If learning condition is the primary driving force, however, I should see clear evidence of gender-based anticipation in the sequence first group, and little or no evidence for anticipation in the noun first group, regardless of the stability of each noun's gender representation. It is further possible for these effects to be additive or interactive. If they are additive, I should observe prediction benefits both for gender-stable nouns and for sequence first learning. If the effects are interactive, I expect predictive benefits for gender-stable nouns in the sequence first group, with little or no benefit for stable nouns in the noun first group.

2.1.6. Results

2.1.6.1. Gender assignment performance

Figure 1 shows the proportion of nouns for which participants in each learning group assigned the correct grammatical gender on both iterations of the forced-choice task. This is taken as a measure of gender stability. In contrast to the hypothesis that the sequence first group would be more successful at learning grammatical gender, their mean gender stability was numerically lower than that of the noun first group (Sequence First: *M* = 0.359, *SD* = 0.131; Noun First: *M* = 0.362, *SD* = 0.192). Results from the model fit to these data are summarized in Table 1. These results show no evidence that learning condition had an impact on gender assignment stability, as indicated by the effect size of approximately zero, and the near chance probability of this effect being different from zero. Results do, however, find a 100% posterior probability of the mean performance being greater than chance (25%, or approximately 1.1 on the log odds scale). In short, participants' gender stability was above chance, but did not differ by learning condition.

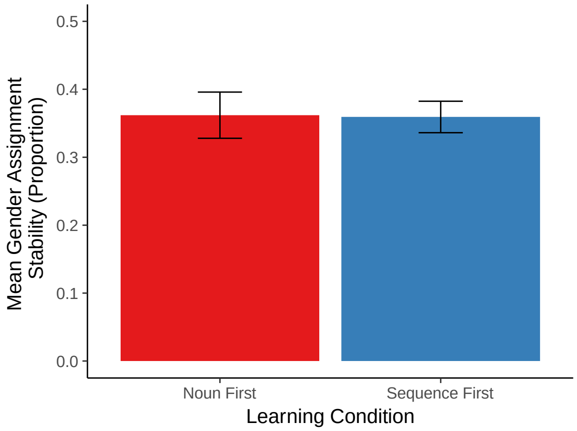

Figure 1. Mean proportion of nouns (with standard errors) for which participants in each group assigned the correct grammatical gender in both forced-choice tasks.

Table 1. Summary of model results for gender assignment stability.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-0.61	0.11	-0.82	-0.40	1
Learning Condition (Noun First)	0.00	0.10	-0.19	0.20	0.514

2.1.6.2. Eye tracking results

Figure 2 shows the difference in proportion of looks to the target noun and the proportion of looks to the distractor over time. These were calculated as looks to target minus looks to distractor for every sample spanning from the onset of the sentence out to 4000 ms. Thus, a positive difference indicates that participants looked more at the target than the distractor. The dashed vertical lines mark the points 200 ms after the acoustic onset of the article and 200 ms after the acoustic onset of the noun. Thus, these demarcate the “prediction window”, during which linguistically mediated eye movements can only plausibly be guided by information on the articles. Visual inspection of the plots indicates that neither group starts to preferentially look at the target on unstable trials until after the prediction window, irrespective of condition. That is, neither group shows evidence of anticipatory looks to the target when they do not have stable gender assignments for the target and distractor nouns on a given trial. On trials with stable gender assignments, both groups show evidence of increased looks to the target that start earlier on informative trials than uninformative trials. While the sequence first group seems to show a greater difference for informative compared to uninformative trials during the prediction window compared to the noun first group, the noun first group shows an earlier increase in their preferential looks to the target than the sequence first group. Consequently, it is unclear to what extent the visual pattern of the gaze data reflects any differences across groups in anticipatory use of grammatical gender.

The probabilities of fixating the target image during the prediction window are summarized in Table 2 for each condition on gender-stable and gender-unstable trials. Descriptively, both groups show a clear increase in proportions of looks to the target on stable informative trials relative to stable uninformative trials. The size of this effect, moreover, appears larger for the sequence first group compared to the noun first group. On unstable trials, there is no clear evidence of anticipatory looks to the target for either group.

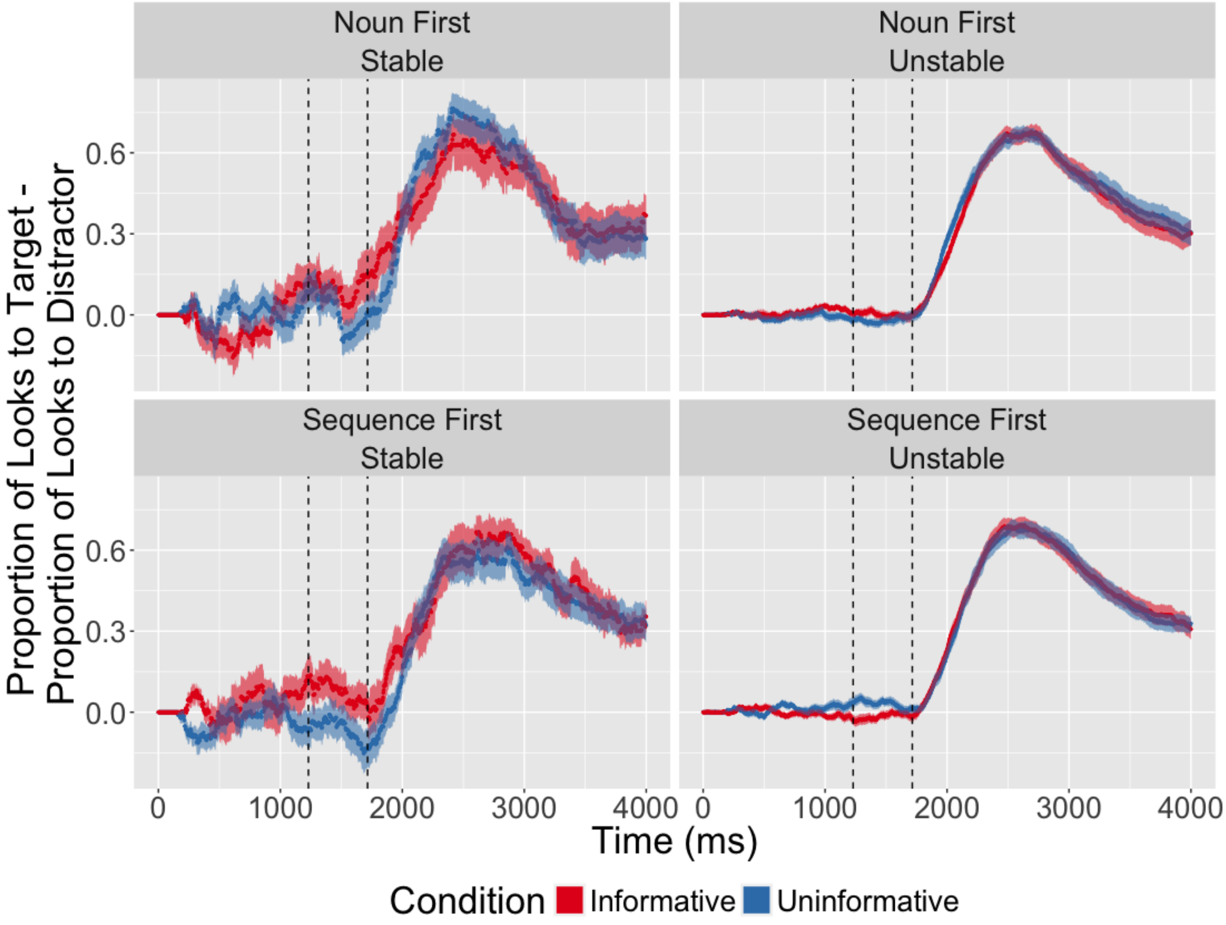

Figure 2. Difference in proportion of looks to target and proportion of looks to distractor, calculated as proportion of looks to target minus proportion of looks to distractor at each sample. Standard errors are shown around each sample. Dashed vertical lines are located at 200 ms after the acoustic onset of the article and 200 ms after the acoustic onset of the nouns. These thus mark the "prediction window", during which linguistically mediated eye movements can only be guided by information on the article.

Table 2. Mean proportion of fixations to target noun for each group in each condition for stable and unstable trials in the time window extending from 200 ms after the onset of the article until 200 ms after the onset of the noun. Standard deviations are shown in parentheses.

	Stable Trials		Unstable Trials	
	Informative	Uninformative	Informative	Uninformative
Noun First	0.484 (0.383)	0.401 (0.318)	0.356 (0.212)	0.345 (0.214)
Sequence First	0.447 (0.301)	0.326 (0.295)	0.368 (0.247)	0.393 (0.235)

Results of the model fit to these data are summarized in Table 3. While the raw data are consistent with an interaction between condition, learning group and stability in the direction predicted by the LGLH, this is not borne out by the model, which finds only extremely weak evidence for the three-way interaction. Moreover, as shown in Figure 3, the posterior group means estimated by the model for this interaction, indicate that the noun first group shows a larger effect of condition than the sequence first group, contra the predictions of the LGLH. Table 4, shows the posterior group difference for the contrasts of theoretical interest. These find weak evidence that the noun first group is more likely to predictively fixate the target on informative trials than on uninformative trials when gender assignments are stable. The evidence for all other contrasts is most consistent with no effects of condition. In addition, though the noun first group shows a numerically larger effect of condition, the posterior probability that the effect of condition is larger for the noun first group than for the sequence first group suggests that this difference is not likely reliable (Mean = 0.06, SD = 0.27, 2.5% = -0.44, 97.5% = 0.60, P(sign) = 0.584).

Table 3. Summary of model results for the fixation probability data.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-0.77	0.17	-1.09	-0.42	1
Learning Condition (Noun First)	0.00	0.17	-0.32	0.32	0.513
Trial Stability (Stable)	0.07	0.06	-0.05	0.19	0.876
Condition (Uninformative)	-0.01	0.05	-0.11	0.08	0.604
Learning Group (Noun First) x Trial Stability (Stable)	0.05	0.06	-0.06	0.16	0.792
Learning Group (Noun First) x Condition (Uninformative)	-0.03	0.05	-0.12	0.07	0.706
Trial Stability (Stable) x Condition (Uninformative)	-0.03	0.05	-0.13	0.06	0.763
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	0.02	0.05	-0.08	0.12	0.668

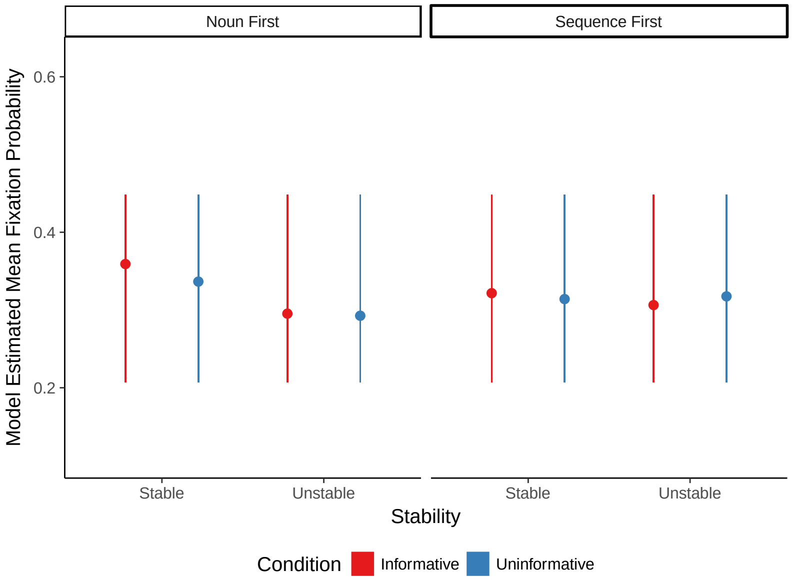

Figure 3. Model estimated posterior group means for fixation probability by condition, learning group and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 4. Posterior group differences for each group by stability and condition for fixation probabilities.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P > 0
Noun First	Stable	Informative - Uninformative	0.10	0.24	-0.36	0.57	0.673
	Unstable	Informative - Uninformative	0.01	0.13	-0.24	0.27	0.530
Sequence First	Stable	Informative - Uninformative	0.04	0.20	-0.35	0.43	0.573
	Unstable	Informative - Uninformative	-0.05	0.19	-0.44	0.33	0.394

Overall, the fixation probability data find some, albeit very weak evidence that gender stability influences the anticipatory use of grammatical gender, as predicted by the LGLH. These data do not, however, find evidence that learning condition influences gender stability or the predictive use of grammatical gender in the ways predicted by the LGLH.

First fixation latencies are summarized in Table 5. Descriptively, these data appear consistent with the LGLH. The sequence first learning group is faster overall to fixate the target on informative trials than on uninformative trials, and this difference between conditions is larger on stable trials compared to unstable trials. In contrast, the noun first group shows only a 3 ms advantage on stable informative trials over stable uninformative trials, but only as measured by the mean.

Table 5. Summary of first fixation latencies by learning group, condition and stability.

Learning Condition	Trial Stability	Condition	Mean	Median	SD
Noun First	Stable	Informative	916.34	938.14	345.79
		Uninformative	919.52	896.00	268.81
	Unstable	Informative	1024.84	981.25	223.45
		Uninformative	1004.45	998.63	227.72
Sequence First	Stable	Informative	1058.73	873.00	580.66
		Uninformative	1116.61	1005.00	468.12
	Unstable	Informative	1022.83	1003.05	279.00
		Uninformative	1064.23	1054.38	320.73

Modeling results for first fixation latencies are summarized in Table 6. Like the fixation probability data, this model finds only weak evidence for an interaction between learning group, stability and condition. Moreover, as is evident from examining the posterior group means plotted in Figure 4, the nature of this interaction is not consistent with the LGLH. Though the sequence first group shows earlier first fixation latencies on informative trials compared to uninformative trials, this difference is actually larger for unstable than for stable trials. The noun first group, in contrast, shows slower first fixation latencies on informative trials compared to uninformative trials, regardless of stability. Table 7 gives the posterior group differences of theoretical interest for this three-way interaction. Examining the posterior probabilities that targets are fixated earlier on informative trials compared to uninformative trials, the data do not favor the LGLH. Again, for the noun first group the evidence actually favors the alternative hypothesis. For the sequence first group there is only weak evidence for this hypothesis on unstable trials, but on stable trials the evidence more strongly favors a null effect.

Table 6. Summary of model results for the first fixation latencies.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	667.43	31.77	605.20	730.63	1
Learning Condition (Noun First)	-17.70	30.33	-78.33	40.32	0.721
Trial Stability (Stable)	-9.63	11.70	-33.11	13.31	0.802
Condition (Uninformative)	-3.47	9.19	-22.10	14.30	0.643
Learning Group (Noun First) x Trial Stability (Stable)	-16.84	9.98	-36.21	3.29	0.948
Learning Group (Noun First) x Condition (Uninformative)	-6.90	9.17	-25.63	10.94	0.768
Trial Stability (Stable) x Condition (Uninformative)	-0.54	9.56	-19.07	17.82	0.517
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	-5.66	9.20	-23.81	12.03	0.730

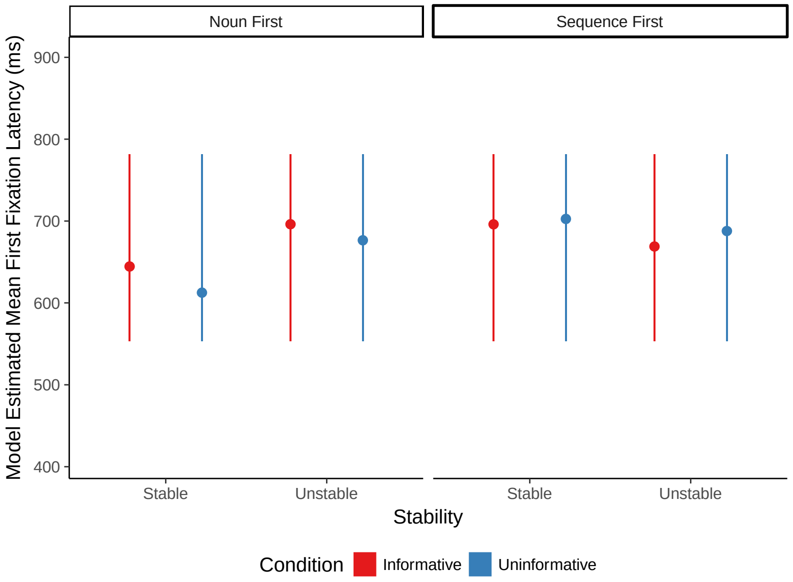

Figure 4. Model estimated posterior group means for first fixation latencies by learning group, condition and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 7. Posterior group differences for learning group by stability by condition for first fixation latencies.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P < 0
Noun First	Stable	Informative - Uninformative	33.15	43.81	-52.85	119.40	0.221
	Unstable	Informative - Uninformative	19.65	24.42	-27.07	68.08	0.212
Sequence First	Stable	Informative - Uninformative	-5.76	40.18	-82.84	73.42	0.567
	Unstable	Informative - Uninformative	-19.26	36.74	-91.23	53.98	0.699

In short, the first fixation latency results find only weak evidence that a more L1-like learning condition can lead to more predictive use of grammatical gender. Like the fixation probability data, there is no evidence that learning condition interacts with gender stability to produce greater anticipatory effects for stable trials when participants had a more L1-like learning context.

Finally, the response times for clicking on the target image relative to the onset of the article are summarized in Table 8. Descriptively, both groups of participants show faster response times on informative trials relative to uninformative trials when gender assignments are stable. The sequence first group also shows faster response times to informative trials when gender assignments are unstable, whereas the noun first group does not. As measured by the mean, moreover, the sequence first group's informativity effect is larger for stable trials compared to unstable trials. Thus, the raw data appear consistent with the LGLH.

Table 8. Summary of latencies for clicking on the target image by learning group, condition and stability.

Learning Condition	Trial Stability	Condition	Mean	Median	SD
Noun First	Stable	Informative	1907.04	1753.65	538.98
		Uninformative	1953.22	1772.28	604.60
	Unstable	Informative	2013.99	2000.18	345.31
		Uninformative	2003.80	1964.71	316.21
Sequence First	Stable	Informative	1960.01	1868.91	372.14
		Uninformative	2071.18	1917.86	431.91
	Unstable	Informative	1956.30	1947.65	232.99
		Uninformative	1973.98	2003.99	254.22

Model results for the response times are summarized in Table 9. Results show that the posterior probability for the interaction between learning group, stability and condition is most consistent with a null effect. The posterior group means for this interaction are shown in Figure 5. As can be seen, both groups are faster to select the target image on informative trials than uninformative trials when gender assignments are stable. In contrast, response times across conditions are very similar when gender assignments are unstable. Table 10 summarizes the posterior group differences for the contrasts of theoretical interest. Consistent with the strong evidence for an interaction between trial stability and condition found by the model, both groups show a high posterior probability for stable informative trials having faster response times than stable uninformative trials, whereas for unstable trials both groups are more likely to respond more slowly on informative trials than uninformative trials. While the evidence for the effect of informativity on stable trials is stronger for the sequence first group than the noun first group, the difference in effect sizes is only about 5 ms, which is unlikely to reflect a meaningful cognitive difference across these groups.

Table 9. Summary of model results for the response times.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	1445.35	31.48	1382.72	1508.47	1
Learning Condition (Noun First)	-22.70	26.02	-72.85	29.96	0.811
Trial Stability (Stable)	1.29	10.11	-18.60	21.12	0.550
Condition (Uninformative)	11.78	6.55	-0.71	24.83	0.968
Learning Group (Noun First) x Trial Stability (Stable)	-8.25	10.28	-28.51	11.23	0.783
Learning Group (Noun First) x Condition (Uninformative)	-2.02	6.70	-15.33	10.57	0.610
Trial Stability (Stable) x Condition (Uninformative)	14.83	6.52	2.03	27.32	0.988
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	1.55	6.66	-11.56	14.29	0.593

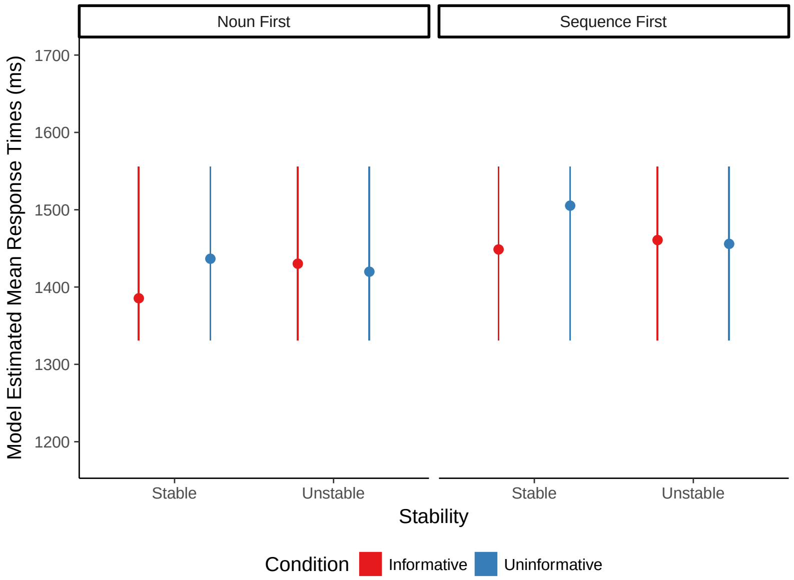

Figure 5. Model estimated posterior group means for response times by learning group, condition and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 10. Posterior group differences for learning group by stability by condition for response times.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P < 0
Noun First	Stable	Informative - Uninformative	-52.29	33.19	-116.61	13.74	0.941
	Unstable	Informative - Uninformative	10.13	17.15	-23.04	44.56	0.282
Sequence First	Stable	Informative - Uninformative	-57.25	27.82	-113.26	-3.76	0.982
	Unstable	Informative - Uninformative	5.17	26.66	-48.50	55.68	0.417

Summarizing, the response time data find strong evidence for an interaction between stability and condition, such that participants are faster to select the target image on stable informative trials than stable uninformative trials, but do not show this difference for unstable trials. The model does not find strong evidence for an interaction between learning group, stability and condition. Thus, these data find support for effects of gender stability on the anticipatory use of grammatical gender, but not learning context.

2.1.7. Experiment 1 summary and discussion

Experiment 1 used an artificial language learning task in conjunction with a visual world eye tracking task to examine the extent to which learning a language in a more L1-like manner would lead to better learning of grammatical gender and to greater anticipatory use of grammatical gender, as well as the extent to which the stability of gender representations impacts the anticipatory use of grammatical gender. The LGLH and prior work (i.e. Arnon & Ramscar, 2012), predict that 1) participants in the sequence first group should have stable gender assignments for more nouns than participants in the noun first group; 2) the sequence first learning group should show greater anticipatory effects than the noun first group; and 3) anticipatory effects should be larger on trials with stable gender assignments.

Results of Experiment 1 only find clear evidence for the third prediction. The gender assignment results found only a 51.4% posterior probability for an effect of learning group on gender assignment stability, which is most consistent with a null effect. Moreover, even if the effect were real, it was the noun first group, not the sequence first group, that showed a higher likelihood of having stable gender assignments. These results thus do not provide support the first claim made by the LGLH, that more L1-like learning contexts lead to better learning of grammatical gender. That said, it is important to also bear in mind that these results do not provide strong evidence *against* the first claim of the LGLH, as it is possible that learning context does matter, but that its effects do not stem from the use of different cues to words and word boundaries.

The results of this experiment also contrast with prior work using the same artificial language learning task (Arnon & Ramscar, 2012; Siegelman & Arnon, 2015), which have found that learning conditions that reduce the cues to word boundaries lead to better learning of grammatical gender (see also Paul & Grüter, 2016 for similar findings with Chinese classifiers). One reason for this discrepancy between previous studies and the current one may be the difficulty of the learning task. The previous experiments with grammatical gender used 14 and 12 nouns respectively, whereas the current experiment used 24. Though the number of repetitions of each noun in each learning condition was the same in all experiments, it is possible that the larger number of nouns in this experiment made learning more difficult overall, and that this may have eliminated any benefits conferred by the sequence first learning condition. If so, we should expect to see effects of learning condition if more time is given for learning, or perhaps if time is allowed for offline consolidation, which has been shown to play an important role in word learning (e.g. Bakker, Takashima, van Hell, Janzen, & McQueen, 2015; Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010). Another possibility is to make learning easier, by reducing the number of items or perhaps by providing some amount of explicit information, which has generally been shown to lead to more successful learning, at least in tasks involving a relatively short amount of overall exposure (see Norris & Ortega, 2000 for a review).

It is also important to note that the current experiment differs from the previous studies in that gender learning was not directly assessed after the learning task, but rather the visual world eye tracking task intervened between training and learning assessment. This effectively provided further opportunity for participants to continue learning the artificial grammar while also testing their knowledge. Thus, to whatever extent learning condition affected the initial learning of the artificial grammar, additional learning during the eye tracking task may have eliminated these effects.

The results of Experiment 1 also do not support the second claim of the LGLH, that more L1-like learning conditions should lead to greater anticipatory use of grammatical gender. None of the models fit to the eye tracking data found strong evidence for an interaction between learning group and condition, nor for an interaction between learning group, condition and stability. To the extent that there was evidence for an effect of learning group, the direction of the effects was not consistent with the LGLH. In the fixation probability data, it was the noun first group, not the sequence first group that showed greater effects of gender assignment stability on their anticipatory use of grammatical gender, though again the evidence for this effect was overall quite weak. In the first fixation latencies, the effect of condition for the noun first group was in the wrong direction for anticipatory effects (i.e. they fixated the target faster on uninformative trials). For the sequence first group the effect of condition was in the appropriate direction for anticipatory effects, however the effect was larger on unstable trials than on stable trials, which is the opposite direction predicted by the LGLH if stable gender assignments truly lead to larger anticipatory effects. Lastly, the model fit to the response time data did not find evidence for any effects of learning condition.

Given the fact that no effects of learning condition were found on gender assignment stability, it is not surprising to also find no effects of learning condition on the anticipatory use of gender. As was discussed above, the lack of these effects may be due to the difficulty of the learning task and/or additional learning taking place during the eye tracking task. Another possibility that was not discussed above is that the effects of learning condition reported in previous experiments may at least partly reflect the use probabilistic cues to gender. In particular, whereas care was taken in this experiment to ensure that word-form did not provide a reliable cue to gender, the prior studies did not report controlling for this. Arnon and Ramscar, (2012) provide their materials, however they do not report how nouns were divided into gender classes. It is therefore not possible to assess the plausibility of this explanation for their results. Siegelman and Arnon (2015) do report how nouns were divided into gender classes, and there are indeed numerous imbalance across gender classes in word form properties that could have provided additional cues to grammatical gender (for example, 6/6 nouns in one class have a CVC ending, whereas 3/6 nouns in the other class end in an open syllable. 5/6 nouns in the first class also end with an <o> followed by a consonant, whereas only one noun in the other class has this type of ending). It is therefore possible that the effects of learning condition observed in these studies may not reflect chunk-based learning, as hypothesized by Arnon and Christiansen, (2017), but rather may reflect differences in associative, form-based learning. If so, the fact that form-based cues to gender were tightly controlled in the current experiment could explain the lack of any effects of learning group.

Finally, whereas Experiment 1 did not find evidence for the first two claims of the LGLH, it did find evidence for the third claim, that gender stability modulates the anticipatory use of grammatical gender. In particular, there was weak evidence in the fixation probability data for an interaction between stability and condition, such that participants were more likely to fixate the target on informative trials than uninformative trials during the prediction window when assignments were stable, but not when assignments were unstable. The fixation latency data do not find evidence for this effect, but the response time data showed strong evidence for an interaction between stability and condition in the direction predicted by the LGLH. Specifically, participants were faster to click on the target image on informative trials than on uninformative trials when gender assignments were stable, but not when assignments were unstable.

Though this pattern of results is clearly consistent with an effect of gender stability, the fact that there were no effects in the fixation latency data and only weak effects in the fixation probability data make it somewhat difficult to confidently claim that this is anticipation, per se. An alternative account of gender facilitation effects on nouns that has been put forward in the literature is the post-lexical checking account (Bates et al., 1996; Friederici & Jacobsen, 1999). Under this account, upon encountering nouns with pre-nominal gender marking, comprehenders check the gender of the noun against the active gender information conveyed by preceding gender-marking. If this information is congruent, the checking process proceeds normally and is completed quickly. If this information is incongruent, costs are incurred which result in longer processing of the noun. Critically, this account claims that apparent gender facilitation effects reflect inhibition rather than facilitation, and that the locus of these effects is entirely post-lexical (i.e. after lexical access has occurred). Thus, while this account is compatible with effects of gender stability, it should not predict an effect of informativity. That is, if gender facilitation effects reflect post-lexical rather than pre-lexical processes, then it should be just as easy to check a noun against active gender information when there are two nouns belonging to the same gender class as when there are two nouns of different gender classes. Similarly, though a facilitated integration account could also potentially be compatible with the effect of gender stability, it is not clear that such an account would predict an effect of informativity. In particular, facilitated integration accounts claim that apparent anticipatory effects reflect the ease of integrating linguistic content into a sentence or discourse context, and that ease of integration depends on how plausible an item is given the preceding context. Given that all target nouns in this experiment occurred in the same context, the only linguistic content that should mediate their plausibility is the gender marked article. On informative trials, the target noun that is congruent with the gender marking would be more plausible than the gender-incongruent distractor, and thus would be easier to integrate than the distractor would be. On uninformative trials, however, both nouns are equally plausible, but critically should also be just as plausible given the linguistic content as the target noun on informative trials. Consequently, a facilitated integration account does not seem to predict the effect of informativity observed in the response times.

Given the lack of a clear alternative explanation, and the strong evidence for the role of anticipatory processes in language comprehension, it seems reasonable to assume that the response time data do, in fact, reflect anticipation. This, however, raises the question of why no clear anticipatory effects were seen in the eye movement data. One possible explanation is that there was simply not enough time between the article and the onset of the noun for robust anticipatory looks to manifest given the limited experience that participants had with the artificial grammar. If this were the case, however, we should expect to find little or no evidence for anticipation in the fixation probability data, but we should still see evidence for anticipation in the first fixation latencies given that this experiment did not use a restricted time window. Contrary to this, Experiment 1 showed weak evidence for anticipation in the fixation probability data, but no clear evidence for anticipation in the fixation latency data. Thus, it does not seem likely that the eye movement results can be explained by the time course of anticipatory processing.

A more likely explanation that is also in line with an explanation put forward for the lack of group effects in the gender assignment data is that eye movements during the visual world task were most strongly guided by learning processes rather than anticipatory processes. If so, as discussed earlier, making the learning task easier, longer and/or allowing for consolidation should lead to greater evidence for anticipation in the eye movement data. Another way to assess this possibility is to use participants whose L1 has grammatical gender. In light of prior research indicating that having gender in one's L1 leads to better learning of gender in an L2 (Sabourin et al., 2006) and to more effective use of gender as an anticipatory cue in an L2 (Dussias et al., 2013), Experiment 2 will be able to assess the plausibility of this explanation, albeit indirectly.

2.2. Experiment 2

2.2.1. Participants

Participants in Experiment 2 were 41 native speakers of German (29 Female; mean age = 23.63 years; range: 18-37), all of whom reported normal or corrected-to-normal vision and no known hearing impairments. Data for 30 of these participants were collected at a university in Germany. Data for the remaining participants were collected at a large university in the United States.

2.2.2. Materials

Materials were identical to those used in Experiment 1.

2.2.3. Procedure

The procedure for the training task and the gender assignment tasks was identical to Experiment 1. For participants whose data were collected in the United States, eye movements were recorded from each participant's right eye using a desk-mounted Eyelink 1000+ eye tracker (SR Research) with a 500 Hz sampling rate. A 9-point calibration was used at the beginning of the experiment; a drift check was performed at the beginning of each trial. For participants whose data were collected in Germany, eye movements were recorded from participant's right and left eyes using a desk-mounted SMI Red 500 eye tracker (SensoMotoric Instruments) with a 500 Hz sampling rate. A 9-point calibration was used at the beginning of the experiment. No drift checks were performed due to limitations of the SMI equipment. Only data from the right eye were analyzed. All participants were seated at an approximately 100 cm viewing distance from the computer monitor.

2.2.4. Data preprocessing and analysis

Data processing and analyses were identical to Experiment 1. Native German speakers' production data were not analyzed for the same reasons as in Experiment 1. In addition, technical issues during online recording with the SMI recording software resulted in the loss of a large number of response times for clicks on the target image.

2.2.5. Hypotheses

Hypotheses with respect to the LGLH are the same as in Experiment 1. In addition, because the presence of gender in the L1 has been shown to facilitate L2 gender learning and processing (Franceschina, 2005; Sabourin & Stowe, 2008; Sabourin et al., 2006), I expect to observe greater gender assignment stability and more robust anticipation in the L1-German speakers compared to the L1-English speakers.

2.2.6. Results

2.2.6.1. Gender assignment performance

Figure 6 shows the proportion of nouns for which participants in each learning group assigned the correct grammatical gender in both iterations of the forced-choice task. As in Experiment 1, results find a 100% posterior probability of the mean performance being greater than chance. In contrast to Experiment 1, the sequence first group is numerically more likely to have stable gender assignments for more nouns than the noun first group (Sequence First: *M* = 0.363, *SD* = 0.145; Noun First: *M* = 0.343, *SD* = 0.117). Results from the model fit to these data are summarized in Table 11. These results do find evidence that learning condition had an impact on gender assignment stability, such that the sequence first group is more likely to have stable gender assignments for a noun. The evidence for this effect, however, is very weak, and the effect size is so small (~1%) that it may not be practically meaningful.

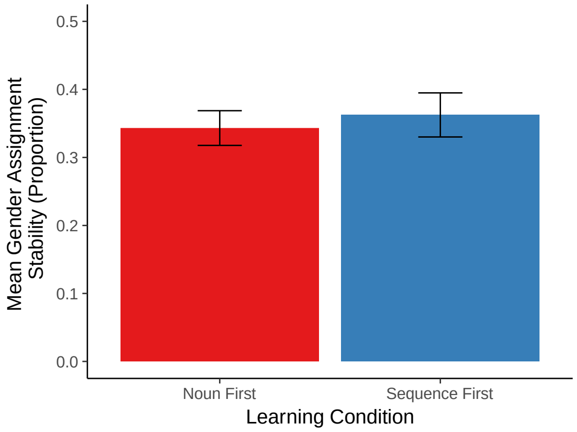

Figure 6. Mean proportion of nouns (with standard errors) for which participants in each group assigned the correct grammatical gender in both forced-choice tasks.

Table 11. Summary of model results for gender assignment stability.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-0.64	0.11	-0.86	-0.43	1
Learning Condition (Noun First)	-0.04	0.10	-0.24	0.15	0.672

2.2.6.2. Eye tracking results

The difference in proportion of looks to the target noun and the proportion

of looks to the distractor over time is shown in Figure 7 for each group by condition and stability. On unstable trials, neither group shows evidence of preferentially orienting to the target window until after the prediction window, regardless of condition. On stable trials, it does appear that the sequence first group may start to preferentially look toward the target during the prediction window, however the large standard errors the high degree of jitter across nearby samples make it difficult to be confident that this is not noise.

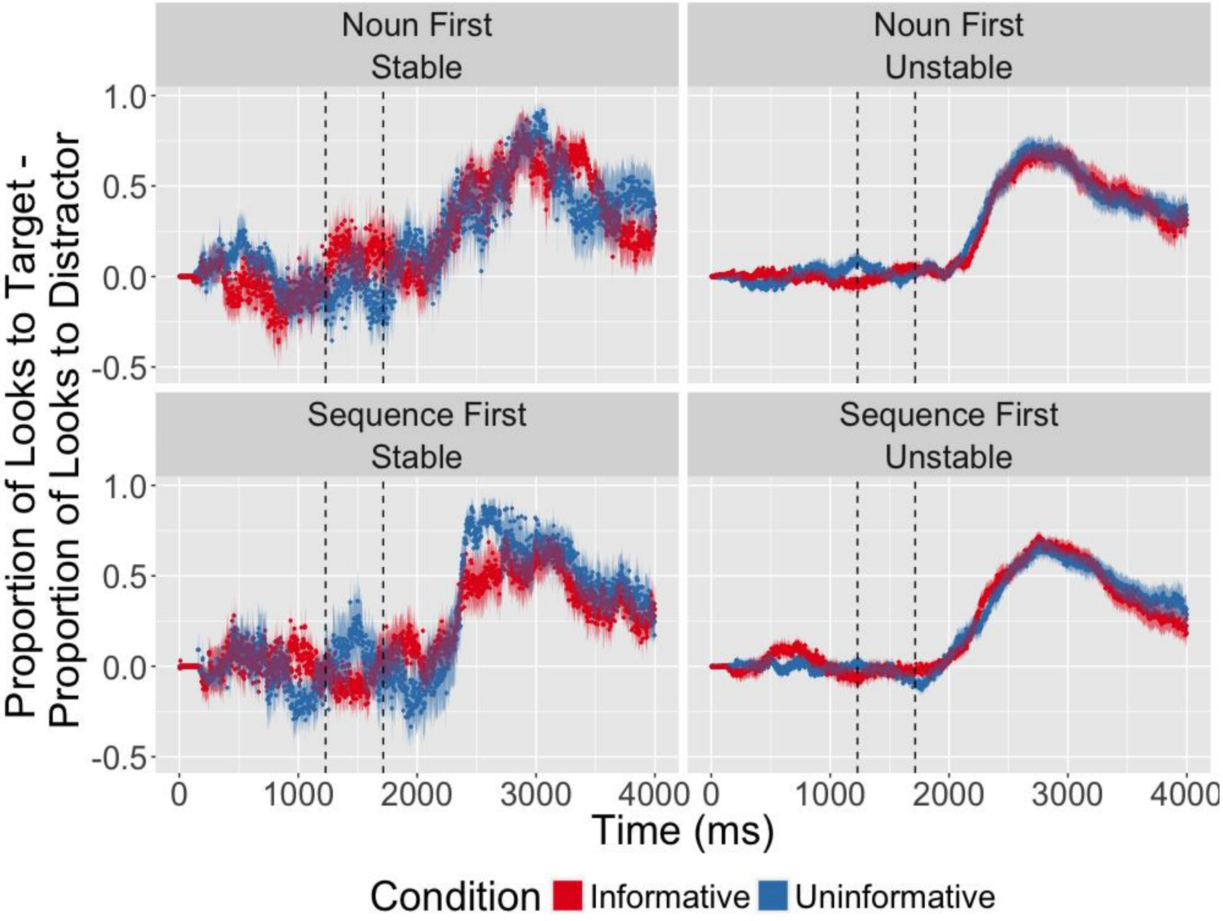

Figure 7. Difference in proportion of looks to target and proportion of looks to distractor, calculated as proportion of looks to target minus proportion of looks to distractor at each sample. Standard errors are shown around each sample. Dashed vertical lines are located at 200 ms after the acoustic onset of the article and 200 ms after the acoustic onset of the nouns. These thus mark the “prediction window”, during which linguistically mediated eye movements can only be guided by information on the article.

Table 12 summarizes the probabilities of fixating the target during the prediction window by learning group, stability and condition. Unlike Experiment 1, only the noun first group shows a greater probability of fixating the target during the prediction window for stable informative trials compared to uninformative trials. The results of the model fit to these data are summarized in Table 13. The posterior probability for the group x stability x condition interaction estimated by the model is 84.6%, comprising weak evidence for this effect. As can be seen in Figure 8, however, the posterior group means are not consistent with the LGLH. The noun first group is more likely to fixate the target on informative trials when gender assignments are stable and when they are unstable, and this difference is larger for stable trials. The sequence first group, in contrast, only shows an increased probability of fixating the target on informative over uninformative trials when gender assignments are stable. The posterior group differences for the these effects are given in Table 14. These show that the posterior probability for an informativity effect on stable trials in the direction predicted by the LGLH is 93.5% for the noun first group compared to only 76.4% for the sequence first group. This effect, moreover, is nearly three times as large for the noun first group as for the sequence first group. In short, these results are consistent with an effect of stability on the anticipatory use of grammatical gender, but the evidence does not support the effect of learning condition predicted by the LGLH.

Table 12. Mean proportion of fixations to target noun for each group in each condition for stable and unstable trials in the time window extending from 200 ms after the onset of the article until 200 ms after the onset of the noun. Standard deviations are shown in parentheses.

	Stable Trials		Unstable Trials	
	Informative	Uninformative	Informative	Uninformative
Noun First	0.251 (0.263)	0.193 (0.214)	0.227 (0.175)	0.222 (0.168)
Sequence First	0.306 (0.336)	0.337 (0.378)	0.229 (0.159)	0.213 (0.135)

Table 13. Summary of model results for the fixation probability data.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-1.70	0.25	-2.22	-1.22	1
Learning Condition (Noun First)	-0.10	0.24	-0.58	0.37	0.666
Trial Stability (Stable)	-0.05	0.11	-0.29	0.14	0.684
Condition (Uninformative)	-0.09	0.08	-0.25	0.07	0.873
Learning Group (Noun First) x Trial Stability (Stable)	-0.09	0.10	-0.30	0.11	0.812
Learning Group (Noun First) x Condition (Uninformative)	-0.06	0.09	-0.23	0.10	0.761
Trial Stability (Stable) x Condition (Uninformative)	-0.09	0.08	-0.24	0.06	0.863
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	-0.08	0.08	-0.24	0.07	0.846

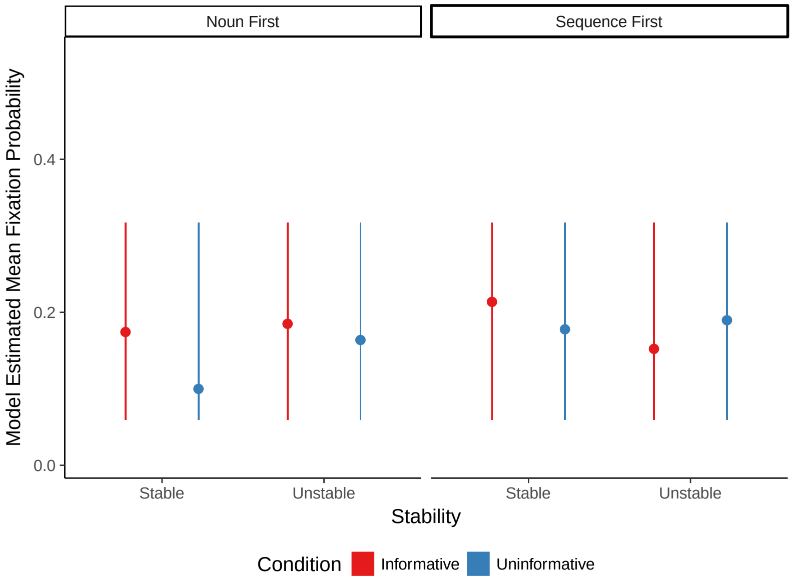

Figure 8. Model estimated posterior group means for fixation probability by condition, learning group and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 14. Posterior group differences for group by stability by condition for fixation probabilities.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P > 0
Noun First	Stable	Informative - Uninformative	0.64	0.43	-0.20	1.49	0.935
	Unstable	Informative - Uninformative	0.14	0.21	-0.29	0.55	0.735
Sequence First	Stable	Informative - Uninformative	0.24	0.33	-0.42	0.90	0.764
	Unstable	Informative - Uninformative	-0.27	0.33	-0.90	0.37	0.216

First fixation latencies are summarized in Table 15 for each learning group by condition and stability. The results of the model fit to these data are summarized in Table 16. The model finds only weak evidence for an interaction between learning group, condition and stability. Examining the posterior group means plotted in Figure 9, it is apparent that the nature of this interaction is not consistent with the LGLH. Both groups appear to show effects of stability, however the noun first group shows a larger effect of condition than the sequence first group, contra the predictions of the LGLH. These observations are confirmed in Table 17, which summarizes the posterior group differences of theoretical interest. This table further indicates that only the noun first group shows evidence for an effect of condition in the direction predicted by the LGLH, whereas the evidence for the sequence first group most strongly favor a null effect.

Table 15. Summary of first fixation latencies by learning group, condition and stability, aggregated over participants.

Learning Condition	Trial Stability	Condition	Mean	Median	SD
Noun First	Stable	Informative	1138.78	1069.56	400.95
		Uninformative	1147.42	1194.19	303.33
	Unstable	Informative	1141.31	1117.50	261.63
		Uninformative	1165.47	1107.96	260.67
Sequence First	Stable	Informative	1170.70	1114.24	467.07
		Uninformative	982.35	961.31	373.63
	Unstable	Informative	1161.32	1151.75	256.06
		Uninformative	1194.29	1136.48	259.54

Table 16. Summary of model results for the first fixation latencies.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	821.94	43.77	736.53	908.44	1
Learning Condition (Noun First)	-2.68	43.06	-84.89	82.72	0.520
Trial Stability (Stable)	-7.32	14.04	-34.44	19.80	0.701
Condition (Uninformative)	6.08	12.31	-18.12	30.26	0.692
Learning Group (Noun First) x Trial Stability (Stable)	10.18	13.56	-16.04	37.50	0.781
Learning Group (Noun First) x Condition (Uninformative)	6.23	12.41	-17.70	30.52	0.694
Trial Stability (Stable) x Condition (Uninformative)	1.20	12.19	-22.70	24.91	0.543
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	5.62	11.86	-18.03	28.46	0.688

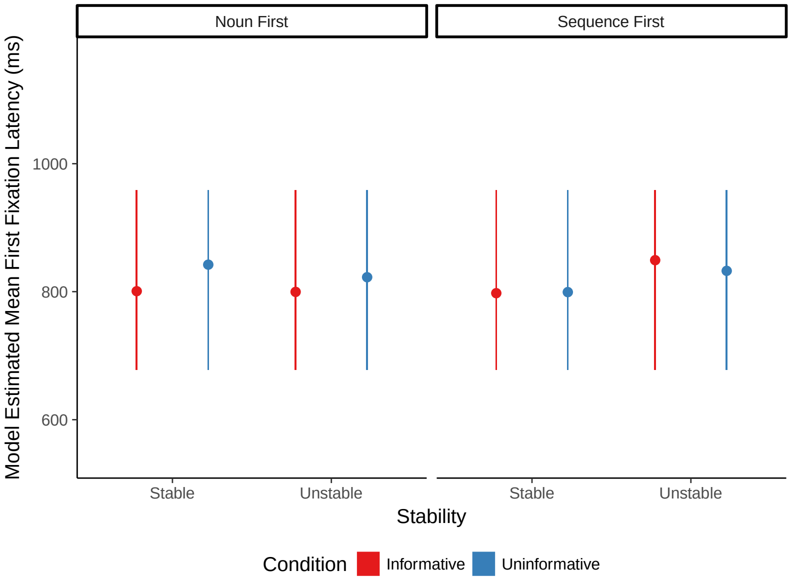

Figure 9. Model estimated posterior group means for first fixation latencies by learning group, condition and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 17. Posterior group differences for learning group by stability by condition for first fixation latencies.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P < 0
Noun First	Stable	Informative - Uninformative	-38.26	66.63	-167.60	91.50	0.718
	Unstable	Informative - Uninformative	-22.21	30.77	-81.75	37.93	0.769
Sequence First	Stable	Informative - Uninformative	-2.10	49.90	-97.00	95.11	0.511
	Unstable	Informative - Uninformative	13.95	47.67	-77.52	105.35	0.385

Finally, the response time latencies for clicking on the target image are summarized in Table 18 for each learning group by condition and stability. Table 19 summarizes the model results for these data. The posterior probability for a learning group x stability x condition effect most strongly favors a null effect. There is, however, strong evidence for an effect of condition and moderate evidence for an interaction between stability and condition. The posterior group means are plotted in Figure 10 for each group by condition and stability. These show clear effects of condition and stability, such that response times are faster on informative trials than uninformative trials, and this difference is larger on stable trials than on unstable trials. The posterior group differences shown in Table 20 are consistent with these observations. For the sequence first group, there is moderate evidence on stable trials that response times are faster on informative trials than uninformative trials; on unstable trials, the evidence best favors a null effect. For the noun first group, there is moderate evidence for an overall effect of condition, with response times being faster on informative trials compared to uninformative trials, irrespective of stability. Both groups show larger effects of condition on stable trials compared to uninformative trials, with approximately equal magnitudes (93.6 ms effect for the noun first group; 93.59 ms effect for the sequence first group). The evidence thus does not support an effect of learning condition on the interaction between stability and condition. The posterior probability that the condition effect is larger on stable than on unstable trials is 89.9%, irrespective of group. Thus, there is moderate evidence in favor of the effect of gender assignment stability on condition that is predicted by the LGLH.

Table 18. Summary of latencies for clicking on the target image by learning group, condition and stability, aggregated over participants.

Learning Condition	Trial Stability	Condition	Mean	Median	SD
Noun First	Stable	Informative	2186.79	2103.66	460.05
		Uninformative	2225.40	2179.83	280.74
	Unstable	Informative	2232.63	2163.89	425.07
		Uninformative	2547.98	2391.63	489.16
Sequence First	Stable	Informative	2143.23	2150.35	545.84
		Uninformative	2006.63	1936.80	320.10
	Unstable	Informative	2308.73	2368.39	326.29
		Uninformative	2657.63	2453.21	760.72

Table 19. Summary of model results for the response times.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	1666.98	43.00	1586.24	1756.04	1
Learning Condition (Noun First)	-3.40	41.64	-89.17	78.79	0.528
Trial Stability (Stable)	-42.99	19.34	-81.69	-4.38	0.985
Condition (Uninformative)	42.27	15.60	12.94	74.12	0.997
Learning Group (Noun First) x Trial Stability (Stable)	1.98	16.70	-32.21	33.75	0.557
Learning Group (Noun First) x Condition (Uninformative)	9.10	14.73	-20.26	37.78	0.733
Trial Stability (Stable) x Condition (Uninformative)	20.53	15.98	-12.74	51.67	0.901
Learning Group (Noun First) x Trial Stability (Stable) x Condition (Uninformative)	5.73	14.87	-23.29	34.35	0.654

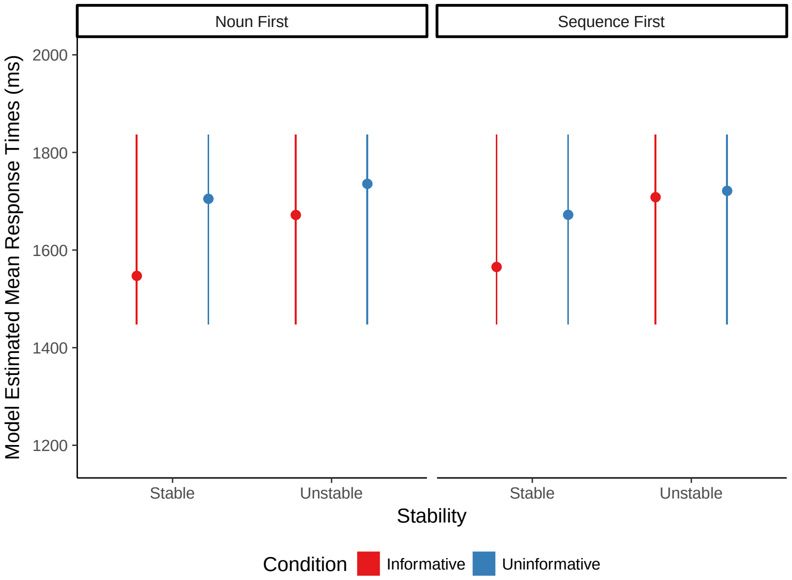

Figure 10. Model estimated posterior group means for response times by learning group, condition and stability. Dots show the mean probability of fixating the target. Vertical lines show the 95% credible intervals.

Table 20. Posterior group differences for learning group by stability by condition for response times.

Learning Condition	Stability	Contrast	Mean	SD	2.5%	97.5%	P < 0
Noun First	Stable	Informative - Uninformative	-155.28	83.90	-320.23	12.46	0.967
	Unstable	Informative - Uninformative	-61.68	40.32	-141.60	17.14	0.933
Sequence First	Stable	Informative - Uninformative	-107.40	60.39	-228.41	11.09	0.964
	Unstable	Informative - Uninformative	-13.81	60.63	-133.61	102.95	0.585

2.2.7. Experiment 2 summary

Using the same design as Experiment 1, Experiment 2 examined whether having grammatical gender in one's L1 might lead to better learning of grammatical gender and whether this in turn might lead to clearer effects of learning context and gender stability on the anticipatory use of grammatical gender, as predicted by the LGLH. Like Experiment 1, the results of Experiment 2 are largely inconsistent with the claim that a more L1-like learning context leads to better learning of grammatical gender and better use of gender as an anticipatory cue. Though the gender assignment data are numerically consistent with what was predicted based on the LGLH and prior work, the weak evidence for this effect and the very small effect size suggest that this may just reflect noise. Examining the interactions between learning group, condition and stability, Experiment 2 found, at best, only moderate evidence for this. As in Experiment 1, though, it was the noun first group rather than the sequence first group that showed the best evidence for using gender as an anticipatory cue. This is thus not consistent with the predictions of the LGLH with respect to learning context. Finally, Experiment 2 did find evidence that gender stability modulates the anticipatory use of grammatical gender, though as in Experiment 1, the evidence for this was weak or absent in the eye tracking data, but moderately strong in the response time data. These interactions will be discussed in more detail below. Taken together, the data for Experiment 2 are consistent with the data from Experiment 1 in finding no support for the effects of learning context predicted by the LGLH, but in finding clear support that the extent to which gender is used as an anticipatory cue depends critically on the stability of gender representations.

2.3. Discussion of Experiments 1 and 2

Considering the results of Experiments 1 and 2 together, the results of Experiment 2 are overall very similar to Experiment 1. With respect to gender assignment stability, the model estimated mean probability of having stable gender assignments was actually lower for Exp. 2 (34.5%) than for Exp.1 (35.2%). Consequently, these data do not support the hypothesis that having gender in one's L1 leads to better learning of grammatical gender in an L2. An important caveat to this, however, is that the cumulative experience each participant had with the artificial grammar was very limited, whereas prior work finding effects of L1 on gender learning and use have used learners of natural languages with much more experience with the language. Thus, it will be important for future work to disentangle the extent to which the lack of clear effects of L1 on gender learning found in this study reflect a true absence of any effects and the extent to which such effects may emerge over time with longer exposure and offline consolidation.

As in Experiment 1, the results of this Experiment 2 are not consistent with the effect of learning context predicted by the LGLH. The fixation probability and first fixation latency data both found weak evidence for an interaction between learning group, condition and stability, however it was the noun first group that showed larger effects of condition, contra the predictions of the LGLH. The posterior probability of this three-way interaction in the reaction time data is most consistent with a null effect, but even if the evidence were better, it was again the noun first group showing larger effects of condition. Weighed together, the fact that the results of Experiments 1 and 2 are not consistent with the effects of learning context predicted by the LGLH provides stronger evidence against the existence of an effect of learning context on gender learning and the anticipatory use of gender, at least under conditions of very limited exposure. As was discussed for Experiment 1, however, it will be important to determine whether the absence of these effects might reflect task difficulty and/or a lack of time for offline consolidation. Minimally, to whatever extent learning difficulty may have played a role, the fact that Experiment 2 still does not find clear evidence for a group effect indicates that having gender in one's L1 in insufficient to overcome any such difficulties.

Finally, Experiments 1 and 2 both find evidence that the anticipatory use of grammatical gender is mediated by gender stability. The evidence for this, moreover, is cumulatively somewhat stronger in Experiment 2 than in Experiment 1. In the fixation probability data, there is weak evidence for an interaction between stability and condition in both experiments. In comparison to Experiment 1, though, the posterior probability and effect size for this interaction are both larger in Experiment 2 (Exp. 1: Beta = -0.03, P(sign) = 76.3%; Exp. 2: Beta = -0.09, P(sign) = 86.3%). Whereas the first fixation latency results for Experiment 1 did not find evidence for anticipation, in Experiment 2 the pattern of fixation latencies is consistent with anticipation, showing earlier latencies on informative trials, with larger effects when gender assignments were stable. That said, the evidence for these effects is extremely weak. Finally, both experiments found clear evidence for anticipation in the response time data. In Experiment 1, there was moderate evidence for a main effect of condition (P(sign) = 96.8%) with an effect size of 11.78 ms. In Experiment 2, there was strong evidence for a main effect of condition (P(sign) = 99.7%) with an effect size nearly four times as large (42.27 ms). Both experiments, moreover, showed good evidence that the effects of condition were qualified by an interaction with gender stability. The evidence for this interaction is somewhat weaker in Experiment 2 compared to Experiment 1 (Exp. 1: P(sign) = 98.8%; Exp. 2: P(sign) = 90.01), but the data suggest that this difference may be due to a different nature of this interaction across experiments. In Exp. 1, response times to informative trials were only faster than uninformative trials when gender assignments were stable. On unstable trials, response times were actually slower on informative trials. These data thus indicate that the native English speakers only show reliable anticipatory use of grammatical gender when gender assignments are stable. In contrast, the native German speakers show faster response times on informative trials compared to uninformative trials irrespective of gender stability, but the size of this difference is larger on stable trials compared to unstable trials. Thus, whereas the native English speakers appear not to use grammatical gender as an anticipatory cue when gender assignments are unstable, the native German speakers seem to exhibit reduced rather than totally absent anticipation. This latter finding suggests that having grammatical gender in one's L1 may indeed lead to earlier and/or more robust use of grammatical gender as an anticipatory cue, consistent with earlier findings from Dussias et al. (2013).

The fact that the native German speakers showed greater evidence for anticipation, and larger effect sizes further suggests that having grammatical gender in one's L1 facilitates gender learning. On the one hand, this seems to contrast with the finding that L1 German speakers were no more successful than the L1 English speakers in learning the genders of nouns. However it may be that having gender in the L1 does not benefit all aspects of gender learning. In particular, these data suggest that having an L1 with grammatical gender may not make one more efficient or better at learning the gender of a noun in an L2. The benefits of L1 may instead reflect better associative learning, such that the L1 German speakers were more efficient at learning to associate a noun directly with its gender representations, or that L1 German speakers formed stronger associations between nouns and their gender representations. One reason for this may be that lifelong experience with grammatical gender makes L2 learners with grammatical gender marking in their L1 more likely to attend to gender cues early on. Another potential reason may have to do with language-specific experience with grammatical gender. Anecdotally, a number of the L1 English participants reported attempting to identify patterns in the nouns' semantics and/or word-forms that would help them classify the nouns into gender classes. Of the L1 English participants who had experience learning a language with grammatical gender, nearly all of these participants had learned Spanish, which employs a highly transparent gender-marking system in which the word-form is a highly reliable cue to gender class. Thus, their previous experience with gender-marking may have biased them toward expecting and searching for transparent cues to gender. In contrast, gender-marking in German is largely opaque, which may have biased the L1 German speakers toward a strategy of directly associating nouns with gender-marking words, rather than attempting to identify probabilistic relationships between gender classes and the properties of the nouns belonging to each class. Future work could evaluate these possibilities by including participants whose L1 uses a more transparent gender-marking system. If the differences in anticipatory behavior that were observed for the L1-German compared to L1-English participants reflect an effect of L1 experience influencing the types of cues that are attended to, participants with an L1 like Spanish should perform more similarly to the L1-German speakers than the L1-English speakers. To the extent that the observed differences may reflect strategies due to any previous experience with gender-marking languages, irrespective of L1, we might expect L1-Spanish speakers to perform either more like the L1-English speakers, or to exhibit intermediate effect sizes.

In summary, the cumulative results of Experiments 1 and 2 provide clear evidence that gender stability plays a critical role in the anticipatory use of grammatical gender in an L2. Consistent with one of the LGLH's claims, stronger evidence for anticipation was observed when gender assignments were stable compared to when gender assignments were unstable. The strength of the evidence and the size of these effects, moreover, tended to be larger for the L1-German speakers compared to the L1-English speakers. Contrary to the claims of the LGLH about learning context, Experiments 1 and 2 did not find evidence that more L1-like learning contexts which minimize cues to word boundaries and increase the likelihood of chunking determiner + noun sequences lead to better learning of grammatical gender and more robust use of gender marking as an anticipatory cue.

CHAPTER 3: NEURAL INDICES OF GENDER-BASED ANTICIPATION IN L1 AND L2 GERMAN

Experiments 1 and 2 reported evidence from an artificial grammar that gender stability modulates the anticipatory use of grammatical gender in an L2. In the present experiment, the claims of the LGLH with respect to the role of gender stability were tested using L2 learners of German in order to assess the extent to which the findings from Experiments 1 and 2 generalize to natural language. In addition, Experiments 1 and 2 found only weak evidence for anticipation and for effects of gender stability in the eye movement data. However, behavioral measures have been shown – at least in some cases – to underestimate performance in an L2, while neural measures can be more sensitive to subtle processing differences (McLaughlin et al., 2004; Tokowicz & MacWhinney, 2005). The use of event-related potentials may therefore reveal stronger evidence for gender-based anticipation *prior* to the onset of a target noun, in addition to which ERPs may be more sensitive to any modulating effects of gender stability. Experiment 3 therefore combines ERPs with a cued lateralized picture monitoring task to elicit the N2pc as an index of anticipatory processing.

The N2pc, or N2-posterior-contralateral, is an ERP component related to attention that consists of an enhanced negativity over posterior electrodes that are contralateral to a visually attended stimulus (see Luck, 2012 for a review). For example, if an attended stimulus is in the right visual field, the N2pc will be largest over left posterior sites. The N2pc is isolated by subtracting the waveform that is ipsilateral to a visually attended stimulus from the waveform that is contralateral to the stimulus. By examining when the N2pc deviates from zero, inferences can be drawn about the time course of shifts in covert visual attention (e.g. Kiss, Driver, & Eimer, 2009; Rommers, Meyer, & Praamstra, 2017; Woodman & Luck, 1999). This experiment thus uses the N2pc to track the time course of language-mediated shifts in covert visual attention while participants listen to short utterances and monitor sets of lateralized images.

Woodman, Arita, & Luck (2009) showed that when a target location is cued, attention can shift to the cued location prior to the onset of a target stimulus. When participants in their study viewed an array of boxes and were given a valid cue about the location of an impending stimulus, the N2pc was found to develop approximately 200 ms prior to the onset of the imperative stimulus. This finding thus provided strong evidence that participants were using the cues to anticipate the location of the imperative stimuli. The present experiment therefore adapts this type of cuing paradigm to language in order to examine whether grammatical gender marking can be used as a cue triggering covert shifts in attention to a target image prior to the onset of a target noun. It further examines whether variability in gender assignment across multiple tasks modulates this ability.

3.1. Participants

Data were collected from 42 native speakers of English. One of these participant was removed due to being a heritage speaker of German (see Montrul et al., 2014). Because ERP results can be strongly influenced by imbalances in noise across condition, ERP analysis was restricted to the 30 participants who had stable gender assignments for enough nouns that there were no fewer than 25 usable trials in each condition (stable/unstable x informative/uninformative x target left/target right). Data from all participants except the heritage speaker, however, were retained in the behavioral analyses. All L2 participants except one had a minimum of three years' experience learning German. The other participant had two years' experience, of which 11 months had been spent abroad on an exchange to Germany. Participants also included 13 native speakers of German (7 female). All participants reported normal or corrected-to-normal vision, no history of neurological impairment, no known hearing problems or color blindness, and did not report being on any psychoactive medication. Handedness was assessed with the Edinburgh Handedness Inventory (Oldfield, 1971), but left-handers were not excluded from participating. Seven L2 speakers and one L1 speaker considered themselves left-handed.

All participants completed a modified version of the Language Experience and Proficiency Questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007); German proficiency was assessed with the Goethe Institut Test. Demographic and proficiency information is summarized in Table 21.

Table 21. Summary of demographic and proficiency information for native and non-native speakers

		Mean	SD	Range
	Non-Native Speakers			
Demographics	Age (years)	22.2	5.0	18-40
	AoA (years)	13.8	2.4	10-21
	Length of Exposure (years)	8.0	3.3	2-19
Proficiency Measure	Goethe Institut Test (30 max)	17.7	5.2	7-28
Self-Rated German Proficiency	Speaking (10 max)	6.6	1.5	4-9
	Listening (10 max)	7.0	1.5	4-9
	Writing (10 max)	6.8	1.4	4-10
	Reading (10 max)	7.4	1.1	4-10
	Native Speakers			
Demographics	Age (years)	25.7	5.0	20-37
Proficiency Measure	Goethe Institut Test (30 max)	27.8	1.0	25-29
Self-Rated German Proficiency	Speaking (10 max)	9.9	0.3	9-10
	Listening (10 max)	10	-	-
	Writing (10 max)	9.8	0.6	8-10
	Reading (10 max)	10	-	-

3.2. Materials

The materials for Exp. 3 consisted of 60 grey-scale images taken from the Multilingual Picture Databank (Duñabeitia et al., 2017). All images depict concrete, morphologically simple nouns that have high German naming agreement (Duñabeitia et al., 2017) and mid–high frequency in German (Brysbaert et al., 2011). The depicted nouns were equally distributed across German's three gender classes, such that there were 20 masculine, 20 feminine, and 20 neuter nouns. Nouns were matched across gender classes for frequency, naming agreement, phonological neighborhood density (Marian, Bartolotti, Chabal, & Shook, 2012), and phonemic length (see Appendix B). Images were resized to 200x200 pixels.

Each image was paired with eight other images from the set of 60 to yield four pairs depicting nouns that were matched for grammatical gender, and four pairs that were mismatched in grammatical gender. Care was taken so that paired nouns did not share word-initial phonological cohorts. For the mismatching pairs, each noun was paired equally often with nouns of the two other gender classes. For example, a masculine noun was paired with two feminine nouns and two neuter nouns. This yielded a total of 480 noun-pairs. Each noun-pair occurred once in an experimental list, so that target nouns were not predictable on the basis of having previously seen a noun-pair.

To create the auditory stimuli for this experiment, each noun was embedded into a short phrase consisting of an imperative (*Guck mal!* or *Schau mal!* meaning "Look!"), a singular, definite, gender marked determiner in the nominative case, the adjective, *dargestellte*, (meaning *depicted*), and the noun. Note that in German, attributive adjectives take the same suffix (*-e*) when preceded by a definite, nominative gender-marked determiner. Thus, the only pre-nominal gender-marking information available to participants is on the determiner. A female native speaker of German was recorded producing each phrase at a slow–moderate pace with neutral intonation. Phrases were recorded in a sound-attenuating booth with a condenser microphone at 16-bit with a 44,100 Hz sampling rate. A single, representative token of each sentence element was spliced out of the recordings so that determiners and the adjective were identical across recordings. Determiners were matched in duration so that the interval between determiner onsets and noun onsets was identical in all conditions. In order to match determiners in duration, 10 ms of mid-vowel glottal pulses were added to masculine determiner *der*, 38.4 ms of mid-vowel glottal pulses were added to the feminine determiner *die*, and 1.097 ms of high frequency spectral noise were removed from the /s/ in the neuter determiner *das*. Final durations were 131.8 ms for *der*, 132.3 ms for *die*, an 132.2 ms for *das* (maximum difference = 0.5 ms). Auditory stimuli for the experiment were then be created by concatenating the imperatives, determiners, the adjective, and the nouns into phrases.

3.3. Procedure

To ensure familiarity with the images and the nouns, participants first performed a picture naming task in which they were asked to provide the name for each depicted object along with a gender-marked determiner. The task had two phases. In the first phase, images were shown one at a time, and remained on screen for either 10 seconds or until the participant had made their response and pressed a button to proceed. Next, the noun was printed underneath the picture, providing participants feedback about whether they used the expected word. If the expected word was not provided, participants were asked to produce the written word with its gender-marked article before proceeding to the next image. Once each image had been named in this manner, participants then named each object once more, this time without any written feedback. Responses in the picture naming task were recorded on a hand-held recording device, and were used to assess gender stability and knowledge of the nouns used in the primary task.

Figure 11 presents a schematization of the procedure for the lateralized picture monitoring task. For this task, participants were shown a fixation cross and two images at a time from the 480 pairs. One image was presented to the left of the fixation cross, and the other to the right, each at 2° of eccentricity from the center of the screen. Each image was surrounded by a colored frame of equal luminance. Colored frames were either be red, blue, green or violet. Each trial began with a fixation cross, followed 300 ms later by the appearance of the images and colored frames. The fixation cross, images and frames remained on screen for the duration of the trial. 300 ms after the appearance of the images, a phrase was presented over headphones which contained one of the depicted nouns. Concurrent with the acoustic onset of the noun, the color of the frame around the image depicting the target noun changed for 200 ms, after which a blank screen appeared, lasting until the acoustic offset of the noun. This was followed by a 500 ms blank screen and then a question asking participants to provide either the first or second color that framed the target image. Participants were asked to maintain a central fixation for the duration of the trial sequence until the question appears on screen. Each noun occurred equally often as a target and as a distractor. Nouns also occurred equally often on the left and right sides of the screen as a target and distractor. Target status and left/right position within noun-pairs were counterbalanced across experimental lists.

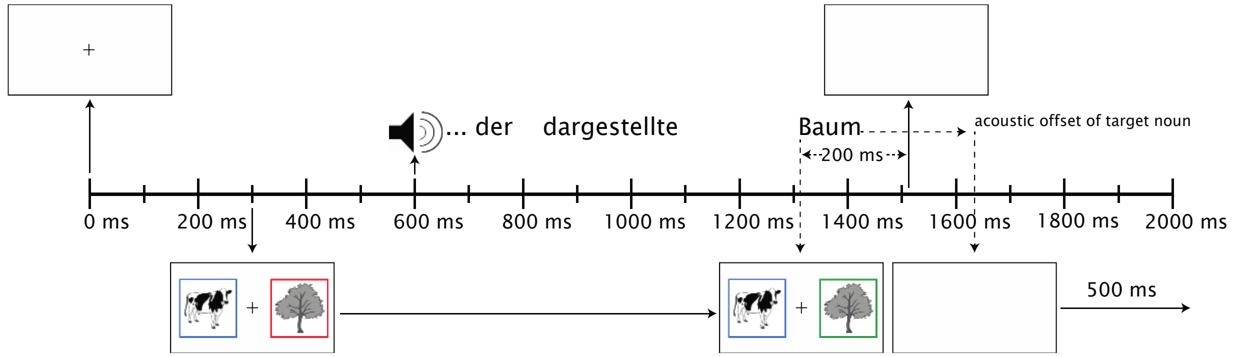

Figure 11. Schematization of the trial procedure for the lateralized picture monitoring task.

Prior to the main tasks, participants were given two blocks of practice with twelve additional objects to train them on maintaining a central fixation. During the practice, the EOG was monitored to ensure successful maintenance of fixation, and participants were given feedback on their performance. Participants were allowed to repeat each block of practice as many times as they wanted until they felt that they could maintain a steady fixation on the fixation cross. Participants who had difficulty with this at first were asked to repeat the practice until they could maintain fixation. All participants were successful at fixating the fixation cross for the duration of a trial by the end of the practice.

After the lateralized picture monitoring task, participants completed an EOG calibration similar to the one used by Wlotko and Federmeier (2007). For this task, trials began with a fixation cross that participants were instructed to fixate, after which an ‘x’ would appear to the left or right of the fixation cross at 1, 2 or 4 degrees of visual angle. Participants were instructed to shift their gaze to the ‘x’ when it appeared and to fixate it. These data were used to select participant-specific thresholds for rejecting trials contaminated by saccadic artifacts in the main task. Following the EOG calibration, participants performed a gender-decision task on all 60 nouns used in the main task. Nouns were presented individually in their written form in a randomized order over two repetitions. A noun was not repeated until all nouns had been presented once. Participants responded by pressing one of three pre-specified buttons on a gamepad to indicate if the gender of a word was feminine, masculine, or neuter. Each trial began with a fixation cross in the center of the screen lasting for 300 ms, followed by a blank screen of 200 ms and then the word, which remained on screen until a response was made. After participants made their response, there was a blank screen with a variable interstimulus interval between 1500 and 1800 ms before the next trial. Data from this task were used to assess gender stability, and to examine whether the speed of gender retrieval correlates with gender stability.

3.4. Data acquisition and analysis

Continuous EEG were recorded from 37 tin scalp electrodes mounted in an elastic cap (Electro-cap International), in accordance with the extended 10-20 system (Jasper, 1958). The locations of the scalp electrodes were FP1, FP2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P5, P3, P1, Pz, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2. Additional electrodes were placed below the left eye and at the outer canthus of each eye to monitor eye movements. EEG was recorded online with a 0.016–250 Hz bandpass filter and digitized at a 1000 Hz sampling rate, and filtered offline with a 0.1-30 Hz bandpass filter. The EEG was amplified with a BrainAmpDc bioamplifier system (Brain Products, Gilching, Germany). Impedances were held below 10 kΩ at all electrode sites. Data were referenced online to the left mastoid and re-referenced offline to the average of activity recorded over the left and right mastoids.

The EEG was processed offline with EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes. Artifact rejection was carried out in multiple steps to ensure that any N2pc effects were not contaminated by artifacts related to saccades. First, using the data from each participant's EOG calibration, EEGLAB's step-like artifact detection algorithm was run repeatedly with different thresholds until a threshold was found that reliably flagged all trials containing saccades greater than 1 degree of visual angle for removal. Next, these participant-specific thresholds were used to run the step-like artifact detection algorithm on each participant's EEG from the lateralized picture monitoring task to identify and remove any trials containing saccades greater than 1 degree of visual angle. The rationale behind this is that saccades greater than 1 degree of visual angle likely mean the participant was fixating one of the images. If that image was the target image, then participants would be able to direct overt visual attention to that image, given that it was fixated, eliminating the possibility of any N2pc being elicited. In contrast, saccades smaller than 1 degree of visual angle would not have been large enough to fixate one of the images, and thus it would still be possible to direct covert visual attention toward an image, even if the saccade resulted in a fixation point closer to that image than the fixation cross. While it might arguably be better to have employed a stricter threshold for rejecting trials contaminated by saccades, a balance ultimately had to be struck between rejecting trials on which N2pcs would be substantially reduced or absent due to the location of a fixation, and avoiding rejecting trials on which saccades were small enough that the fixation cross was still, effectively fixated, or on which high frequency noise rather than true saccades caused the trial to be flagged for rejection. At this point in the artifact rejection process, trials were also removed if they contained blinks that occurred within 200 ms of the onset of the noun, which was when the color of the target frame changed.

Given that trials containing small saccades were preserved, it is important to consider that saccades toward a target are much more likely on informative than uninformative trials, which potentially creates an imbalance in recorded EEG activity across these two conditions due to ocular activity. To correct for this imbalance, independent component analysis (ICA) was used to isolate and remove EEG activity related to eye movements from the scalp electrodes. ICA was run on the data prior to artifact rejection, in order to maximize the likelihood of achieving good source separation. Components were identified on the basis of their time course, scalp topography, and component spectra. Following ICA, the artifact rejection steps described above were carried out using the VEOG and HEOG, which were left out of the ICA, to identify blinks occurring within 200 ms of the onset of the noun and saccades larger than 1 degree of visual angle. Trials were also screened to remove epochs with excessive drift, muscle activity, or other artifacts that remained after ICA.

ERPs were calculated over a 1600-ms window time-locked to the acoustic onset of the determiner for each participant for each condition over each electrode, relative to a 200-ms prestimulus baseline. ERP analysis focused on the N2pc, which was computed at electrodes PO7 and PO8 by subtracting waveforms ipsilateral to a target noun from waveforms contralateral to the target noun. To assess whether attention had been covertly directed toward the target image prior to the onset of the noun, N2pcs were quantified as the mean amplitude for each condition in the 400 ms preceding the noun onset (440-840 ms). This time window was selected based on visual inspection of the waveforms, which showed apparent N2pcs starting to consistently deviate negatively from zero around this time.

For the L2 data, trials were classified as having stable or unstable gender assignments. Whereas previous studies sorting trials by gender assignments have done so on the basis of single measures with single repetitions of a noun (Hopp, 2013, 2016; Lemhöfer et al., 2014), this experiment used a composite measure of gender stability across multiple repetitions of nouns in the two gender assignment tasks described above in order to more directly test the extent to which variability in gender assignment at the item level impacts anticipatory processing. Gender representations were considered stable if participants assigned the correct gender to a noun on at least three of the four repetitions of each noun across the two gender assignment task. In all other cases, gender representations were considered unstable. In order to analyze the data from the primary task, trials were classified as stable trials if the gender representations for the target and distractor noun were both stable; otherwise trials were classified as unstable.

All data were analyzed using Bayesian mixed models with the *brms* package (Bürkner, 2017) in R. For the ERP data, separate models were fit to the L1 and L2 data. The dependent variable in each model was the mean amplitude in each condition for each participant from 440-840 ms after the onset of the determiner. Both models were fit using a Gaussian distribution (i.e. normal). The L1 model contained fixed effects for the intercept and condition (informative or uninformative), as well as a random intercept for participant. Thus, this model was effectively a Bayesian implementation of a mixed ANOVA. The L2 model contained fixed effects for the intercept, condition and stability (stable or unstable), a random intercept for participant, and by participant random slopes for condition and stability. In both ERP models, the outcome of interest is the probability that the mean amplitude in a given condition is less than zero (i.e. negative), which would indicate that covert attention had been directed toward the target image.

For the behavioral data, two separate models were fit to the accuracy data for identifying the requested frame color from the main task: one for the native speakers and one for the non-native speakers. The model for L1 participants included fixed effects for the intercept and condition (informative or uninformative). The variance parameters were random intercepts for participant, target item and distractor item, as well as random slopes for condition by each of the random intercept terms. The model fit to the L2 data contained fixed effects terms for the intercept, the three-way interaction between condition, trial stability (stable or unstable) and proficiency as determined by the Goethe test, as well as all subordinate interactions and main effects for these terms. While the LGLH does not make any predictions about the role of proficiency, this is an important term to include in the models given that proficiency has been shown to correlate with gender knowledge (e.g. Hopp, 2013; White et al., 2004). Indeed, in the present study, participants' Goethe scores were found to positively correlate with the number of correct gender assignments made across the gender assignment tasks ($\rho = 0.26$, $p < 0.001$). As a consequence, higher proficiency participants are more likely to have trials classified as stable, which creates a confound between proficiency and trial stability. By including proficiency in interactions with the other parameters and as a main effect in the L2 model, this confound is controlled for statistically, helping to disentangle any potential effects of proficiency from gender stability on the anticipatory use of grammatical gender. In addition to the fixed effects described above, the L2 accuracy model also contained random intercepts for participant, target item and distractor item, by participant slopes for the interaction between condition and stability as well as the main effects of each, and by target and by distractor slopes for the condition x stability x proficiency interaction, its subordinate interactions and main effects. Both accuracy models were fit using a Bernoulli distribution.

The other behavioral model that was fit examined the L2 response times as a function of gender stability. Gender stability was quantified using the unalikeability coefficient (see Kader & Perry, 2007 for a very accessible overview) in order to calculate the variability in gender assignments for each item on a per-subject basis. Briefly, unalikeability is a measure of variability in categorical data. It yields a coefficient between 0 and 1 which quantifies the extent to which observations in a dataset are different from one another. The more differences there are (i.e. the more variability there is), the greater the coefficient. Hence, nouns with low unalikeability coefficients are considered to have high gender stability, and nouns with high coefficients are considered gender-unstable. Because there was no a priori reason to expect that unalikeability should have a linear relationship with response times, and because this variable only contained five different values, it was coded as an ordinal factor to assess possible non-linear effects. The model fit to these data include fixed effects terms for the intercept, the interaction between unalikeability and proficiency as well as their main effects. Each parameter involving unalikeability included a linear, quadratic, cubic and quartic term. Variance terms included random intercepts for participant with by participant slopes for unalikeability. There was also a random intercept for item with by item slopes for the interaction between unalikeability and proficiency, as well as their main effects. This model was fit using an exponentially modified Gaussian distribution.

All models were fit using the default, weakly informative priors used by *brms*. Categorical predictors were sum coded, and continuous variables were centered and scaled. Each model was run with 4 chains. Each chain consisted of a total of 2000 iterations, 1000 of which were warm-up sampling. For the models fit to the ERP data, the priors for the fixed effects, standard deviations and variance parameters were set to Student-*t* distribution with a center at 0, a scale of 10 and 3 degrees of freedom. For the logistic models, priors for the fixed effects and standard deviations were set to a Student-*t* distribution with a center at 0, a scale of 10 and 3 degrees of freedom. The correlation parameters in all models had an *LKJ*(1) prior. For the model fit to the reaction time data, priors for the intercepts were set to Student-*t* distributions with 3 degrees of freedom, centers located at the median latency value for each dataset, and scales equal to the standard deviation of each dataset. Priors for the fixed effects parameters were set to a gamma distribution with a shape parameter of 1 and a scale parameter of 0.1. Finally, in the model for the reaction time data, the variance parameter had a prior with a Student-*t* distribution with 3 degrees of freedom, a center of 0, and a scale equal to the standard deviation of each dataset.

Convergence of the models was assessed by ensuring that there were no divergent transitions post-warmup, by examining the traceplots for good mixing, and by ensuring that the Gelman-Rubin *Rhat* statistic was below the recommended 1.1 (Gelman & Rubin, 1992). No models gave warnings about divergent transitions, all traceplots showed good mixing, and all *Rhat* values were below 1.1. Thus, all models were deemed to have converged.

For discussing the results, the same descriptions will be used as in Chapter 2. That is, when the credible interval for an effect does not overlap with zero, the evidence for that effect will be described as strong. When credible intervals do contain zero, if the posterior probability for the effect is greater than 90%, this will be described as moderate evidence; if the posterior probability is between 66% and 90%, this will be described as weak evidence. Posterior probabilities between 33% and 66% will be considered evidence for the null hypothesis, and, when relevant, posterior probabilities lower than 33% will be considered evidence for the opposite effect from the one being tested.

3.5. Hypotheses

For the native speakers, because prior work has shown that the N2pc can emerge in anticipation of a target (Woodman et al., 2009), and because the task requires that attention be directed to the target image in order to apprehend the color change, I expect that N2pcs will develop prior to the onset of the noun on informative trials, but not for uninformative trials. This would show that gender cues can be used predictively to direct covert visual attention. In addition, when asked to report the color of the target image, native speakers should be more accurate on informative trials than on uninformative trials, as gender-marking would allow them to direct their attention to the target image earlier, making it easier for them to apprehend the color change, in addition to which gender-marking would inform them that they can stop maintaining the color of the distractor in working memory, which would reduce memory load.

For the L2 speakers, the LGLH makes clear predictions about the use of gender-based anticipation in L2 users. Based on these, I expect to observe an interaction between condition (informative or uninformative) and gender stability (stable or unstable). Pairwise comparisons should show little to no evidence of gender-based anticipation in L2 users on unstable trials (Grüter et al., 2012; Lew-Williams & Fernald, 2010). Thus, I expect that N2pc amplitudes will not reliably differ from zero on informative or uninformative trials prior to the onset of the noun, or that their amplitude on unstable informative trials will be smaller than on stable informative trials. This follows from the LGLH's claim about the role of learning context in L2 gender-based anticipation; hence the classroom L2 learners in this study should have weaker links between nouns and their respective gender representations, resulting in a reduction in the use of gender marking as an anticipatory cue. Second, when comparing trials for which the nouns have stable gender representations for the correct gender, the LGLH predicts that gender marking will serve as a robust anticipatory cue (Hopp, 2013, 2016). In this case, N2pc effects should parallel those seen in the L1 participants. For trials with unstable gender representations (i.e. those for which participants did not consistently assign the same gender in the gender assignment tasks), the LGLH predicts little or no evidence of anticipation. Alternatively, if robust anticipation is found for gender-unstable nouns, it would provide an important caveat to the LGLH's proposal that the stability of gender representations is the culprit for L2 users' reduced ability to use gender marking predictively.

Based on the LGLH, I further expect that the accuracy of L2 speakers in reporting the color of the target image should parallel the results for the N2pc. In other words, if gender assignment variability reduces or eliminates the anticipatory use of gender marking, L2 participants should be approximately equally accurate in reporting the color of the target image for informative and uninformative trials when gender assignments are unstable. On trials with stable gender assignments, however, accuracy in reporting the color of the target image should be greater on informative trials than on uninformative trials, as participants should show robust use of gender marking as an anticipatory cue, and therefore direct their attention to target nouns earlier on informative trials.

Finally, testing Grüter et al.'s (2012) claim about the consequences of weaker links, if delayed retrieval of gender information is responsible for gender assignment errors, I should observe slower response latencies as gender stability decreases. If, on the other hand, gender assignment errors do not specifically reflect delayed retrieval of grammatical gender, we should find no relationship between gender decision latencies and gender stability.

3.6. Results

3.6.1. Lateralized picture monitoring task

3.6.1.1. Behavioral results

The accuracy for selecting the correct color in the main task is summarized in Table 22 for native speakers in each condition and for L2 speakers in each condition by gender stability. As can be seen, both groups of participants were overall highly accurate in making their responses, and both groups showed greater accuracy on informative compared to uninformative trials. For the L2 speakers, moreover, the effect of informativity on accuracy is larger on stable compared to unstable trials.

Table 22. Summary of accuracy data in the lateralized picture monitoring task by language group, condition and stability.

Language Group	Stability	Condition	Mean (%)	SD (%)
L1		Informative	88.3	8.0
		Uninformative	87.3	9.1
L2	Stable	Informative	90.2	6.6
		Uninformative	88.0	8.0
	Unstable	Informative	89.2	6.7
		Uninformative	88.3	8.1

Results of the accuracy model fit to the L1 data are summarized in Table 23. This model finds weak evidence that L1 participants are more accurate at selecting the correct color on informative trials compared to uninformative trials. Table 24 summarizes the accuracy model fit to the L2 data. This model finds moderate evidence for a main effect of condition, such that accuracy is greater on informative than on uninformative trials. The model further finds weak evidence that the effect of condition is qualified by an interaction with stability. The posterior group means for this interaction are shown in Figure 12, which shows a larger effect of condition on stable trials compared to unstable trials. Table 25 summarizes the mean posterior group differences for these effects. These show a 95.8% chance that L2 participants responded more accurately on stable informative trials than on stable uninformative trials, compared to a 74.2% chance that responses were more accurate on unstable informative trials than on unstable uninformative trials.

Table 23. Summary of model results for L1 accuracy data in the lateralize picture monitoring task.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	2.27	0.30	1.70	2.88	1
Condition (Informative)	0.05	0.06	-0.08	0.17	0.784

Table 24. Summary of model results for the L2 accuracy data in the lateralized picture monitoring task.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	2.33	0.11	2.11	2.54	1
Condition (Informative)	0.07	0.05	-0.02	0.16	0.940
Stability (Stable)	0.02	0.04	-0.05	0.09	0.727
Proficiency	0.04	0.15	-0.26	0.32	0.597
Condition (Informative) x Stability (Stable)	0.03	0.04	-0.04	0.11	0.807
Condition (Informative) x Proficiency	-0.04	0.04	-0.11	0.04	0.824
Stability (Stable) x Proficiency	-0.01	0.04	-0.09	0.08	0.562
Condition (Informative) x Stability (Stable) x Proficiency	-0.04	0.04	-0.11	0.04	0.848

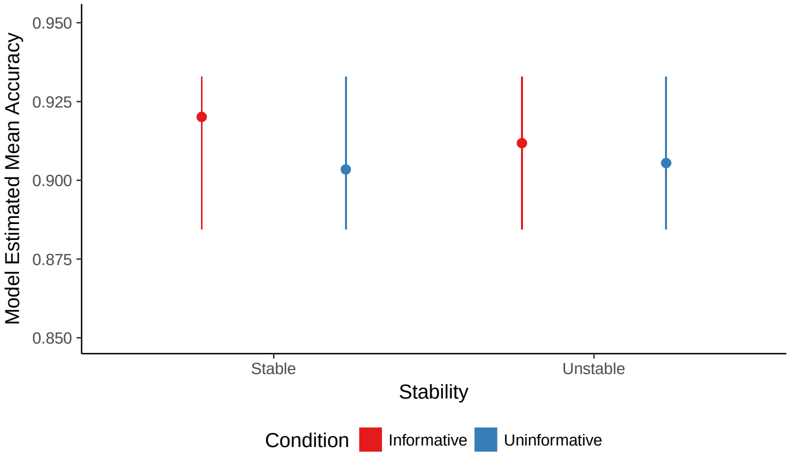

Figure 12. Model estimated posterior group means for condition by stability for L2 accuracy in the lateralized picture monitoring task.

Table 25. Model estimated posterior group differences for condition by stability for the L2 accuracy data in the lateralized picture monitoring task.

Stability	Contrast	Mean	SD	2.5%	97.5%	P > 0
Stable	Informative - Uninformative	0.21	0.12	-0.03	0.44	0.958
Unstable	Informative - Uninformative	0.08	0.12	-0.15	0.31	0.742

The model for these data also finds weak evidence that the interaction between condition and stability is affected by proficiency. This interaction, visualized in Figure 13, appears to be driven by lower proficiency participants, who exhibit larger gains in accuracy on informative trials when gender assignments are stable. In contrast, higher proficiency participants do not appear to show much of an effect of condition on stable trials. For unstable trials, accuracy is uniformly higher on informative than on uninformative trials, regardless of proficiency.

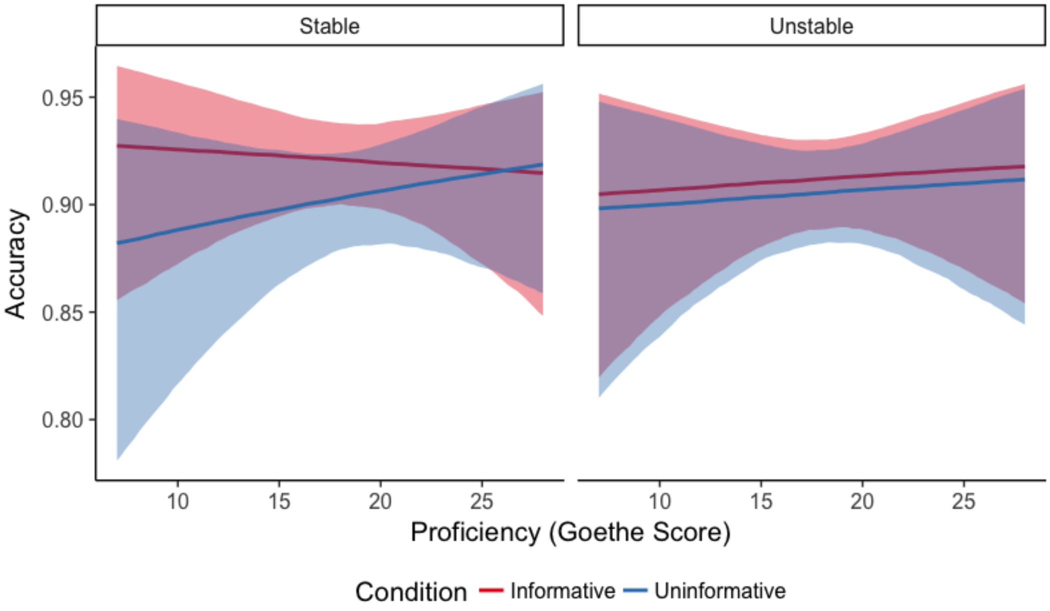

Figure 13. Model estimated marginal effects for condition by stability by proficiency for L2 accuracy in the lateralized picture monitoring task.

3.6.1.2. ERP results

Figure 14 shows the N2pcs on informative and uninformative trials for the L1 participants time-locked to the acoustic onset of the determiner. Visual inspection of the waveforms indicates that the N2pc does not appear to deviate consistently from zero on uninformative trails until after the onset of the noun, at which point the color change elicits large visual components. On informative trials, the waveform starts to show a sustained negative deflection beginning around 400 ms and extending until the onset of the noun. Results of the model fit to these data are summarized in Table 26. Table 27 summarizes the posterior group means for each condition and provides the probability that the mean amplitude in each condition is negative. These results show a 96.5% chance that the N2pc is negative on informative trials, with an estimated effect size of 0.2 µV. On uninformative trials, in contrast, the probability that the N2pc is negative is 65.8%, which is most consistent with a null effect.

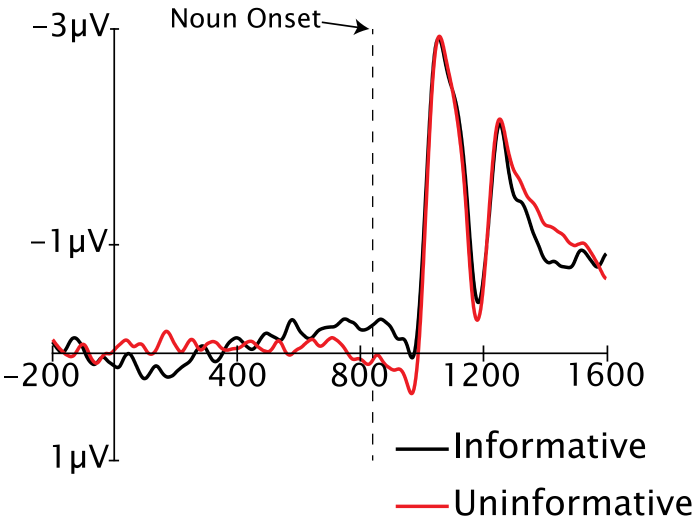

Figure 14. N2pc for informative and uninformative trials for L1 participants. Waveforms are time-locked to the acoustic onset of the determiner. The dashed vertical line represents the acoustic onset of the target nouns at 840 ms. Negative is plotted up. Waveforms were low-pass filtered at 12 Hz for visualization.

Table 26. Summary of the model fit to the mean amplitudes of the L1 N2pc data from 440-840 ms after the onset of the determiner.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-0.12	0.09	-0.29	0.05	0.914
Condition (Informative)	-0.8	0.06	-0.20	0.05	0.901

Table 27. Summary of the posterior group means for each condition and the probabilities that each effect has a negative value.

Condition	Mean	SD	2.5%	97.5%	P < 0
Informative	-0.20	0.11	-0.42	0.02	0.965
Uninformative	-0.04	0.11	-0.25	0.17	0.658

The N2pcs for L2 participants are shown in Figure 15 for each condition on stable and unstable trials. Visual inspection suggests that uninformative trials do not differ consistently from zero until after the onset of the noun, regardless of stability. Informative trials appear to exhibit reliable negative excursions that are sustained until the onset of the noun on both stable and unstable trials. This negativity appears to begin slightly earlier on stable trials, however the negativity seems larger on unstable trials. Results of the model fit to these data are summarized in Table 28. These results find moderate evidence for an effect of condition, but no evidence that stability mediates this effect. That said, to assess whether or not the N2pc has onset, it is necessary to evaluate whether the mean amplitude of the N2pc in a given condition is more negative than zero, not whether it differs from the mean amplitude of the N2pc in another condition. Table 29 provides the posterior group means for each combination of condition and stability, and the posterior probability that each effect is more negative than zero. These estimates show that the data in the uninformative conditions are most consistent with null effects. On stable informative trials, the model shows weak evidence that the N2pc is more negative than zero. In addition, the model finds strong evidence that the N2pc is negative on unstable informative trials, with an effect size more than double that found for informative stable trials. Finally, the posterior probability that the N2pc on informative stable trials is more negative than the N2pc on informative unstable trials is 18.7% (*M* = 0.24, *SD* = 0.27, *CI*: -0.29, 0.77), indicating that it is much more likely that N2pcs were more negative on informative unstable trials. Thus, the ERP data do not appear consistent with the LGLH.

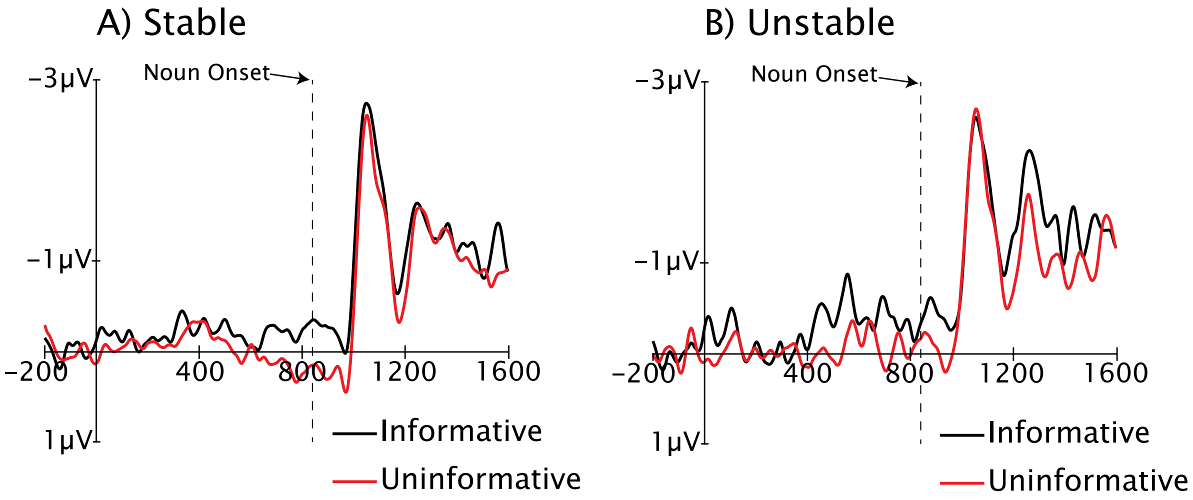

Figure 15. N2pc for informative and uninformative trials for L2 participants when gender assignments are stable (Panel A) and unstable (Panel B). Waveforms are time-locked to the acoustic onset of the determiner. The dashed vertical line represents the acoustic onset of the target nouns at 840 ms. Negative is plotted up. Waveforms were low-pass filtered at 12 Hz for visualization.

Table 28. Summary of the model fit to the mean amplitudes of the L2 N2pc data from 440-840 ms after the onset of the determiner.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	-0.17	0.10	-0.36	0.03	0.952
Condition (Informative)	-0.16	0.12	-0.39	0.07	0.918
Stability (Stable)	0.09	0.10	-0.10	0.28	0.834
Condition (Informative) x Stability (Stable)	0.02	0.10	-0.16	0.21	0.592

Table 29. Summary of the posterior group means for each condition by stability and the probabilities that each effect has a negative value.

Stability	Condition	Mean	SD	2.5%	97.5%	P < 0
Stable	Informative	-0.21	0.20	-0.61	0.19	0.847
	Uninformative	0.06	0.20	-0.33	0.48	0.390
Unstable	Informative	-0.45	0.20	-0.85	-0.04	0.985
	Uninformative	-0.08	0.22	-0.50	0.34	0.639

3.6.2. Gender decision task

Response times in the gender decision task are shown in Figure 16 for the L2 participants by unalikeability. Overall, there is a clear increase in response times as gender assignment variability increases. Results of the model fit to these data are summarized in Table 30.

Consistent with the raw data, this model finds strong evidence for a linear effect of unalikeability, such that response times become longer as unalikeability increases. There is weak evidence that the effect of gender assignment variability may be non-linear, however this is not explored any further given that the evidence for a linear effect is so strong. The size of the main effect of unalikeability estimated by the model is strikingly large, however it is important not to interpret this effect in isolation, given that the interaction with proficiency was modeled. There is, moreover, weak evidence that the effect of unalikability is mediated by proficiency. This interaction is shown in Figure 17. As can be seen in this figure, when gender assignment variability is low-to-moderate, response times appear to decrease as proficiency increases, and response times become longer as variability increases. When variability is high, however, the effect of proficiency reverses such that there is a very large increase in response times as proficiency increases. At the lowest proficiency levels, moreover, high variability nouns were responded to the fastest. These data are thus largely consistent with the claim of the LGLH that weaker links between nouns and their gender representations lead to slower gender retrieval.

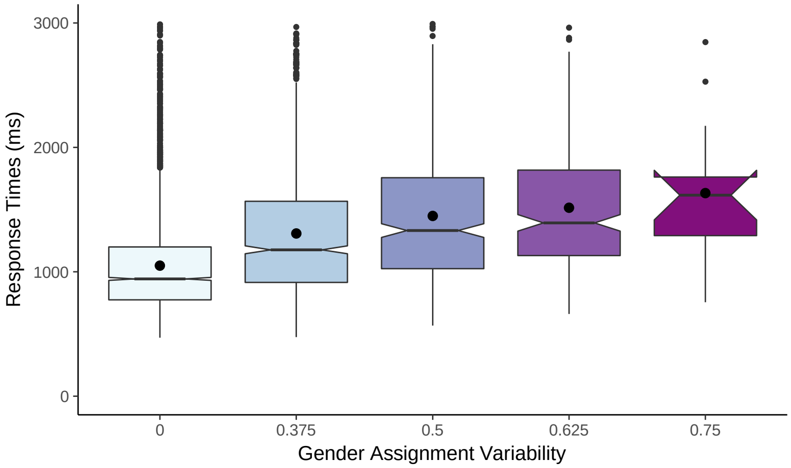

Figure 16. L2 response times in the gender decision task by accuracy and gender assignment variability as measured by unalikeability coefficients. Lower unalikeability coefficients correspond to lower variability. The dots contained within each boxplot represent the mean.

Table 30. Summary of model results for L2 response times in the gender decision task.

Fixed Effect	Mean	SD	2.5%	97.5%	P(sign)
Intercept	910.44	55.95	802.99	1014.53	1
Variability (Linear)	411.88	159.16	112.82	717.89	0.998
Variability (Quadratic)	115.60	133.54	-139.11	374.24	0.796
Variability (Cubic)	87.10	81.93	-72.58	246.80	0.851
Variability (Quartic)	43.00	36.99	-29.39	116.46	0.878
Proficiency	53.08	75.26	-73.78	215.24	0.736
Variability (Linear) x Proficiency	238.42	226.91	-133.17	722.60	0.855
Variability (Quadratic) x Proficiency	192.76	190.99	-117.05	604.17	0.834
Variability (Cubic) x Proficiency	92.05	114.57	-95.88	340.94	0.764
Variability (Quartic) x Proficiency	17.43	46.58	-61.78	116.00	0.616

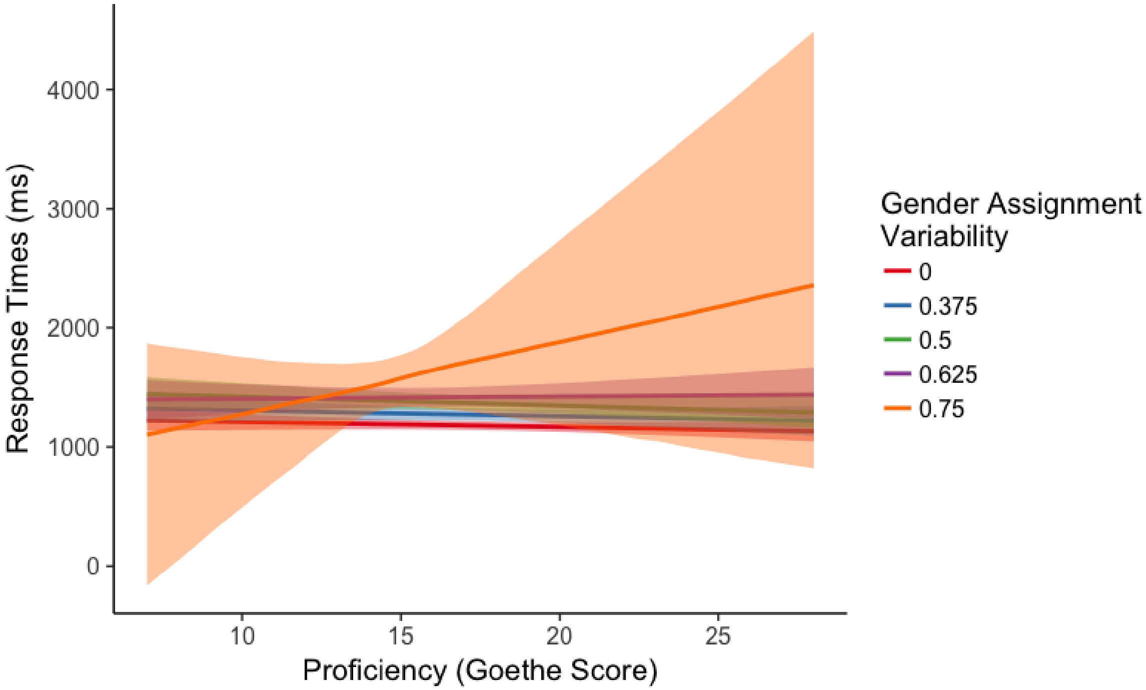

Figure 17. Model estimated marginal effects of proficiency by unalikeability on L2 response times in the gender decision task. Shading shows the 95% credible interval around each estimated line.

3.7. Discussion of experiment 3

3.7.1. Lateralized picture monitoring task

Experiment 3 used a lateralized picture monitoring task in combination with repeated measures of gender assignment in order to assess the extent to which variability in gender assignment modulates the anticipatory use of grammatical gender in German. Results found evidence that the native speakers were more accurate at selecting the appropriate frame color when trials provided informative gender information compared to when gender marking was not informative about where the color change would occur. Native speakers further showed evidence, as measured by the N2pc, that they used gender marking to direct covert visual attention to the target noun prior to the onset of the noun. While the evidence for anticipation in the native speakers was not strong, this was likely due to the fact that there are currently data from only 13 native speakers, and thus there may simply be too little power at present to find strong evidence for anticipatory effects. In addition, the high accuracy suggests that the task may have been too easy. In particular, the fact that participants were on average nearly 90% accurate on uninformative trials suggests that they had little difficulty with the task even when gender marking could not help them to direct their attention earlier to the location of the color change. Making the duration of the color change shorter might therefore help to encourage greater anticipatory use of grammatical gender if it becomes much more difficult to reliably detect and report the color change. Importantly, both the accuracy data and the ERP results are consistent with anticipatory use of grammatical gender. Thus, the task appears to have been successful in eliciting anticipatory behavior.

For the L2 speakers, the results present a somewhat contradictory picture. The accuracy data provide clear evidence that L2 participants were more accurate on informative trials, and that the size of this effect was larger on stable compared to unstable trials. These data are thus consistent with the claim of the LGLH that L2ers are more effective at using gender as an anticipatory cue when gender representations are stable. In the ERP data, however, while there is clear evidence that L2 participants used gender to direct their covert visual attention to the target noun in anticipation of the color change, this effect is actually larger on unstable trials than on stable trials. This apparent contradiction across measures could be partly explained by the fact that there are two distinct components involved in accurately performing the task: an attentional/perceptual component and a memory component. With respect to the former, accurate performance requires perceiving the color change in order to accurately report the second color of the frame around the target. To facilitate accurate perception of this color change, it is strategically beneficial for participants to utilize any cues that allow them to reliably deploy attention to where the color change will happen – this is what the N2pc measures. For the memory component, accurate performance requires that participants maintain all colors in working memory that may be required for the response. On uninformative trials, participants must maintain the color around the target, the color around the distractor, and the information about which color corresponds to which noun for the entire duration of the trial until the noun onset. In contrast, on informative trials, participants must only maintain this information until the determiner, at which point they can selectively forget the information about the distractor in order to free up memory resources, reducing cognitive load. Whereas the N2pc measures the deployment of attention, the accuracy data measure both the attentional/perceptual component of the task and the memory component of the task. The discrepancy across measures with respect to the role of gender stability may therefore reflect differences in how gender stability affects the use of gender information to perform each of these subcomponents. The N2pc data suggest that gender stability may not play a critical role in the ability to use gender information to predictively deploy attentional resources. This might further suggest that the bottom-up use of grammatical gender as an anticipatory cue does not depend on gender stability, though the nature of this task and the somewhat long duration between the onset of the article and the onset of the noun make it likely that top-down information was at least partially involved in how attention was deployed. In the case of accuracy, these data may indicate that gender stability plays an important role in controlled, top-down processes. Future work will be necessary to assess this possibility.

While the above explanation offers an appealing account of the discrepancy between the N2pc data and the accuracy data with respect to the effects of gender stability, it does not explain why N2pc amplitudes were more negative for unstable informative trials compared to stable informative trials. One possibility is that these results are related to learning that occurred over the course of the lateralized picture monitoring task. Given that nouns were repeated multiple times, it is likely that participants' gender knowledge improved over the course of the experiment. This could potentially explain the N2pc results in two ways: first, recall that the naming task was performed prior to the lateralized picture monitoring task, whereas the second gender assignment task was performed *after* the lateralized picture monitoring task. This may have led to an underestimation of gender stability for some items if participants assigned the incorrect genders on the naming task, but after the main task had accurately learned the genders of those nouns and correctly assigned gender to those items on the gender decision task. That is, if participants correctly assigned grammatical gender to an item on two out of four iterations of the gender assignment tasks, but if the correct assignments both occurred after the lateralized picture monitoring task, this item would be classified as unstable, but it may in fact have a stable representation due to learning over the course of the main task. Thus, trials classified as unstable may in fact often actually have stable representations. Second, if learning occurred during the lateralized picture monitoring task, we should expect it to be most prominent for nouns that would be classified as unstable, as genders of the stable nouns were presumably already well-known. This may have led participants to allocate greater attentional resources to target nouns whose gender they were less confident about, and/or participants on unstable trials may have directed their attention to the image of the noun rather than just to the frame, which would result in a larger N2pc due to the image being further from the fixation cross than the frame.

Alternatively, the competition resolution hypothesis of the N2pc (cf. Luck, 2012, p. 356) proposes that the N2pc reflects a process by which the neural representation of a target is enhanced by minimizing interference from distractor items. This hypothesis predicts larger N2pc amplitudes when more effort is needed to resolve competition between a target and any distractors. It may therefore be that unstable informative trials elicit a larger N2pc amplitude because it is more difficult for participants to suppress interference from the distractor noun, perhaps because they are less certain about the gender of each noun, even when they are able to predictively direct attention to the target. More work will be needed to assess the extent to which competition resolution vs learning might explain the larger N2pc amplitudes on unstable informative trials.

Finally, the accuracy data also interacted with proficiency. On unstable trials, accuracy was greater on informative trials, and accuracy increased steadily as proficiency increased. On stable trials, accuracy increased with proficiency on uninformative trials, but actually decreased with proficiency on informative trials. This decrease appears not to be due to higher proficiency participants being worse at using gender as an anticipatory cue, but rather due to lower proficiency participants benefiting much more from using gender predictively. Given the high accuracy overall, it is possible that this reflects a ceiling effect whereby higher proficiency participants are already so accurate that their performance cannot benefit substantially from using gender as an anticipatory cue, whereas the lower accuracy on uninformative trials for the lower proficiency participants leaves more room for gender marking to improve accuracy. While this explanation would predict a flat slope on stable informative trials, it is important to note that there were very few participants with proficiency scores at the extremes, and thus the slope of this line may be somewhat misleading. More data would be necessary to evaluate this explanation. With the exception of the stable informative condition, the general increase in accuracy as proficiency increases is likely due to the fact that more proficient participants also likely have more experience with German and are more efficient at processing German. Thus, higher proficiency participants may have been under lower cognitive load and may have been more efficient at using phonological information from the noun onset to rapidly shift attention to the target noun and suppress information from the distractor.

Summarizing, the data from the lateralized picture monitoring task showed clear evidence that advanced L2 speakers of German can effectively use grammatical gender as an anticipatory cue. The accuracy data are consistent with the LGLH in suggesting that the anticipatory use of grammatical gender in an L2 is more effective when nouns have stable gender representations. The ERP data, in contrast, are less clear in that they show evidence for an effect of gender stability, but not in the direction that was initially predicted. Future work will be necessary to examine the extent to which these ERP results may reflect learning or competition resolution versus being incompatible with the LGLH.

3.7.2. Gender decision task

Data from the gender decision task were used to test the claim of the LGLH that weaker links between nouns and their gender representations lead to delayed retrieval of gender information, which results in increased variability in gender assignments. Results from these data found strong evidence that response latencies in the gender decision task increased linearly as gender assignment variability increased. This is consistent with what the LGLH predicts, and thus lends plausibility to the claim that weaker links play a role in gender assignment variability. That said, because the strength of associations between nouns and their gender representations were not manipulated directly, these data cannot be used to make any causal claims about the nature of this effect. In addition, the response latencies in this task reflect the culmination multiple processes of which gender retrieval is only one. Consequently, these data cannot tell us whether the locus of the observed delays might reflect slower retrieval of grammatical gender, post-retrieval processes, or some combination of these.

Results from this task also found that gender assignment variability interacted with proficiency, with the most proficient participants showing large increases in response latencies for highly variable items, whereas the lowest proficiency participants were actually faster at responding to highly variable nouns. It is possible that this effect reflects different strategies adopted by higher and lower proficiency participants when faced with uncertainty about a noun’s gender. For the lower proficiency participants, when they are highly uncertain about the gender of a noun, they may simply give up and resort to quickly guessing at the gender of the noun. Higher proficiency participants, in contrast, may be more willing to invest extra time in attempting to accurately determine the gender of nouns. That said, the estimated effect of proficiency on response times at the highest level of gender assignment variability should be interpreted with caution at this point, as there are relatively few data points with unalikeability scores of 0.75.

In short, results from the gender decision task are consistent with the claim of the LGLH that weaker links lead to slower retrieval of gender information and greater variability in gender assignment. More work, however, is needed to understand the causal dynamics of these data and the locus of the observed delays.

CHAPTER 4: GENERAL DISCUSSION

Across three experiments, I tested the claims of the LGLH that 1) differences in L1 vs L2 learning contexts lead to weaker associations between nouns and their respective gender representations in an L2, 2) these weaker links lead to increased variability in gender assignment, and 3) this increased variability leads to a reduced or absent use of grammatical gender marking as an anticipatory cue. With respect to the first two claims, Experiments 1 and 2 did not find evidence consistent with the LGLH. Given that prior research has found effects of learning condition on the success of gender learning using this paradigm, I suggested that the lack of effects found in this dissertation may have been due to the difficulty of the learning task or to learning over the course of the visual world task, which may have eliminated group effects. Ongoing work is assessing the extent to which making learning easier by providing explicit information might affect learning outcomes and the predictive use of grammatical gender. In addition, it would be fruitful to attempt to replicate the original Arnon and Ramscar (2012) findings with the 24 items from Experiments 1 and 2 to examine whether the number of items did, in fact, make the task too challenging. If group effects can still be found with 24 items, this would suggest that learning during the visual world task is at least partly responsible for the lack of any effects of learning condition in the direction predicted by the LGLH.

Another approach to testing the first claim of the LGLH would be to use L1, L2 and heritage speakers of a natural language. If early language experience truly plays a role in the ability to use gender as an anticipatory cue in an L2, heritage speakers should then show greater anticipatory effects, at least for nouns learned early in life. Indeed, work by Montrul et al. (2014) suggests that this may be the case. Another important consideration with respect to language experience is that Grüter et al. (2012) found comparable anticipatory effects in L1 and L2 speakers of Spanish for novel Spanish nouns that were learned in full sentences. This finding raises questions about the extent to which early language experience with a gender-marking system is responsible for the discrepancies observed in L1 and L2 populations, and the extent to which individual experience with individual items might shape gender knowledge and the ability to use gender as an anticipatory cue to that specific noun. Heritage speakers of a gender-marking language are an ideal population with which to tease these possibilities apart, given that they have early language experience with a gender-marking system, but their experience with individual words may be more L1-like or L2-like depending on when and in what context they learned that word (e.g. under naturalistic conditions in early life or in an L2 classroom; cf. Montrul, 2015; Montrul, 2008).

Turning to the third claim of the LGLH, that increased gender variability leads to a reduction in the use of gender marking as an anticipatory cue, the evidence found in this dissertation is somewhat mixed. Experiments 1 and 2 found clear evidence from the button press data that gender stability affects the anticipatory use of grammatical gender, however the eye movement data yielded weak evidence for this, at best. In Experiment 3, the accuracy data provided evidence consistent with the LGLH's claim, but the ERP data, while showing clear anticipatory effects, were ambiguous about whether stable gender assignments lead to greater use of gender as an anticipatory cue. One possibility suggested by these results is that gender stability does impact the predictive use of grammatical gender, but that its effects are limited to or most robust for later or more top-down processes. If so, we might expect to observe no or very small effects of gender stability in gender priming tasks where top-down predictions or decisions are not necessary for accurate performance.

It was also suggested, however, that the lack of clear effects of gender stability prior to the target nouns in these experiments might reflect learning that occurred over the course of each experiment. This could be straightforwardly assessed by using designs in which nouns are only used once as a target or distractor. If the lack of stability effects is, in fact, due to learning, we should expect robust effects of stability to emerge when items are only repeated once, thus eliminating the possibility for learning to occur over multiple repetitions of a noun.

In addition, it is important to consider how well gender stability was actually measured in these experiments. Though the use of multiple repetitions makes this unlikely, it is certainly possible that some trials classified as stable were the product of pure guessing behavior. Without knowing how participants were making their decisions, it is impossible to know the extent to which stability may have been overestimated due to this. One way of tapping into this might be to combine gender assignment performance with ratings of how confident participants were in their gender assignments to yield a composite measure of gender assignment variability and confidence. If participants were guessing, we should expect lower confidence ratings compared when gender assignments reflect the use of prior experience and knowledge. The confidence ratings available from Experiments 1 and 2 did not reveal any differences across learning conditions, and hence were not discussed due to not being central to testing the LGLH. If, however, combining these ratings with gender assignment variability does indeed provide a better measure of gender stability, we might expect to find stronger effects of stability in the data from these experiments. Analyses are planned to examine this possibility.

Confidence ratings were not collected for Experiment 3 due to time constraints, however it may be particularly important for future work to include such ratings in measures of gender stability, given that probabilistic cues to gender class can be used to produce stable gender assignments without necessarily entailing that participants have directly associated a noun with its appropriate gender representation. If the predictive use of grammatical gender depends primarily or solely on a direct association between gender nodes and a noun rather than a form-based route (cf. Gollan & Frost, 2001), any noise introduced to the gender stability measure as a result of the use of probabilistic information could lead to an underestimation of the effects of gender stability on the anticipatory use of grammatical gender. In addition, to whatever extent the measure of gender stability used in these experiments was imprecise, any noise in this measure would be compounded in the analyses due to the fact that trials were classified as stable or unstable based on assignments for both the target and the distractor noun on each trial. It would therefore be useful for future work to also examine the predictive use of grammatical gender in contexts where gender stability on a given trial can be determined based solely on a target noun.

Finally, data from Experiment 3 were used to assess the plausibility of a corollary to the second claim of the LGLH, that weaker links between nouns and their gender representations lead to slower retrieval of grammatical gender, which in turn leads to greater variability in gender assignment. Consistent with this claim, Experiment 3 found that a positive relationship between response times for gender assignment and the variability a participant exhibited in assigning gender to that noun. As was discussed in Chapter 3, however, these data cannot determine the causal nature of this relationship. In addition, recent ERP work has found that late L2 speakers are able to retrieve grammatical gender information on a similar time course to native speakers, at least for highly familiar nouns, even when those same L2 speakers are slower at assigning gender to nouns in a separate gender decision task (Shantz & Tanner, 2016). It will therefore be important for future work to examine whether the slower response times for more variable nouns observed in Exp. 3 might reflect delayed retrieval, slower post-retrieval processes, or some combination of these. Understanding the nature of these delays may, moreover, lead to a better understanding of online deficits in L2 gender processing in comprehension.

CHAPTER 5: CONCLUSION

This dissertation used a combination of artificial language learning, eye tracking and event-related potentials to test the claims of the Lexical Gender Learning Hypothesis about why second language learners struggle to learn gender and to use gender as an anticipatory cue. The results from three experiments indicate that the stability of an individual's gender representations likely affects their use of gender marking as an anticipatory cue to that noun, though the data further suggest that these effects may be limited to later or more top-down processes. Results did not find any evidence that learning context affects the stability of gender representations nor the anticipatory use of gender information in the ways predicted by the LGLH.

Going forward, the work described in this dissertation raises numerous questions that future work should address in order to attain a comprehensive understanding the precise nature of the effects of gender stability. These include identifying the representational nature of gender stability, the time course of its effects on gender retrieval and on the anticipatory use of grammatical gender, as well as how gender representations come to have varying stability in the learning process.

Finally, the finding that gender stability impacts the predictive use of grammatical gender in an L2 has important methodological implications for future L2 research. Specifically, these data complement other recent work (Hopp, 2013; Lemhöfer et al., 2014) in demonstrating the critical importance of taking L2 knowledge into account when investigating L1-L2 processing differences, as failure to do so may result in an underestimation of L2 performance.

CHAPTER 6: REFERENCES

Alarcón, I. V. (2011). Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Bilingualism: Language and Cognition*, *14*(3), 332–350. https://doi.org/10.1017/S1366728910000222

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

Alemán Bañón, J., Fiorentino, R., & Gabriele, A. (2014). Morphosyntactic processing in advanced second language (L2) learners: An event-related potential investigation of the effects of L1-L2 similarity and structural distance. *Second Language Research, 30*(3), 275–306. https://doi.org/10.1177/0267658313515671

Alemán Bañón, J., Miller, D., & Rothman, J. (2017). Morphological variability in second language learners: An examination of electrophysiological and production data. *Journal of Experimental Psychology: Learning Memory and Cognition*.

Alemán Bañón, J., & Rothman, J. (2016). The role of morphological markedness in the processing of number and gender agreement in Spanish : an event-related potential investigation. *Language, Cognition and Neuroscience*, *31*(10), 1273–1298. https://doi.org/10.1080/23273798.2016.1218032

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*(4), 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582. https://doi.org/10.1111/j.0956-7976.2004.00723.x

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science*, 1–22.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305. https://doi.org/10.1016/j.cognition.2011.10.009

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.1287/mksc.12.4.395

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2015). Tracking lexical consolidation with ERPs: Lexical and semantic-priming effects on N400 and LPC responses to newly-learned words. *Neuropsychologia*, *79*, 33–41. https://doi.org/10.1016/j.neuropsychologia.2015.10.020

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

Barber, H. A., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish:

an ERP comparison. *Journal of Cognitive Neuroscience*, *17*(1), 137–153. https://doi.org/10.1162/0898929052880101

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), arXiv:1406.5823. https://doi.org/10.18637/jss.v067.i01

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

Bates, E. A., Devescovi, A., Pizzamiglio, L., D'amico, S., & Hernandez, A. E. (1995). Gender and lexical access in Italian. *Perception & Psychophysics, 57*(6), 847–862. https://doi.org/10.3758/BF03206800

Bates, E. A., Elman, J., & Li, P. (1994). Language in, on, and about time. In M. Haith, J. Benson, R. Roberts, & B. Pennington (Eds.), *The Development of Future-Oriented Processes* (pp. 293–321). Chicago: The University of Chicago Press.

Bialystok, E., Craik, F. I. M., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest*, *10*(3), 89–129. https://doi.org/10.1177/1529100610387084

Bölte, J., & Connine, C. M. (2004). Grammatical gender in spoken word recognition in German. *Perception & Psychophysics*, *66*(6), 1018–1032. https://doi.org/10.3758/BF03194992

Bordag, D., Opitz, A., & Pechmann, T. (2006). Gender processing in first and second languages: The role of noun termination. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(5), 1090–1101. https://doi.org/10.1037/0278-7393.32.5.1090

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

Bosker, H. R., Quené, H., Sanders, T., & de Jong, N. H. (2014). Native "um"s elicit prediction of low-frequency referents, but non-native "um"s do not. *Journal of Memory and Language*, *75*, 104–116. https://doi.org/10.1016/j.jml.2014.05.004

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*(C), 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*. Elsevier B.V. https://doi.org/10.1016/j.brainres.2012.01.055

Bruhn de Garavito, J., & White, L. (2002). The second language acquisition of Spanish DPs: The status of grammatical features. In A. T. Pérez-Leroux & J. M. Liceras (Eds.), *The Acquisition of Spanish Morphosyntax* (pp. 153–178). Springer Netherlands.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–90.

https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–11. https://doi.org/10.3758/s13428-013-0403-5

Bürkner, P.-C. (2017). brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Caffarra, S., & Barber, H. A. (2015). Does the ending matter? The role of gender-to-ending consistency in sentence reading. *Brain Research*, *1605*(1), 83–92. https://doi.org/10.1016/j.brainres.2015.02.018

Caffarra, S., Janssen, N., & Barber, H. A. (2014). Two sides of gender: ERP evidence for the presence of two routes during gender agreement processing. *Neuropsychologia*, *63*, 124–34. https://doi.org/10.1016/j.neuropsychologia.2014.08.016

Caffarra, S., Siyanova-Chanturia, A., Pesciarelli, F., Vespignani, F., & Cacciari, C. (2015). Is the noun ending a cue to grammatical gender processing? An ERP study on sentences in Italian. *Psychophysiology*, *52*, n/a-n/a. https://doi.org/10.1111/psyp.12429

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: evidence from the "tip-of-the-tongue" phenomenon. *Cognition*, *64*(3), 309–343. https://doi.org/10.1016/S0010-0277(97)00031-0

Carreiras, M., Garnham, A., & Oakhill, J. (1993). The use of superficial and meaning-based representations in interpreting pronouns: Evidence from Spanish. *European Journal of Cognitive Psychology*, *5*(1), 93–116. https://doi.org/10.1080/09541449308406516

Carroll, S. (1989). Second-language acquisition and the computational paradigm. *Language Learning*, *39*(4), 535–594. https://doi.org/10.1111/j.1467-1770.1989.tb00902.x

Chou, C. J., Huang, H. W., Lee, C. L., & Lee, C. Y. (2014). Effects of semantic constraint and cloze probability on Chinese classifier-noun agreement. *Journal of Neurolinguistics*, *31*, 42–54. https://doi.org/10.1016/j.jneuroling.2014.06.003

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Corbett, G. G. (1991). *Gender*. New York: Cambridge University Press.

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

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*(4), 465–480. https://doi.org/10.1006/jmla.1999.2688

Davidson, D. J., & Indefrey, P. (2009). An event-related potential study on changes of violation and error responses during morphosyntactic learning. *Journal of Cognitive Neuroscience*, *21*(3), 433–446. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/jocn.2008.21031

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, *121*(3), 226–239.

https://doi.org/10.1016/j.bandl.2012.02.006.Thinking

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*(1), 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-Processing in Sentence Comprehension: Sensitivity to Likely Upcoming Meaning and Structure. *Language and Linguistics Compass*, *8*(12), 631–645. https://doi.org/10.1111/lnc3.12093

DeLong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin & Nieuwland (2016). *Language, Cognition, and Neuroscience*. https://doi.org/10.1080/23273798.2017.1279339

Delorme, A., & Makeig, S. (2004). EEGLAB: an open sorce toolbox for analysis of single-trail EEG dynamics including independent component anlaysis. *Journal of Neuroscience Methods*, *134*, 9–21.

Demestre, J., & García-Albea, J. E. (2007). ERP evidence for the rapid assignment of an (appropriate) antecedent to PRO. *Cognitive Science*, *31*(2), 343–354. https://doi.org/10.1080/15326900701221512

Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew : Evidence from ERPs and eye movements. *Journal of Memory and Language*, *45*, 200–224. https://doi.org/10.1006/jmla.2000.2768

Dewaele, J.-M., & Véronique, D. (2001). Gender assignment and gender agreement in advanced French interlanguage: a cross-sectional study. *Bilingualism: Language and Cognition*, *4*(3), 275–297. https://doi.org/10.1017/S136672890100044X

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2016). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, 1–14. https://doi.org/10.1017/S1366728916000547

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: the neglected role of phonology. *Journal of Memory and Language*, *41*, 496–518. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X99926542

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

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, *21*(5), 629–634. https://doi.org/10.1177/0956797610367751

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition*, *110*(3), 293–321. https://doi.org/10.1016/j.cognition.2008.09.008

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-

9280.2007.01845.x

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2017). MultiPic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology*, 1–24. https://doi.org/http://dx.doi.org/10.1080/17470218.2017.1310261

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, *35*(2), 353–387. https://doi.org/10.1017/S0272263112000915

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Perspectives on Morphological Oganization: Data and Analyses.* (pp. 212–239). Leiden: Brill.

Eager, C. D., & Roy, J. (2017). Mixed Effects Models are Sometimes Terrible, 24. Retrieved from http://arxiv.org/abs/1701.04858

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(32), 12203–12208. https://doi.org/10.1073/pnas.0602173103

Favreau, M., & Segalowitz, N. S. (1983). Automatic and controlled processes in the first- and second-language reading of fluent bilinguals. *Memory & Cognition*, *11*(6), 565–574. https://doi.org/10.3758/BF03198281

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., & Kutas, M. (2005). Aging in context: age-related changes in context use during language comprehension. *Psychophysiology*, *42*(2), 133–41. https://doi.org/10.1111/j.1469-8986.2005.00274.x

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161. https://doi.org/10.1016/j.bandl.2010.07.006

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146. https://doi.org/10.1017/S0048577202001373

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*(1), 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, *8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Foote, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Applied Psycholinguistics*, *32*(1), 187–220. https://doi.org/10.1017/S0142716410000342

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

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

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1–9. https://doi.org/10.1037/a0036756

Foucart, A., Ruiz-Tada, E., & Costa, A. (2015). How do you know I was about to say "book"? Anticipation processes affect speech processing and lexical recognition. *Language, Cognition and Neuroscience*, *30*(6), 768–780. https://doi.org/10.1080/23273798.2015.1016047

Foucart, A., Ruiz-Tada, E., & Costa, A. (2016). Anticipation processes in L2 speech comprehension: Evidence from ERPs and lexical recognition task. *Bilingualism: Language and Cognition*, *19*(1), 213–219. https://doi.org/10.1017/S1366728915000486

Franceschina, F. (2005). *Fossilized Second Language Grammars* (Vol. 38). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lald.38

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*.

Friederici, A. D., & Jacobsen, T. (1999). Processing grammatical gender during language comprehension. *Journal of Psycholinguistic Research*, *28*(5), 467–484.

Fruchter, J., Linzen, T., Westerlund, M., & Marantz, A. (2015). Lexical preactivation in basic linguistic phrases. *Journal of Cognitive Neuroscience*, *27*(10), 1912–1935. https://doi.org/10.1162/jocn_a_00822

Gabriele, A., Fiorentino, R., & Alemán Bañón, J. (2013). Examining second language development using event-related potentials: A cross-sectional study on the processing of gender and number agreement. *Linguistic Approaches to Bilingualism*, *3*(2), 213–232. https://doi.org/10.1075/lab.3.2.04gab

Garnham, A., Oakhill, J., Ehrlich, M. F., & Carreiras, M. (1995). Representations and processes in the interpretation of pronouns: New evidence from Spanish and French. *Journal of Memory and Language*, *34*(1), 41–62. https://doi.org/http://dx.doi.org/10.1006/jmla.1995.1003

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

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

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

Goldberg, D., Looney, D., & Lusin, N. (2015). *Enrollments in Languages Other Than English in United States Institutions of Higher Education , Fall 2013*. *Modern Language Association*. New York. Retrieved from https://www.mla.org/content/download/31180/1452509/2013_enrollment_survey.pdf

Gollan, T. H., & Frost, R. (2001). Two routes to grammatical gender: evidence from Hebrew. *Journal of Psycholinguistic Research, 30*(6), 627–51. https://doi.org/10.1023/A:1014235223566

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814. https://doi.org/10.1016/j.jml.2007.07.001

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

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

Grüter, T., Rohde, H., & Schafer, A. J. (2016). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism*. https://doi.org/10.1075/lab.15011.gru

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

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*(4), 556–568. https://doi.org/10.1162/089892900562336

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

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

Hawkins, R., & Chan, Y. (1997). The partial availability of Universal Grammar in second language acquisition: the "failed functional features hypothesis." *Second Language Research*, *13*(3), 187–226.

Hawkins, R., & Franceschina, F. (2004). Explaining the acquisition and non-acquisition of determiner-noun gender concord in French and Spanish. In P. Prévost & J. Paradis (Eds.), *The Acquisition of French in Different Contexts: Focus on Functional Categories* (pp. 175–205). Philadelphia: John Benjamins.

Heath, J. (1975). Some functional relationships in grammar. *Language*, *51*(1), 89–104. https://doi.org/10.2307/413151

Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: Macmillan.

Holmes, V. M., & Dejean De La Bâtie, B. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics*, *20*(4), 479–506. https://doi.org/10.1017/S0142716499004026

Holmes, V. M., & Segui, J. (2004). Sublexical and lexical influences on gender assignment in French. *Journal of Psycholinguistic Research*, *33*(6), 425–457. https://doi.org/10.1007/s10936-004-2665-7

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, 1–11. https://doi.org/10.1177/0956797618754486

Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. https://doi.org/10.1177/0267658312461803

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307. https://doi.org/10.1177/0267658315624960

Hopp, H., & Lemmerth, N. (2016). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, 1–29. https://doi.org/10.1017/S0272263116000437

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 1–18. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Brouwer, S. (2015). Delayed anticipatory spoken language processing in adults with dyslexia - Evidence from eye-tracking. *Dyslexia*, *21*(2), 97–122. https://doi.org/10.1002/dys.1497

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. https://doi.org/10.1080/23273798.2015.1047459

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica*, *137*(2), 151–71. https://doi.org/10.1016/j.actpsy.2010.11.003

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

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, *21*(2), 251–264. https://doi.org/10.1017/S1366728917000050

Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016a). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *69*(4), 1–12. https://doi.org/10.1080/23273798.2016.1242761

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016b). On predicting form and meaning in a second language. *Journal of Experimental Psychology: Learning Memory and Cognition*.

https://doi.org/10.1037/xlm0000315

Jasper, H. (1958). Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalography and Clinical Neurophysiology*, *10*(2), 370–375. https://doi.org/10.1016/0013-4694(58)90053-1

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. J. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

Kader, G. D., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2), 1–17. Retrieved from http://www.amstat.org/publications/JSE/v15n2/kader.pdf

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive procesing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*(1), 37–55.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning*, *59*(3), 503–535.

Kimball, A. E., Shantz, K., Eager, C. D., & Roy, J. (2018). Confronting quasi-separation in logistic mixed effects for linguistic data: A Bayesian approach. *Journal of Quantitative Linguistics*.

Kiss, M., Driver, J., & Eimer, M. (2009). Reward priority of visual target singletons modulates ERP signatures of attentional selection. *Psychological Science*, *20*(2), 245–251. https://doi.org/10.1111/j.1467-9280.2009.02281.x.Reward

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology. General*, *135*(1), 12–35. https://doi.org/http://dx.doi.org/10.1037/0096-3445.135.1.12

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language*, *100*(3), 223–237. https://doi.org/10.1016/j.bandl.2005.12.006

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the Brain* (pp.

190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065

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

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Lardiere, D. (1998). Case and tense in the “fossilized” steady state. *Second Language Research*, *14*(1), 1–26. https://doi.org/10.1191/026765898674105303

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Lemhöfer, K., Schriefers, H., & Indefrey, P. (2014). Idiosyncratic grammars: Syntactic processing in second language comprehension uses subjective feature representations. *Journal of Cognitive Neuroscience*, *26*(7), 1428–1444. https://doi.org/10.1162/jocn_a_00609

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences*, *22*(1), 1-38-75.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18*(3), 193–198. https://doi.org/10.1111/j.1467-9280.2007.01871.x

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Lewis, A. G., Lemhöfer, K., Schoffelen, J., & Schriefers, H. (2016). Gender agreement violations modulate beta oscillatory dynamics during sentence comprehension: A comparison of second language learners and native speakers. *Neuropsychologia*, *89,* 254–272. https://doi.org/10.1016/j.neuropsychologia.2016.06.031

Loerts, H., Wieling, M., & Schmid, M. S. (2013). Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing. *Journal of Psycholinguistic Research*, *42*(6), 551–570. https://doi.org/10.1007/s10936-012-9234-2

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(April), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2012). Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related electrophysiological correlates. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of ERP Components* (pp. 329–360). New York: Oxford University Press.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, *146,* 349–370. https://doi.org/10.1016/j.cognition.2015.10.012

Lundquist, B., Rodina, Y., Sekerina, I. A., & Westergaard, M. (2016). Gender change in Norwegian dialects: Comprehension is affected before production. *Linguistics Vanguard*, *2*(s1), 69–83. https://doi.org/10.1515/lingvan-2016-0026

Macdonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*(April), 1–16. https://doi.org/10.3389/fpsyg.2013.00226

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Mani, N., & Huettig, F. (2014). Word reading skill predicts anticipation of upcoming spoken language input: A study of children developing proficiency in reading. *Journal of Experimental Child Psychology*, *126*, 264–279. https://doi.org/10.1016/j.jecp.2014.05.004

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, *7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230

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

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588. https://doi.org/10.1016/j.jml.2013.08.001

McCarthy, C. (2008). Morphological variability in the comprehension of agreement: an argument for representation over computation. *Second Language Research*, *24*(4), 459–486. https://doi.org/10.1177/0267658308095737

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience*, *7*(7), 703–704. https://doi.org/10.1038/nn1264

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G. D., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*(s2), 123–150. https://doi.org/10.1111/j.1467-9922.2010.00604.x

Meulman, N., Stowe, L. A., Sprenger, S. A., Bresser, M., & Schmid, M. S. (2014). An ERP study on L2 syntax processing: When do learners fail? *Frontiers in Psychology*, *5*, 1072. https://doi.org/10.3389/fpsyg.2014.01072

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PloS One*, *10*(12), e0143328. https://doi.org/10.1371/journal.pone.0143328

Mishra, R. K., Singh, N., Pandey, A., & Huettig, F. (2012). Spoken language-mediated anticipatory eye- movements are modulated by reading ability - Evidence from Indian low and high literates. *Journal of Eye Movement Research*, *5*(1), 1–10.

Molinaro, N., Vespignani, F., & Job, R. (2008). A deeper reanalysis of a superficial feature: An ERP study on agreement violations. *Brain Research*, *1228*, 161–176. https://doi.org/10.1016/j.brainres.2008.06.064

Montrul, S. A. (2008). *Incomplete Acquisition in Bilingualism: Re-Examining the Age Factor*. John Benjamins Publishing.

Montrul, S. A. (2015). *The Acquisition of Heritage Languages*. Cambridge University Press.

Montrul, S. A., Davidson, J., De la Fuente, I., & Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Bilingualism: Language and Cognition*, *17*(1), 118–138. https://doi.org/10.1017/S1366728913000114

Montrul, S. A., Foote, R., & Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Language Learning*, *58*(3), 503–553. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9922.2008.00449.x/full

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. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9922.2009.00554.x/full

Nevins, A., Dillon, B., Malhotra, S., & Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Research*, *1164*, 81–94. https://doi.org/10.1016/j.brainres.2007.05.058

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2017). Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *BioRxiv*, 111807. https://doi.org/10.1101/111807

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, 1–24. https://doi.org/10.7554/eLife.33468

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

O'Rourke, P. L., & van Petten, C. (2011). Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential. *Brain Research*, *1392*, 62–79. https://doi.org/10.1016/j.brainres.2011.03.071

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *20*(4), 786–803. https://doi.org/10.1037/0278-7393.20.4.786

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning*, *56*(SUPPL. 1), 199–230. https://doi.org/10.1111/j.1467-9922.2006.00361.x

Otten, M., & van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Paul, J. Z., & Grüter, T. (2016). Blocking effects in the learning of Chinese classifiers. *Language Learning*, *66*(4), 972–999. https://doi.org/10.1111/lang.12197

Peereman, R., Dufour, S., & Burt, J. S. (2009). Orthographic influences in spoken word recognition: The consistency effect in semantic and gender categorization tasks.

Psychonomic Bulletin and Review, *16*(2), 363–368. https://doi.org/10.3758/PBR.16.2.363

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, *53*(1), 146–156. https://doi.org/10.1016/j.neuropsychologia.2013.11.014

Piai, V., Roelofs, A., Rommers, J., & Maris, E. (2015). Beta oscillations reflect memory and motor aspects of spoken word production. *Human Brain Mapping*, *36*(7), 2767–2780. https://doi.org/10.1002/hbm.22806

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329–47. https://doi.org/10.1017/S0140525X12001495

Prévost, P., & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, *16*(2), 103–133. https://doi.org/10.1191/026765800677556046

Qian, Z., & Garnsey, S. M. (2016). A sheet of coffee: an event-related brain potential study of the processing of classifier-noun sequences in English and Mandarin. *Language, Cognition and Neuroscience*, *3798*(May), 1–24. https://doi.org/10.1080/23273798.2016.1153116

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514–528. https://doi.org/10.1037/a0020990

Riordan, B., Dye, M., & Jones, M. N. (2015). Grammatical number processing and anticipatory eye movements are not tightly coordinated in English spoken language comprehension. *Frontiers in Psychology*, *6*(JAN), 1–11. https://doi.org/10.3389/fpsyg.2015.00590

Romero-Rivas, C., Martin, C. D., & Costa, A. (2016). Foreign-accented speech modulates linguistic anticipatory processes. *Neuropsychologia*, *85*, 245–255. https://doi.org/10.1016/j.neuropsychologia.2016.03.022

Rommers, J., Dickson, D. S., Norton, J. J. S., Wlotko, E. W., & Federmeier, K. D. (2016). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, 1–14. https://doi.org/10.1080/23273798.2016.1183799

Rommers, J., Meyer, A. S., & Praamstra, P. (2017). Lateralized electrical brain activity reveals covert attention allocation during speaking. *Neuropsychologia*, *95*(December 2016), 101–110. https://doi.org/10.1016/j.neuropsychologia.2016.12.013

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception, 33*(2), 217–236. https://doi.org/10.1068/p5117

Sabourin, L., & Stowe, L. A. (2008). Second language processing: when are first and second languages processed similarly? *Second Language Research*, *24*(3), 397–430. https://doi.org/10.1177/0267658308090186

Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research, 22*(1), 1–29. https://doi.org/10.1191/0267658306sr259oa

Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua*, *120*(8), 2022–2039. https://doi.org/10.1016/j.lingua.2010.02.004

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145

Schriefers, H., & Jescheniak, J. D. (1999). Representation and processing of grammatical gender in language production: A review. *Journal of Psycholinguistic Research*, *28*(6), 575–600. Retrieved from http://link.springer.com/article/10.1023/A:1023264810403

Segalowitz, N. S., & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Appriaches* (pp. 371–388). New York: Oxford University Press.

Shantz, K., & Tanner, D. (2016). Are L2 learners pressed for time? Retrieval of grammatical gender information in L2 lexical access. In J. Scott & D. Waughtal (Eds.), *Proceedings of the 40th Annual Boston University Conference on Language Development* (pp. 331–345). Cascadilla Press.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75. https://doi.org/10.1016/j.jml.2015.07.003

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists, *12*(3), 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Spalek, K., Franck, J., Schriefers, H., & Frauenfelder, U. H. (2008). Phonological regularities and grammatical gender retrieval in spoken word recognition and word production. *Journal of Psycholinguistic Research*, *37*(6), 419–42. https://doi.org/10.1007/s10936-008-9074-2

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*(8), 311–327. https://doi.org/10.1111/lnc3.12151

Szewczyk, J. M., & Schriefers, H. (2013). Prediction in language comprehension beyond specific words: An ERP study on sentence comprehension in Polish. *Journal of Memory and Language*, *68*(4), 297–314. https://doi.org/10.1016/j.jml.2012.12.002

Taft, M., & Meunier, F. (1998). Lexical representation of gender: a quasiregular domain. *Journal of Psycholinguistic Research*, *27*(1), 23–45. https://doi.org/10.1023/A:1023270723066

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *Journal of Neuroscience*, *30*(43), 14356–14360. https://doi.org/10.1523/JNEUROSCI.3028-10.2010

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, *16*(2), 367–382. https://doi.org/10.1017/S1366728912000302

Taraban, R., & Kempe, V. (1999). Gender processing in native and nonnative Russian speakers. *Applied Psycholinguistics*, *20*, 119–148. https://doi.org/10.1017/S0142716499001046

Thornhill, D. E., & van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: an event-related potential investigation. *Studies in Second Language Acquisition*, *27*(2), 173–204. https://doi.org/10.1017/S0272263105050102

Tsimpli, I. M., & Dimitrakopoulou, M. (2007). The Interpretability Hypothesis: evidence from wh-interrogatives in second language acquisition. *Second Language Research*, *23*(2), 215–242.

van Bergen, G., & Flecken, M. (2017). Putting things in new places: Linguistic experience modulates the predictive power of placement verb semantics. *Journal of Memory and Language*, *92*, 26–42. https://doi.org/10.1016/j.jml.2016.05.003

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Weber, A., Grice, M., & Crocker, M. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*(2), B63–B72. https://doi.org/10.1016/j.cognition.2005.07.001

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482.

White, L. (2011). Second language acquisition at the interfaces. *Lingua*, *121*(4), 577–590. https://doi.org/10.1016/j.lingua.2010.05.005

White, L., Valenzuela, E., Kozlowska–Macgregor, M., & Leung, Y.-K. I. (2004). Gender and number agreement in nonnative Spanish. *Applied Psycholinguistics*, *25*(1), 105–133. https://doi.org/10.1017/S0142716404001067

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*(3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). Expecting gender: An event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. *Cortex*, *39*(3), 483–508. https://doi.org/10.1016/S0010-

9452(08)70260-0

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–88. https://doi.org/10.1162/0898929041920487

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia*, *45*(13), 3001–3014. https://doi.org/10.1016/j.neuropsychologia.2007.05.013

Wlotko, E. W., & Federmeier, K. D. (2012a). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology*, *49*(6), 770–85. https://doi.org/10.1111/j.1469-8986.2012.01366.x

Wlotko, E. W., & Federmeier, K. D. (2012b). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, *62*(1), 356–366. https://doi.org/10.1016/j.neuroimage.2012.04.054

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex*, *68*, 20–32. https://doi.org/10.1016/j.cortex.2015.03.014

Woodman, G. F., Arita, J. T., & Luck, S. J. (2009). A cuing study of the N2pc component: An index of attentional deployment to objects rather than spatial locations. *Brain Research*, *1297*, 101–111. https://doi.org/10.1016/j.brainres.2009.08.011

Woodman, G. F., & Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature*, *400*(6747), 867–869. https://doi.org/10.1038/23698

Zubin, D. A., & Köpcke, K.-M. (1986). Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In C. G. Collete (Ed.), *Noun Classes and Categorization: Proceedings of a Symposium on Categorization and Noun Classification, Eugene, Oregon, October 1983* (pp. 139–180). Philadelphia: John Benjamins.

APPENDIX A. ITEMS FOR EXPERIMENTS 1 AND 2

Noun labels for Experiments 1 and 2 grouped by gender class in the artificial grammar with their English and German names (EName, GName), German gender (GGend), log English frequency per million (LogEF; Brysbaert & New, 2009), log German frequency per million (LogDF; Brysbaert et al., 2011), mean concreteness ratings (Brysbaert et al., 2014), and mean age of acquisition (AoA; Kuperman et al., 2012).

Gender Class I							
Noun Label	EName	GName	GGend	LogEF	LogGF	Concreteness	AoA
etkot	shirt	Hemd	Neuter	1.68	1.47	4.94	3.53
fertsot	leaf	Blatt	Neuter	0.79	1.12	5.00	4.60
hekloo	glass	Glas	Neuter	1.79	1.72	4.82	4.47
gesoo	bed	Bett	Neuter	2.27	2.21	5.00	2.89
panjol	tree	Baum	Masculine	1.82	1.62	5.00	3.57
jatree	stove	Herd	Masculine	0.93	0.82	4.96	4.32
perdip	moon	Mond	Masculine	1.71	1.50	4.90	4.83
sodap	ball	Ball	Masculine	2.03	1.76	5.00	2.90
toonbot	jacket	Jacke	Feminine	1.54	1.46	4.86	3.95
viltord	scissors	Schere	Feminine	0.89	0.91	4.85	4.50
romdee	glasses	Brille	Feminine	1.53	1.42	4.90	4.34
sidpal	hand	Hand	Feminine	2.45	2.37	4.72	2.74
Means				1.62	1.53	4.91	3.89
Standard Deviations				0.51	0.44	0.08	0.71
Gender Class II							
Noun Label	EName	GName	GGend	LogEF	LogGF	Concreteness	AoA
pikroo	ear	Ohr	Neuter	1.52	1.49	5.00	3.63
gorok	bread	Brot	Neuter	1.47	1.41	4.92	3.58
slindot	wheel	Rad	Neuter	1.45	1.14	4.86	4.40
tilmoo	window	Fenster	Neuter	1.94	1.93	4.86	4.74
famtog	chair	Stuhl	Masculine	1.70	1.50	4.58	3.43
bagdee	shoe	Schuh	Masculine	1.50	1.19	4.97	2.60
fitloo	pot	Topf	Masculine	1.37	0.88	4.81	5.95
orteep	train	Zug	Masculine	1.98	1.88	4.79	4.00
dalku	bottle	Flasche	Feminine	1.71	1.57	4.91	3.56
pugtee	flower	Blume	Feminine	1.38	1.27	5.00	3.11
gertom	whistle	Pfeife	Feminine	1.22	1.05	4.42	5.42
pridmos	watch	Uhr	Feminine	2.52	2.43	4.61	4.33
Means				1.65	1.48	4.81	4.06
Standard Deviations				0.34	0.42	0.18	0.92

APPENDIX B. ITEMS FOR EXPERIMENT 3

Items for Experiment 3 with their grammatical gender, English name (EName), log frequency per million (LogFreq; Brysbaert et al., 2011), percent naming agreement (NamAgr; Duñabeitia et al., 2017), phonological neighborhood density (PND; Marian et al., 2012) and phonemic length (PLength).

Item	Gender	EName	LogFreq	NamAgr	PND	PLength
Ananas	Feminine	Pineapple	1.79	100.00	0	6
Bank	Feminine	Bench	3.16	91.00	10	4
Birne	Feminine	Pear	2.18	100.00	2	5
Bombe	Feminine	Bomb	3.17	95.96	3	5
Brille	Feminine	Glasses	2.81	100.00	4	5
Ente	Feminine	Duck	2.43	85.57	7	4
Faust	Feminine	Fist	2.52	100.00	7	4
Giraffe	Feminine	Giraffe	1.49	100.00	0	6
Kanone	Feminine	Canon	2.52	95.79	1	6
Kette	Feminine	Chain	2.69	95.92	11	4
Mauer	Feminine	Wall	2.63	92.93	1	3
Nase	Feminine	Nose	3.26	100.00	1	4
Paprika	Feminine	Pepper	1.48	100.00	0	7
Pfeife	Feminine	Pipe	2.41	100.00	7	4
Pyramide	Feminine	Pyramid	1.91	85.06	1	8
Rakete	Feminine	Rocket	2.61	100.00	1	6
Schnecke	Feminine	Snail	1.88	95.00	4	5
Tasche	Feminine	Purse	3.25	73.00	9	4
Tomate	Feminine	Tomato	1.71	100.00	1	6
Zwiebel	Feminine	Onion	1.60	100.00	1	6
Arm	Masculine	Arm	3.32	100.00	12	3
Ball	Masculine	Ball	3.16	89.58	34	3
Baum	Masculine	Tree	3.01	100.00	10	
Delfin	Masculine	Dolphin	1.60	100.00	2	6
Drache	Masculine	Dragon	2.43	100.00	2	5
Drucker	Masculine	Printer	1.61	93.88	5	5
Fisch	Masculine	Fish	3.04	78.79	9	3
Hammer	Masculine	Hammer	2.58	100.00	6	4
Koffer	Masculine	Suitcase	2.98	87.00	7	4
Mantel	Masculine	Coat	2.73	97.00	2	6
Pinguin	Masculine	Penguin	1.78	100.00	1	6
Sarg	Masculine	Coffin	2.49	100.00	8	4
Schlüssel	Masculine	Key	3.49	10.00	3	6
Schrank	Masculine	Wardrobe	2.76	95.96	5	5
Sessel	Masculine	Armchair	2.07	97.98	2	5
Skorpion	Masculine	Scorpion	2.02	86.73	1	8
Stein	Masculine	Stone	3.05	100.00	9	4
Stuhl	Masculine	Chair	2.89	100.00	10	4
Tiger	Masculine	Tiger	2.64	96.97	6	4

Traktor	Masculine	Tractor	1.71	94.95	0	7
Aquarium	Neuter	Aquarium	1.93	94.95	0	8
Gehirn	Neuter	Brain	3.14	91.00	2	6
Geschenk	Neuter	Gift	3.25	97.96	2	6
Glas	Neuter	Glass	3.11	95.92	9	4
Hemd	Neuter	Shirt	2.86	95.92	9	4
Horn	Neuter	Horn	2.16	92.13	8	4
Kamel	Neuter	Camel	1.91	90.82	4	5
Kleid	Neuter	Dress	3.04	97.96	4	4
Knie	Neuter	Knee	3.00	100.00	13	3
Lineal	Neuter	Ruler	1.40	100.00	0	6
Netz	Neuter	Net	2.65	96.84	11	3
Pferd	Neuter	Horse	3.11	91.75	7	4
Puzzle	Neuter	Puzzle	1.85	100.00	2	5
Radio	Neuter	Radio	2.94	92.71	1	5
Schwein	Neuter	Pig	3.20	86.87	9	4
Schwert	Neuter	Sword	3.21	90.72	7	5
Skelett	Neuter	Skeleton	1.91	100.00	1	6
Sofa	Neuter	Couch	2.45	93.94	6	4
Walross	Neuter	Walrus	1.58	79.57	0	7
Zebra	Neuter	Zebra	1.63	100.00	1	5
			Mean (SD)			
		Feminine	2.37 (0.59_	95.51 (7.12)	4.6 (4.2)	5.1 (1.3)
		Masculine	2.57 (0.59)	95.94 (5.96)	6.7 (7.4)	4.8 (1.4)
		Neuter	2.52 (0.65)	94.45 (5.25)	4.8 (4.1)	4.9 (1.3)