ABSTRACT

Title of dissertation: MODELING THE EFFECTS OF ENTRENCHMENT AND MEMORY DEVELOPMENT ON SECOND LANGUAGE ACQUISITION

Peter Daniel Osthus, Doctor of Philosophy, 2019

Dissertation directed by: Professor Robert DeKeyser
Department of Second Language Acquisition

The observation that language learning outcomes are less consistent the older one becomes has motivated a large portion of second language acquisition research (e.g., Hartshorne, Tenenbaum, & Pinker, 2018; DeKeyser, 2012). Hypotheses about the underlying mechanisms which lead to age-related declines are traditionally tested with human subjects; however, many hypotheses cannot be fully evaluated in the natural world due to maturational and environmental constraints. In these scenarios, computational simulations provide a convenient way to test these hypotheses.

In the present work, recurrent neural networks are used to study the effects of linguistic entrenchment and memory development on second language acquisition. Previous computational studies have found mixed results regarding these factors. Three computational experiments using a range of languages were conducted to understand better the role of entrenchment and memory development in learning several linguistic sub-tasks: grammatical gender assignment, grammatical gender agreement, and word boundary identification.

Linguistic entrenchment consistently had a negative, but marginal, influence on second language learning outcomes in the gender assignment experiment. In the gender agreement and word boundary experiments, entrenchment rarely affected learning outcomes. Starting with fewer memory resources consistently led to poorer outcomes across learning tasks and languages. The complexity of the learning task and the regularity of the formal cues present in the linguistic input affected outcomes. In the gender assignment experiment, the first language influenced second language outcomes, especially when the second language had fewer gender classes than the first language. These results suggest that the effects of entrenchment and memory development on second language learning may be dependent upon the language pairs and the difficulty of the modeling task.

MODELING THE EFFECTS OF ENTRENCHMENT AND MEMORY DEVELOPMENT ON SECOND LANGUAGE ACQUISITION

by

Peter Daniel Osthus

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Robert DeKeyser (Chair)
Professor Steve Ross
Professor Marine Carpuat
Dr. Jared Linck
Professor Naomi Feldman (Dean's Representative)

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Humans are exceptional language learners when they are young. However, repeating the successes of the first language (L1) in a second language (L2) becomes less likely as time goes on (e.g., Hartshorne et al., 2018; DeKeyser, 2012). Over the past five decades, researchers have proffered many explanations of what causes these age-related declines in L2 learning outcomes. The vocabulary used by a researcher to describe age-related effects identifies how they think about the underlying causal mechanisms, which frequently center around maturational and experiential causes.

Maturational explanations often invoke the concept of a critical period (or sensitive period) to stress the underlying biological origins of cognition. Maturational mechanisms have been related to neurobiological factors, like hemispheric specialization (Lenneberg, 1967) and myelination (Long, 1990; Gao et al., 2009), as well as cognitive factors, like working memory development (Newport, 1988, 1990) and the transition from procedural to declarative learning (Ullman, 2014). Experiential explanations, on the other hand, emphasize the influence of previous experiences and knowledge on the system. Linguistic entrenchment (MacWhinney, 2005, 2016) and factors like the amount and type of L2 input (Jia, Aaronson, & Wu, 2002; Jia & Aaronson, 2003) fall within this group. Other researchers have looked at individual

differences as a hint to the change in learning processes that underlie age-related declines in learning outcomes, like motivation to sound native (Bley-Vroman, 1988) and learning aptitude (Abrahamsson & Hyltenstam, 2008; DeKeyser, Alfi-Shabtay, & Ravid, 2010).

The drive to discover the root causes of age-related effects on L2 learning has led to a mountain of data. Unfortunately, this interest has not yielded a consensus concerning which mechanisms explain why children typically have better language learning outcomes in the long run, how long the critical period lasts, and the shape of the language learning curve (Hartshorne et al., 2018). The dynamic nature of language learning makes it difficult to isolate and control many of these variables (DeKeyser & Larson-Hall, 2005; DeKeyser, 2012), which is why most of these studies have either been correlational or lacked sufficient statistical power for the claims they make (Hartshorne et al., 2018). Many of the factors of interest correlate with each other, and to make matters more complicated, they often develop concurrently. These developmental constraints make it impossible to test all possible hypotheses using human participants. Computational simulations, therefore, have been pursued as a method to test hypotheses that are hard, or impossible, to manipulate in the natural environment.

Models in cognitive science come in three general flavors: verbal-conceptual, computational, and mathematical (Sun, Coward, & Zenzen, 2005). Most models in L2 learning research are verbal-conceptual. Computational models are much less common, and mathematical models are even less common (see Gold, 1967). In computational models, algorithms are often specified with pseudo-code, but any im-

plementation of the model must use programming languages that can be used to perform computations on a von Neumann machine (i.e., a computing device). A key advantage of computational models over verbal-conceptual models is that algorithmic implementations require fine-grained specificity. Verbal-conceptual models are often vague and imprecise, whereas computational models can provide clarity and precision to the concepts represented in the models.

This dissertation uses computational models to understand better the causal mechanisms underlying age-related declines in L2 learning outcomes. Recurrent neural networks were trained to test two developmental hypotheses that develop concurrently throughout childhood. The first hypothesis concerns linguistic entrenchment, an experiential factor that expects greater levels of L1 knowledge and experience to lead to poorer L2 learning outcomes. The second hypothesis concerns the *less-is-more* hypothesis of Newport (1988, 1990). This hypothesis claims that starting the learning process with fewer working memory resources provides advantages when learning low-level attributes of the input. It is not possible to thoroughly evaluate these hypotheses in the natural world due to maturational and environmental constraints; therefore, computational simulations provide an opportunity to understand how these factors may influence L2 learning outcomes.

Linguistic input from two languages (L1 and L2) was used to model the learning of three linguistic sub-tasks (grammatical gender assignment, grammatical gender agreement and word boundary identification) while manipulating the amount of linguistic entrenchment before initial exposure to an L2 and how memory resources develop over time. All three experiments in this study used recurrent neu-

ral networks with a long short-term memory architecture (LSTM; Hochreiter & Schmidhuber, 1997; Gers, Schmidhuber, & Cummins, 1999; Gers, Schraudolph, & Schmidhuber, 2002) to represent language learning agents. The influence of within- and between-language characteristics on learning outcomes was explored in relation to the factors of entrenchment and working memory development.

The remainder of this chapter reviews research on the statistical learning of language and its contribution to the entrenchment and less-is-more hypotheses. Previous computational research addressing the effects of entrenchment and memory development in language learning is reviewed. The experiments of Monner, Vatz, Morini, Hwang, and DeKeyser (2013), which inspired the research presented here, are discussed in detail.

1.1 Statistical Learning

The accumulation of distributional information from perceptual input, commonly referred to as statistical learning, has been at the forefront of language acquisition research for nearly four decades. Newport (2016) describes the statistical learning of language as the ability to extract statistical information from the linguistic distribution of elements in speech. These learned statistical regularities underpin core linguistic capabilities, like the ability to identify which sound patterns form the meaningful units of a language, determine the syntactic constraints of a language, identify the grammatical categories of linguistic elements, and compose the phrasal organization of a language (Newport, 2016).

In two seminal studies by Saffran, Aslin, and Newport (1996) and Saffran, Newport, and Aslin (1996), infants and adults were able to learn to segment a continuous stream of synthesized speech into words based upon the distribution of the sound segments (see also Aslin, Saffran, & Newport, 1998). The synthesized speech provided to learners was controlled so that only the statistical regularities of the sounds were manipulated. No cues were provided in the audio signal to indicate when a word boundary occurred. Training input manipulated the transitional probabilities (see Equation 1.1) of within-word sequences and between-word sequences so that relatively high transitional probabilities were the words and those with relatively low transitional probabilities were the boundaries between words. These studies showed that both infants and adults tended to select sequences with relatively high transitional probabilities as words, suggesting that the subjects used the transitional probabilities between temporally adjacent phonemes to infer word boundaries.

$$p(x_{ij}) = P(X_n = j | X_{n-1} = i) \tag{1.1}$$

Subsequent studies have expanded this research agenda from adjacent dependencies to non-adjacent dependencies in sound (Newport & Aslin, 2004) and syntactic patterns (Thompson & Newport, 2007; Wonnacott, Newport, & Tanenhaus, 2008; Reeder, Newport, & Aslin, 2013; Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018). These studies find that learners, both young and old, can learn from the statistical patterns in the input. However, when input is less consistent, there

are age-related differences in performance measures.

A significant difference between the performance of children and adults lies in their ability to regularize input or apply rules to the input. Adults closely reproduce the statistics of the input, especially any inconsistencies, whereas children show a preference for the most regularly used form (Newport, 2016). Research has found that these patterns of performance change over time. As children get older, there is a gradual shift from an emphasis on rule consistency and generalization to one of replication of the statistics of the input. In other words, there may be an age-related transition from generalization tendencies to pure statistical replication. Results from multiple studies indicate that this is a pattern found across multiple modalities (and across species; see Wilson et al., 2018) and, therefore, a central mechanism available to the entire cognitive system (see Siegelman et al., 2018).

1.2 Entrenchment

Computational models in multiple domains suggest that the typical accumulation of experience may be the primary cause of age-related effects (Ellis & Lambon Ralph, 2000; Zevin & Seidenberg, 2002; Hernandez & Li, 2007; Thomas & Johnson, 2008; Monner et al., 2013). Seidenberg and Zevin (2006) call this the *paradox of success*; the more one learns, the harder it becomes to learn something new. This process, often referred to as entrenchment, has been used to describe age-related declines in language learning outcomes. For example, Langacker (1987) considers the frequency of language use to be the primary driver of linguistic en-

trenchment in cognitive systems. Language skills in the L1 become entrenched over time. The more entrenched the L1 skills become, the more difficult it becomes to either vary or hinder how one uses that skill. This is precisely the scenario L2 learners face, especially if they begin learning later in life. Learning the statistical properties of an L2 becomes more difficult as knowledge of the statistical properties of an L1 becomes entrenched through continued use of and exposure to the L1. Even if the L1 is no longer used often or at all by a learner, the L1 will persist, and entrenchment effects can still be found even after years of only using the L2.

In computational studies of L2 acquisition, both supervised (Monner et al., 2013) and unsupervised neural networks (Li, Farkas, & MacWhinney, 2004; Li, Zhao, & Mac Whinney, 2007; Zhao & Li, 2010; Li & Zhao, 2013) have been shown to demonstrate entrenchment effects (see MacWhinney, 2016). These computational studies illustrate a gradual decline in L2 learning outcomes as levels of entrenchment increase. The DevLex and DevLex II simulations provide a graphical example of entrenchment effects (Li et al., 2004, 2007; Li & Zhao, 2013). In the DevLex model (Li et al., 2004), two self-organizing maps (SOM; Kohonen, 1990) modeled how the topographical organization of English lexical items changes as the lexicon grows. The models linked a semantic SOM and a phonological SOM with associative connections using Hebbian learning. The activation patterns produced by English lexical items were loosely organized by part-of-speech on the topographical map of the SOM. The organization of these items was plastic and moved around the topographical map during the early stages of learning. As training progressed, the activation patterns stabilized, and regions associated with the different part-of-speech categories became

more stable. The results showed entrenchment effects related to word density and semantic similarity that mirror empirical results in children (Gershkoff-Stowe & Smith, 1997; Ellis & Lambon Ralph, 2000).

Subsequent studies using a similar computational framework focused on network adaptation and change due to the introduction of an L2. In one such study using the DevLex II model, Zhao and Li (2010) trained models on 1,000 lexical items from two different languages, Chinese (Cantonese) and English, under three developmental conditions: 1) simultaneous L1-L2 learning, 2) early L2 learning, and 3) late L2 learning. Their results found consistent and progressive entrenchment effects in bilingual networks, similar to the results reported in Li et al. (2004). As entrenchment increased, so did dispersion in the organization of the L2 lexicon. Simultaneous and early bilingual networks showed both semantic and phonological entrenchment resulting in large contiguous areas of the topographical map belonging to either the L1 or L2 lexicon. Late bilingual networks consisted of small islands of L2 lexical activation, suggesting that late L2 learners have a more dispersed and less cohesive representation of L2 words.

These computational examples highlight that entrenchment can be used to account for age-related declines in L2 learning outcomes. These declines are gradual, indicating that the plasticity of the network decreases slowly but monotonically over time.

When an entrenched cognitive system is exposed to input drastically different from what it has encountered up to that point, the system can undergo *catastrophic interference* (see MacWhinney, 2005, 2016). In the case of language learning, en-

trenched linguistic patterns can block the learning of new patterns needed for an L2. For example, phonological rules, like the difficulty Japanese learners of English have in producing and perceiving /r/ and /l/ (MacWhinney, 2008; Schatz, Feldman, Goldwater, Cao, & Dupoux, 2019), are often difficult for L2 learners to acquire. This phonemic distinction, /r/ versus /l/, is not present in Japanese. The monolingual Japanese speaker does not have a system that distinguishes between these two sounds; therefore, their system often applies one deliberate category to what speakers of other languages containing these two phonemes would easily differentiate.

There are many ways in which knowledge of one language can influence the learning of another. In the case of L2 learning, all L2 learning must initially rely upon the cognitive system that developed through exposure to the L1. In some cases, languages can be parasitic on each other. That is, they can transfer knowledge from the L1 to the L2. The transfer of knowledge is potentially bi-directional (L1 to L2 and L2 to L1); however, during the initial stages of L2 learning transfer occurs from L1 to L2. The transfer of this knowledge can have either positive or negative effects with regards to the ultimate learning of an L2. If the transfer of knowledge is positive, it can be a powerful catalyst when acquiring an L2, which has many similarities with the L1. For example, L1 Spanish speakers learning Italian will be able to apply a sizable portion of their Spanish grammar, phonetic rules, and lexicon to Italian. However, Spanish speakers learning Norwegian do not have as many rules and do not have many lexical items that overlap. The shared presence of similar patterns and rules in Spanish and Italian may lead to positive L1 transfer,

ultimately mitigating potentially adverse effects of entrenchment.

Nevertheless, the similarity between two languages does not always convey positive results and can sometimes lead to difficulties in acquiring specific aspects of an L2 (e.g., rule oversimplification, false cognates). For any given language pair, there will be certain elements that lead to positive transfer and certain problem areas that interfere with the learning process.

1.3 Memory Development

Human memory is often divided into two inter-related, but distinguishable biological systems: working memory and long-term memory. Long-term memory is responsible for the permanent storage of information in the cognitive system, whereas working memory is dedicated to the processing, maintenance, and control of information relevant to immediate demands (Bayliss, Jarrold, Baddeley, Gunn, & Leigh, 2005; Conway, Jarrold, Kane, Miyake, & Towse, 2007; Linck, Osthus, Koeth, & Bunting, 2014). There are many theoretical models of working memory (Miyake & Shah, 1999), but its function to organize, store, and coordinate sensory input while integrating this information into cognitive processes is consistent across models.

Working memory is often described in terms of system capacity. That is, how much information can be stored and manipulated to perform a task. The capacity of the working memory system is defined across at least two dimensions: 1) amount of information that can be recalled and 2) speed of access to relevant

information. Often, working memory capacity is measured using a psychological task that requires the individual to retain some information over a period of time while continuing to receive new and sometimes contradictory information. The more pieces of information an individual retains at the time of recall, the greater their working memory capacity. During childhood, working memory capacities across these two dimensions increase steadily (Cowan & Alloway, 1997; Gathercole, 1999). Language learners with larger working memory capacities are associated with better outcomes on performance measures, suggesting that working memory is important for both L1 and L2 language learning and processing (Daneman & Merikle, 1996; Linck et al., 2014).

A meta-analysis by Linck et al. (2014) provided a synthesis of 748 effect sizes addressing working memory and L2 processing and proficiency. The analysis yielded a modest population effect size estimated at 0.255, with executive functions more strongly correlated to L2 outcomes than storage-based functions. However, these results are based solely on studies using adult participants who were exposed to their L2 well after reaching a high level of proficiency in their L1. It remains to be determined whether the findings from Linck et al. (2014) would generalize to other populations that are exposed to an L2 earlier during cognitive development.

Newport (1988, 1990) argues that limited non-linguistic cognitive capabilities, like working memory, are crucial to early stages of language development and therefore may be a cause of the age-related declines in L2 learning outcomes. This hypothesis, called less-is-more, argues that the difference in learning outcomes between children and adults stems from the way in which linguistic input is perceived

and stored. Specifically, the relative advantage afforded to children in the componential analysis of linguistic stimuli is argued to be due to their limited cognitive abilities. Morphological structures, for example, are highly componential. Language users must assess the individual parts of an input to determine the intended semantic representation. Newport (1988, 1990) claims that children are better able to acquire patterns in language that are componential since there is a limited time window for processing. However, the less-is-more hypothesis does not apply to all aspects of language learning (Newport, 2016). Language units that are analyzed as a whole or integratively, such as that experienced in word order learning and whole word learning, require greater working memory capacity in order to link elements separated by time and space.

It is difficult to evaluate the less-is-more hypothesis in empirical studies with human learners. Instead, researchers have used computational simulations to model the effect of memory development on language learning. In one of the earliest attempts to assess the role of working memory capacity during development, Elman (1993) trained simple recurrent networks (SRN; Elman, 1990) to use contextual information derived from current and previous words in a sequence to predict the next word. Recurrent neural networks, of which SRNs are a member, are able to selectively retain information within the recent past in order to make judgments about the future. These networks are often described as a type of working memory. Elman (1993) manipulated the developmental state of working memory by controlling the number of time-steps the network was able to maintain actively. As the network matured, the number of time-steps the network could maintain in its context layer

increased. The results of the working memory development simulation in Elman (1993) support the idea that starting with fewer working memory resources leads to better learning outcomes. The results from Elman (1993) have been used to argue in favor of the less-is-more hypothesis (Newport, 1988, 1990). A subsequent study by Rohde and Plaut (1999) failed to replicate the main findings from Elman (1993). Not only did this study fail to replicate the earlier findings, it actually found that starting with fewer working memory resources led to worse performance on the outcome measure.

A more recent computational experiment by Monner et al. (2013) modeled the growth of working memory capabilities differently. Instead of manipulating the number of time-steps the network was able to maintain actively, the size and connectivity of the network is manipulated. This study is discussed below.

1.4 Overview of Monner et al. (2013)

In two experiments, Monner et al.(2013) tested hypotheses related to linguistic entrenchment and memory development. Recurrent neural networks with a generalized implementation of the LSTM architecture (Monner & Reggia, 2012) learned the statistical distributions of sounds (i.e., phonemes) as they related to grammatical gender in French and Spanish. Since the grammatical gender systems of French and Spanish are considered transparent (i.e., formal cues largely predict grammatical gender class membership in Spanish more than in French nouns), all language data used to train the models was represented phonemically.

The first experiment of Monner et al.(2013) trained models to assign a gender class to nouns in French and Spanish. Training input consisted of nouns matched with a gender-appropriate determiner. The orthographic representation of text in French and Spanish was transliterated into a sequence of phonemes represented by the International Phonetic Alphabet (IPA; *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, 1999). Each phoneme was represented as a unique vector of 19 articulatory features (consonantal, sonorant, continuant, strident, nasal, lateral, trill, voice, labial, round, coronal, anterior, distributed, dorsal, high, low, back, radial, adv tongue root). The articulatory features represent how humans physiologically produce speech sounds (see Table 2.2 for a description of the features used in the present experiments). Each feature was either present (1) or not present (0) in the vector. Since spoken language is sequential, each sequence was represented as an ordered series of articulatory feature vectors, with each vector representing one phoneme segment. The ordered sequence of vectors was entered as input into the neural network one by one to simulate the temporal aspect of the perception of speech in the natural environment. The model architecture consisted of one input layer which received vectors of 19 articulatory features, one hidden layer (30 LSTM units) representing the working memory of the network and one output layer (9 units) representing the possible determiners in French and Spanish (*le*, *la*, *l'*, *un*, and *une* in French and *el*, *la*, *un*, and *una* in Spanish). The input layer was also connected to the output layer in order to provide raw input directly to the output layer.

In the second experiment, language models were trained with determiner-

noun-adjective or determiner-adjective-noun sets to predict the next two subsequent phonemes in a sequence. The word sets were transformed into a sequence of articulatory feature vectors representing phonemes. The models in the second experiment were evaluated on how well they completed a sequence with a determiner-noun-adjective word set. Models were presented with a sequence representing the determiner, the noun, and the root form of the adjective. Once the input sequence reached the end, the model was prompted to predict the next two phonemes that would correctly complete the sequence. This task was called gender agreement since model performance was evaluated by whether the predicted phonemes at the end of the adjective matched the gender class of the noun. The model architecture in the gender agreement experiment was more complex. It consisted of one input layer which received vectors of 19 articulatory features, two hidden layers (30 LSTM units in each layer) representing the working memory of the network, and two output layers (19 units), each representing the 19 articulatory features. The input layer was also connected to the two output layers in order to provide raw input directly to the output layers.

In both experiments, linguistic entrenchment and memory development were manipulated during the training phase. The first factor, entrenchment, manipulated the number of L1 trials presented as input prior to initial exposure to an L2. There were 15 levels in their entrenchment factor, ranging from no entrenchment period (t=0), also referred to as the native bilingual condition, to a maximal entrenchment period (t=900,000), also referred to as the late L2 learner condition. The number of trials presented to the model post-entrenchment (i.e., the bilingual phase) was

always 2,000,000. During the bilingual phase, the L1 or L2 input was randomly selected with equal probability. The selection of a particular sequence of phonemes was determined by a frequency metric calculated by the authors.

The second factor, memory development, manipulated the architecture of the neural networks. Memory units within the hidden layer(s) and/or their corresponding connections to other units within the layers were added or reset periodically during the first 400,000 trials. The manipulation of the hidden layer simulated different developmental conditions of both the working memory and the long-term memory systems. The purpose of this factor was to test the less-is-more hypothesis (Newport, 1988, 1990). There were four levels in this condition: *No Growth*, *Unit Growth*, *Unit Replacement*, *Connection Growth*.

The No Growth condition represented a network that had a fully developed network architecture, something conceptually similar to a fully developed human adult. In the No Growth condition, the number of units in the hidden layer(s) was not manipulated during training, and the connections between layers remained fully connected throughout the training phases. This condition represented a fully developed network with a constant amount of working memory and long-term memory capacity throughout the learning period.

The Unit Growth condition started with an undeveloped network that periodically added units and any associated connections throughout training until the hidden layer(s) of the network had the same number of units and connections as found in the No Growth condition (i.e., the network reached maturity after 400,000 trials). This condition manipulated the development of both working memory ca-

pacity and long-term memory capacity during the initial stages of language learning. Monner et al. (2013) argue that the growth of the network was meant to represent the *recruitment* of new resources, something similar to dendritic outgrowth (Uylings, 2006), not the *creation* of new memory resources.

The Unit Growth condition confounds the variables of working memory capacity and long-term memory capacity. As new units are added to the network, more information can be processed by the network, therefore increasing the capacity of the network. This is argued to be akin to a working memory capacity in that more information can be processed at a given instant. Adding a new unit increases processing capacity and leads to more connections to and from the new units. The addition of more connections increases the storage capacity of the network, which is similar to a long-term memory store. In order to isolate any potential effects of this factor, two additional memory conditions were added.

The Unit Replacement condition randomly reset units within the hidden layer and their associated connections periodically during the first 400,000 trials. In this condition, the working memory and long-term memory capacities remain unchanged; however, the weights of the connections and the unit state were reset periodically during training. This condition was not intended to represent a specific state of actual human development. Instead, it was included to separate the effects of starting small from those of adding new and untrained resources. This condition randomly forgets so that new learning can take place. Similar to the Unit Growth condition, new resources are added over time, but in the Unit Replacement condition, the capacity of the working memory system (i.e., the size of the hidden layer(s)) remains

unchanged. By setting up the condition in this way, it is possible to determine if providing new untrained resources, without starting small, provide benefits throughout the learning process.

The final condition, Connection Growth, began training with a sparsely connected network. Connections were added periodically over the first 400,000 trials. This condition manipulated the size of the long-term memory stores while controlling for working memory capacity. Some networks in the Connection Growth and Unit Growth conditions were exposed to L2 input during their most immature state, while other networks in these conditions were exposed to the L2 well after reaching a fully mature state. Understanding how the timing of initial L2 exposure affects learning outcomes was stated as being one of the major objectives of Monner et al. (2013).

Monner et al. (2013) made three explicit predictions with regards to the performance of their models across the different conditions. First, performance on the L1 should not decrease as entrenchment increases. Second, models in the native bilingual condition (t=0) should perform equivalent to their respective monolingual baselines. Third, final L2 performance should decrease as entrenchment increases. Monner et al. (2013) did not formalize a prediction for the different memory development conditions, but their discussion of the less-is-more hypothesis implied that they expected that conditions that start small and grow (i.e., Connection Growth and Unit Growth) would help alleviate the negative effects of entrenchment if the L2 is presented within the period of memory development (before 400,000 trials).

The statistical analysis of both experiments used a two-proportion z-test to

compare the difference in L2 learning outcomes due to entrenchment. A separate z-test was conducted for each level of the memory development factor for both the L1 and the L2 outcome measures in the two language pairs. The accuracy values for the native bilingual entrenchment level (no entrenchment; $t = 0$) was compared to the late L2 learner entrenchment level (maximum entrenhcment; $t = 900{,}000$).

In both the gender assignment and agreement experiments, L1 performance did not decrease as a function of increasing levels of entrenchment (except in the Connection Growth condition for L1 Spanish in the gender agreement experiment). Native bilingual networks also reportedly performed on par with monolingual networks in both experiments, suggesting that native bilingual networks perform about the same as monolingual networks in either of the languages. In the gender assignment experiment, entrenchment negatively influenced L2 learning outcomes, but only in the L1 Spanish - L2 French language pair. In the L1 Spanish - L2 French pair, the Unit Growth condition led to a modest increase in performance ($1\% > x < 3\%$). There were no entrenchment effects in the L1 French - L2 Spanish language pair in the gender assignment experiment, but a modest increase in performance ($1\% > x < 3\%$) in the Unit Growth condition.

The authors argued that a null effect of entrenchment in the L1 French - L2 Spanish language pair may be due to the relative regularity of the Spanish gender class system. In other words, mapping Spanish nouns to gender class was so easy that ceiling effects affected any potential effects of the variables. Certain final phonemes in Spanish are highly predictive of gender, whereas French final phonemes are less predictive. Another explanation could be that certain patterns learned

during the L1 French training phase may have facilitated L2 Spanish performance, but not in the other direction. Of course, the two explanations are not mutually exclusive. In the gender agreement experiment, a harder learning task, entrenchment negatively influenced L2 learning outcomes in all language pairs. The magnitude of this negative impact was substantial ($> 12\%$ for the L1 Spanish - L2 French language pair and $> 3\%$ for the L1 French - L2 Spanish language pair). The discrepancy between the two experiments suggests that learning task complexity and/or formal cue regularity may have played a role in the null effects in the gender assignment experiment in the L1 French - L2 Spanish language pair.

Interpreting the influence of memory development on learning outcomes was more complicated due to the way in which the data was analyzed. The two-proportion z-test compared the mean accuracy on the outcome measures between the minimum (t=0, no entrenchment or native bilingual condition) and maximum entrenchment (t=900,000, late L2 learner condition) levels. This analysis can only identify if there is a difference in outcomes between a scenario with no entrenchment versus one with a lot of entrenchment. The analysis did not look at how the effect of entrenchment unfolded over the different levels, and whether models exposed to L2 input within the developmental period ($<$ 400,000 trials) were advantaged by starting small.

In the gender assignment experiment, it was difficult to evaluate memory development effects in the L1 French - L2 Spanish language pair due to the ease with which models learned gender assignment in Spanish. Both language pairs in the gender assignment experiment had a modest increase in performance ($1\% > x <$

3%) in the Unit Growth condition. Only the L1 Spanish - L2 French pair had a modest increase in accuracy due to the Connection Growth condition. In the gender agreement experiment, starting with fewer memory resources mitigated the negative impact entrenchment had on L2 learning outcomes only in the Unit Growth condition of the language pairs.

All of the results reported in Monner et al. (2013) were derived from an evaluation on the data used to train the models at the end of the bilingual training phase. A held-out test set was not used (personal correspondence with the first author). The stated reason for evaluating performance on the training data was that the authors were primarily interested in identifying how effectively the model could internalize patterns on words it had already seen. High-frequency patterns that are presented as training data to neural networks are often overlearned (Marchman, 1993). Generally, the field of machine learning regards overlearning as problematic because it hinders learning that is generalizable to a diverse set of examples. Basically, the process of overlearning leads to memorized training examples and potentially interferes with the process of learning generalizable patterns. The ceiling effects seen in Spanish in the gender assignment experiment may partially be the result of evaluating the model on the same data used to train the model.

1.5 Thesis

The objective of this work is to understand how within- and between-language characteristics affect L2 learning outcomes when the factors of linguistic entrench-

ment and memory development are manipulated. The work presented here expands upon previous research by modeling language learning in a variety of language pairs across three experiments addressing the following linguistic sub-tasks: grammatical gender assignment, grammatical gender agreement, and word boundary identification. In each experiment, recurrent neural networks with an LSTM architecture were trained (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002) to perform these tasks in two languages. Entrenchment was manipulated by delaying the moment networks were initially exposed to the L2 input. Memory development, on the other hand, was manipulated during the early stages of learning within the architecture of the neural network.

The first experiment trained models to assign nouns in French, German, Russian, and Spanish to a specific grammatical gender class using phonologically encoded input. Monner et al. (2013) did not find consistent effects of entrenchment across the French and Spanish language pairs, suggesting that the regularity of the input may influence the effect of certain developmental variables. By including two additional languages, this experiment addressed whether the characteristics of the grammatical gender system affect how L1 entrenchment and memory development influence L2 learning outcomes.

In the second experiment, models were trained to predict the next phoneme in a sequence of phonemes in English, French, Russian, and Spanish. Performance was evaluated on an intrinsic task and an extrinsic task. The extrinsic task, called gender agreement, was conceptually similar to the one used in Monner et al. (2013). This experiment evaluated how language-specific characteristics influence the impact

entrenchment and memory development have on a task considerably more difficult than gender assignment.

The third experiment trained models to predict whether the current phoneme within a sequence represented the end of a word (i.e., a word boundary) in English, French, Russian, and Spanish. The binary nature of word boundary identification makes it possible to control the number of possible outputs produced by a neural network. A diverse set of languages can be modeled while maintaining the same output space. This was not possible in the gender assignment and gender agreement experiments. In the gender assignment experiment, the output space was dependent upon the number of gender classes in the language, while the output space in the gender agreement experiment was dependent upon the phonemes present in a particular language. This experiment ensures input in each language maps to each unit in the output layer. The central question in this experiment is whether having a shared output space between languages affects how L1 entrenchment and memory development influence L2 learning outcomes.

These experiments will be presented in the next three chapters. The structure of each chapter is roughly the same; introduce the research questions, describe the linguistic data and methods used to train the neural networks, and present the analysis of the results. A synthesis of the findings from all three experiments will be presented in the last chapter.

Chapter 2: Gender Assignment

Grammatical gender is a linguistic system that assigns nouns to a particular class following semantic and/or formal (i.e., morphological and/or phonological) principles (Comrie, 1999; Corbett, 2006). Many languages with gender systems utilize either semantic information or a combination of semantic and formal information to assign gender to a noun. No grammatical system, however, solely uses formal principles (Corbett, 2006). Nouns that use semantic features for classification often occur in animate entities where there is a transparent relationship between the biological sex of the referents and the gender class of the noun (Barber & Carreiras, 2005). Formal assignment principles, on the other hand, depend on the form of the nouns instead of their meaning.

Languages differ in the number of classes as well as the formal and semantic transparency (i.e., regularity) of cues to determine class. The number of gender classes in languages can range from two, like that found in Romance languages, to the 17 or more found in several West African languages (Sá-Leite, Fraga, & Comesaña, 2019). The statistical-phonological regularities relating to specific gender classes in a language determine the formal transparency of the gender system. These statistical regularities make learning the mapping between nouns and gender classes easier.

However, not all nouns fit into a regular pattern. There are numerous examples of nouns that do not follow the typical patterns in a language (e.g., *la radio* in Spanish). This subset of nouns can be quite large. For example, certain word endings in French only predict gender class 40-60% of the time. Under these circumstances, the gender class associated with these nouns must be memorized by the learner.

The function of grammatical gender is often considered to be syntactic glue that helps determine the form of other words during sentence construction. In fact, Corbett (2007) only considers a language to have a grammatical gender system if the gender class is used to inform verb and/or adjective agreement. Language users show facilitation effects along different outcome measures when words encoded with grammatical gender cues are present in the input (Holmes & de la Bâtie, 1999; Vatz, 2009).

Studies in multiple languages show that late L2 learners process the gender of nouns slower and are less accurate than native speakers in French (Vatz, 2009; Guillelmon & Grosjean, 2001), German (Scherag, Demuth, Rösler, Neville, & Röder, 2004) and Spanish (Lew-Williams & Fernald, 2010). There is even evidence that children that began bilingual immersion education around the age of six do not perform like native French speakers five or six years later when they are around ten years old (Harley, 1979; Lapkin & Swain, 1977), suggesting that there may be a sensitive period for learning the underlying statistical properties in grammatical gender assignment. Neuropsychological evidence using the Event-Related Potential method indicates that grammatical gender is processed differently by native and non-native speakers (Foucart & Frenck-Mestre, 2011). These differences in performance be-

tween native and non-native speakers of a language have led some to suggest that knowledge of the grammatical gender systems may be stored differently between native and non-native speakers and that this may affect how individuals process language data. Vatz (2009) found considerable differences in performance on gender assignment tasks between native speakers of French and non-native speakers from several different L1 backgrounds. Contrary to her hypothesis, L1 Spanish speakers did not experience positive transfer effects. That is, knowledge of a similar grammatical gender system did not provide any knowledge about the L2 gender system for free. A previous study by Sabourin and Stowe (2008) found evidence of positive transfer on a grammatical judgment task that includes gender agreement as a component. These results suggest that native and non-native speakers represent grammatical gender knowledge differently and/or process gender differently.

In the experiment presented here, nouns in French, German, Russian, and Spanish were used to train recurrent neural networks. The grammatical gender systems of these languages are distinct and vary in the number of gender classes assigned to nouns as well as the formal cues available to aid language users in the assignment of gender class to nouns. Monner et al.(2013) suggested that the regularity of formal cues (either phonological or morphological) may have influenced their results in the gender assignment experiment. Before the linguistic data and computational methods are presented, the grammatical gender system of each language is described briefly.

In French, nouns have two classes: feminine and masculine. The French gender system is less regular than other language gender systems. In French, the ending

phonemes often provide cues to the gender class. For example, only the /z/ phoneme and the /ɛ̃/, /ã/, /ø/, /o/, /ʒ/, /m/, and /ɛ/ phonemes indicate feminine and masculine grammatical gender class, respectively, in 90% of the scenarios (Surridge, 1993, 1995). Some final phonemes, such as /e/, /l/, /p/, /t/, provide no systematic clue to the gender of the noun (Surridge, 1995). Cues provided by final phonemes are not as reliable as they are in the Russian and Spanish systems.

In German, nouns have three classes: feminine, masculine, and neuter. The German gender system has the least regular system of the languages used in the present study. However, there are some phonemic endings that often provide cues to the gender class of the noun. For example, nouns ending in /hait/ (heit), /kait/ (keit), /uŋk/ (ung) are often feminine, those ending in /xən/ (chen), and /lain/ (lein) are neuter, and those ending in /ə/ (er) are masculine (Steinmetz, 1986). German linguistic input was only used for the gender assignment experiment.

Like German, Russian nouns have three classes: feminine, masculine, and neuter. In Russian, most nouns in the nominative form provide transparent cues as to the class of the noun. Feminine nouns consistently end in a, masculine nouns typically end in a consonant, and most neuter nouns end in o. Neuter nouns always follow a semantic and/or formal rule; however, some feminine and masculine nouns do not follow a predictable pattern. These irregular feminine and masculine nouns often end in palatalized consonants. Regarding the symmetric nature of declension and gender, Corbett (1982) notes that it is often possible to use the declension class or the gender class to predict the other.

Like French, Spanish only has two gender classes: feminine and masculine.

Spanish has a highly regular gender system. In Spanish, the grammatical gender of the majority of nouns are determined by their final phoneme (see Teschner & Russell, 1984). Nouns ending in /a/ and /d/ are nearly always feminine, while those ending in /e/, /l/, /o/, /ɾ/, /i/, /m/, /t/, /u/, /x/, /y/, /b/, /c/ and /tʃ/ are usually masculine. Several final phonemes are ambiguous: /n/, /z/, /s/; however, these constitute a small portion of nouns. Morphological endings, such as ción (/sion/), gión (/xion/), nión (/njon/), sión (/sjon/), tión (/tjon/), xión (/sjon/), and ez (/es/), are typically feminine, whereas morphological endings of ón (/on/), az (/as/), oz (/os/), uz (/us/) are typically masculine.

2.1 Research Questions

The results of the two experiments in Monner et al.(2013) were not consistent across language pairs and experimental tasks. Linguistic entrenchment was found to affect L2 learning outcomes in both studies negatively, but failed to do so in the L1 French - L2 Spanish language pair of the gender assignment experiment. The authors noted that the formal gender cues associated with Spanish nouns are more consistent than those in French. They argued that Spanish cues are more transparent and therefore, may have been so easy that in all entrenchment and memory development conditions learning easily reached ceiling level performance within the provided training time. To understand the results of Monner et al.(2013) better, recurrent neural networks representing language learning agents learned to classify nouns in French, German, Russian, and Spanish according to their appropriate gen-

der class. Like the experiments in Monner et al.(2013), linguistic entrenchment and memory development are manipulated in order to identify how these variables impact L2 learning outcomes. Based upon previous research, especially the two studies by Monner et al. (2013), several hypotheses were generated.

First, linguistic entrenchment is expected to have an overall negative effect on L2 learning outcomes; however, this effect is expected to be small (see results in Monner et al., 2013). Starting the learning process with a smaller working memory capacity is not expected to mute the negative effects of entrenchment, nor is it expected to lead to better L2 learning outcomes (for a similar view, see Brooks & Kempe, 2019). In fact, it is expected that models starting small will consistently underperform those beginning with more resources. Supporting this hypothesis is research by Rohde and Plaut (1999), which shows that fewer memory resources actually hinder learning performance in neural networks with a similar architecture (for a different result, see Elman, 1990). Even Monner et al. (2013) did not find consistent evidence in support of the less-is-more hypothesis. Also, adult L2 learners with a greater working memory capacity consistently show modest advantages over those with a smaller capacity on many learning outcomes (Linck et al., 2014). Third, it is expected that models with an L2 that is high in formal cue regularity (e.g., French, Russian and Spanish) will have better L2 learning outcomes than those with low levels of formal cue regularity (e.g., German). Specifically, Spanish, the language with the most formal cue regularity, is expected to outperform both French and Russian, two languages which also have high levels of formal cue regularity (see Monner et al., 2013).

The analysis performed in Monner et al.(2013), which was a series of two-proportion z-tests, did not account for all of the data collected over the different entrenchment levels and also failed to take into account any random effects across the factors of entrenchment and memory development due to individual model variability. This experiment uses mixed effects regression models to more appropriately model the structure of the dataset.

2.2 Method

Recurrent neural networks using an LSTM architecture (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002) were trained to learn the grammatical gender class of nouns in two languages. Four different languages were paired (French, German, Russian, Spanish), creating a total of 12 unique language pairs. Two developmental variables hypothesized to influence L2 learning outcomes were manipulated: linguistic entrenchment and memory development. The model learned solely from encoded phonological features of the nouns; therefore, only the formal cues for grammatical gender in each language can be learned by the models. The inclusion of nouns from French, German, Russian, and Spanish allows for the investigation of the influence of other language-specific factors on learning outcomes, like the number of gender classes and the saliency of formal cues. All corpus pre-processing tasks and model training was accomplished with the Python programming language (van Rossum, 1995).

2.2.1 Long Short-Term Memory Architecture

The computational models used in the present work are recurrent neural networks with an LSTM architecture (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002). Other networks, like the multi-layer perceptron (MLP), maintain information across trials in the connection weight between units. This is similar to long-term memory stores. However, MLPs are not able to pass information across steps in a sequence. Recurrent neural networks are able to do this and are therefore often used to model sequential data, a defining characteristic of linguistic data. This ability to maintain information across time is conceptually similar to the concept of working memory. Information can be selectively maintained in a state of active memory in order to be used at a later time.

The simple recurrent network (SRN; Elman, 1990) is one variant of recurrent neural networks. The LSTM is another variant and is currently considered state of the art in terms of empirical results. The LSTM architecture contains recurrently connected units, often referred to as memory cell block assemblies. Each memory block contains at least one self-connected memory cell (c) and three gating units (input gate (ι), forget gate (φ), output gate (ω)) that function as update, delete and read operations for the cells within the memory block (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002). The input and output gates control the flow of information coming into and leaving the memory cell, while the forget gate controls how much of the memory cell state is passed to the next time-step. A diagram of the LSTM architecture is provided in Figure 2.1. This diagram illustrates how new

information is combined with the previous output of the LSTM cell to update the state of the cell. The A verbal description of the flow of information through the forward pass is provided below. All equations provided below were adapted from Graves (2013).

Input (x_t) that enters the memory cell block is multiplied along weighted connections ($W_{i,}$) to the input gate (ι_t), the forget gate (φ_t), the internal memory cell state (c_t), and the output gate (ω_t). Simultaneously, the previous hidden state (h_t) of the memory cell block, is also multiplied by its weighted connections ($W_{h,}$) to the input gate, forget gate, internal memory cell state, and the output gate. For each gate, the net input of the information sources is summed and then squashed using the sigmoid function[1] (see Equations 2.3, 2.4, 2.6 for the input, forget, and output gates, and 2.1 for the sigmoid function). For the internal memory cell state, the net input is summed and squashed using the hyperbolic tangent function and then the product of the previous cell state and the squashed forget gate is added to the sum (see Equation 2.5 for the internal memory cell and 2.2 for the hyperbolic tangent function). The hidden state of the memory cell block is calculated by multiplying the squashed value of the output gate with the output of the cell state after being squashed by the hyperbolic tangent function (see Equation 2.7).

[1]The purpose of squashing functions (also referred to as activation functions) is to manage the scale of the unit activations. In the case of the sigmoid function, all input values are squashed to be between 0 and 1. The hyperbolic tangent function, on the other hand, squashes input values between -1 and 1.

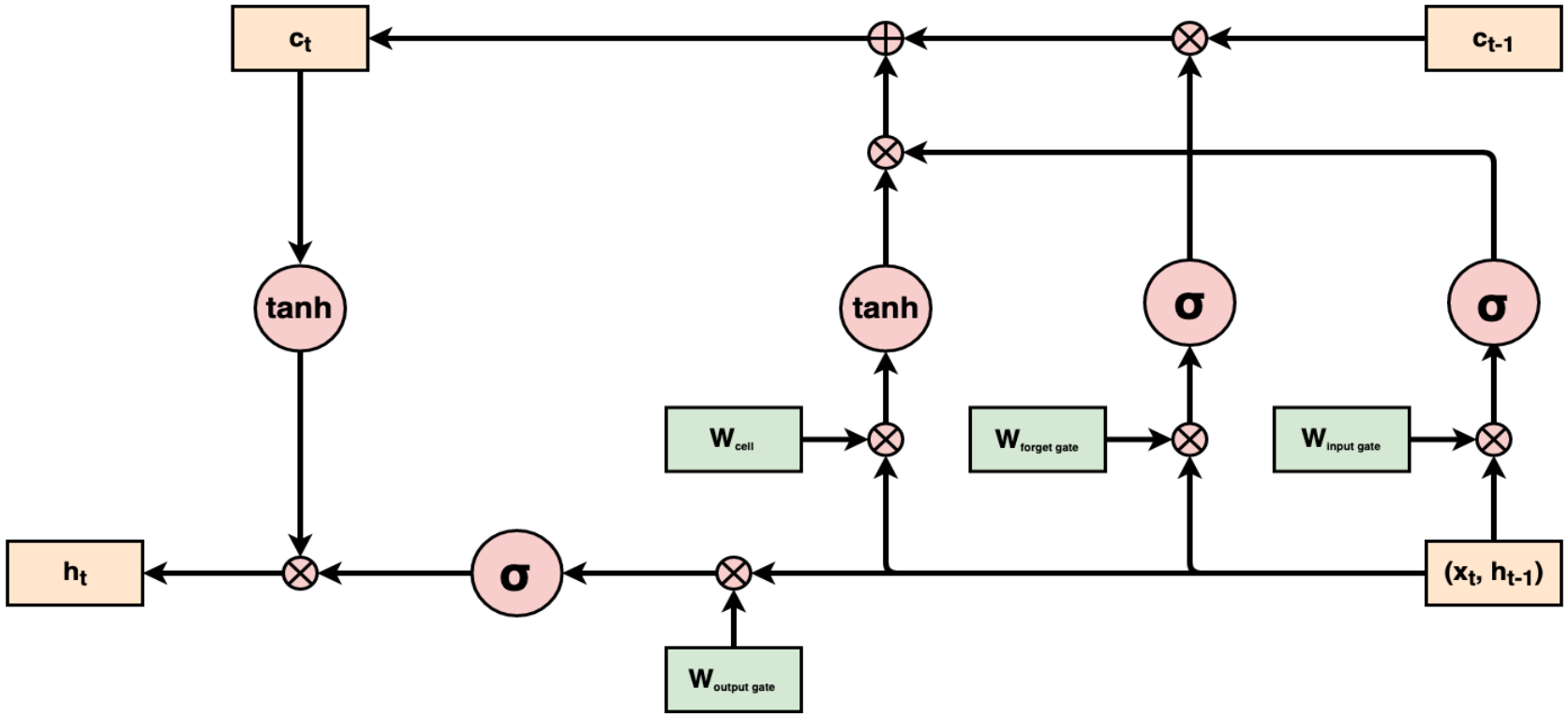

Figure 2.1. Diagram illustrating the flow of information through an LSTM unit. The σ circles indicate the use of the sigmoid squashing function while the tanh circles indicate the use of the hyperbolic tangent squashing function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.2}$$

$$\iota_t = \sigma(W_{i\iota} x_t + b_{i\iota} + W_{h\iota} h_{t-1} + b_{h\iota}) \tag{2.3}$$

$$\varphi_t = \sigma(W_{i\varphi} x_t + b_{i\varphi} + W_{h\varphi} h_{t-1} + b_{h\varphi}) \tag{2.4}$$

$$c_t = \varphi_t * c_{t-1} + \iota_t * tanh(W_{ic} x_t + b_{ic} + W_{hc} h_{t-1} + b_{hc}) \tag{2.5}$$

$$\omega_t = \sigma(W_{i\omega} x_t + b_{i\omega} + W_{h\omega} h_{t-1} + b_{h\omega}) \tag{2.6}$$

$$h_t = \omega_t * tanh(c_t) \tag{2.7}$$

The structure of the LSTM architecture allows memory units to retain information over long sequences while avoiding problems like exploding and vanishing gradients, a problem found in other variations of recurrent networks (Graves & Schmidhuber, 2005). The LSTM architecture is flexible and can be altered to solve a wide range of problems. Monner et al. (2013) used a generalized version of the LSTM architecture (see Monner & Reggia, 2012) to model the effects of entrenchment and memory development in the learning of grammatical gender in French and

Spanish. The LSTM architecture in their paper represented a simplified substrate for working memory. Specifically, the number of memory cell block assemblies signified the working memory capacity of the network. Long-term memory, on the other hand, was represented by the connection weights between units.

Recurrent neural networks with an LSTM architecture were chosen to model language learners because they have robust domain-general mechanisms capable of retaining information over time, a capability possessed by all humans. Other recurrent neural network approaches, like the SRN, have difficulties modeling sequences that extend over many time-steps (e.g., the vanishing and exploding gradient problem). The LSTM architecture is capable of handling longer sequences, which is especially helpful when learning non-adjacent dependencies that can vary considerably in terms of the distance between linguistic elements. Unsupervised algorithms, like the SOMs used in the DevLex II models (Li et al., 2007; Zhao & Li, 2010; Li & Zhao, 2013), do not possess a domain-general mechanism capable of keeping information active over variable lengths of time. For these reasons, the LSTM architecture was chosen to model the learning process.

2.2.2 Nouns

All data used to train and test the models in the gender assignment experiment was a noun. This differs from the data used to train the gender assignment models in Monner et al. (2013), which attached gender-appropriate determiners to the beginning on the noun sequence. Although determiners often accompany nouns,

it is not required that they accompany an associated noun. The consistent use of gender-appropriate determiners is argued to be a distraction from the primary purpose of training a model to recognize the class of a noun based solely upon the phonemic sequence of the noun. For this reason, determiners were not paired with an associated noun to train the models.

The nouns came from two sources. The French, Russian, and Spanish nouns all came from the *United Nations Parallel Corpus v1.0* (UNPC; Ziemski, Junczys-Dowmunt, & Pouliquen, 2016) dataset [2]. Since German was not part of the aligned UNPC, German nouns were gathered using the German lexical frequency dictionary, *SUBTLEX-DE*, which was generated from a large database of movie subtitles in German (Brysbaert et al., 2011).

In order to identify nouns for possible inclusion in the training and test sets, language-specific part-of-speech tagging models using a recurrent neural network architecture (RNNTagger; Schmid, 2019) were applied to each of the corpora in French, German, Russian and Spanish. Each part-of-speech tagging model produced a part-of-speech tag and a lemma (i.e., dictionary form) for each word. These part-of-speech tags were used to generate a list of unique nouns for each language. Lexical frequency dictionaries were used to select eligible nouns for the training and test sets (Brysbaert et al., 2011; Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011; New,

[2]The UNCP dataset contains manually translated United Nations (UN) documents aligned across all six official UN languages from 1990 to 2014: Modern Standard Arabic, Chinese, English, French, Russian, Spanish. The aligned corpus is comprised of 11,365,709 phrases manually translated from 86,307 documents for each of the six languages.

Pallier, Brysbaert, & Ferrand, 2004; Sharoff, 2002). If a given noun was present in the language-specific lexical frequency dictionary, it was set aside for possible inclusion in either the training or test set.

Only one noun from a lemma was allowed in either the training or the test set. That is, once a lemma was used in the training set, another noun with the same lemma was ineligible for inclusion in either the training set or the test set. This ensured that each noun in the sets was associated with only one lemma. This is especially important for Russian, a morphologically rich language. Since language is a stochastic process, each noun was associated with the lexical frequency (i.e., a weight) from the appropriate lexical frequency dictionary. Nouns with a higher frequency had a higher weight and were more likely to be selected as input during the training phase (for similar approaches, see Seidenberg & McClelland, 1989; Zevin & Seidenberg, 2002; Monner et al., 2013).

In order to ensure exact phonological cognates were avoided, only unique IPA representations of nouns were eligible for inclusion across the training and test sets. To determine this, the orthographic text of each noun was transliterated into phonemic segments represented by IPA symbols. The Python programming language package *Epitran*[3][4] (Mortensen et al., 2016) was used to convert orthographic text into phonemic segments represented by the IPA. If the IPA representation of the noun in one language was identical to that of a noun in another language, the noun was only selected for inclusion in one language dataset.

[3]https://github.com/dmort27/epitran

[4]The transcription approach taken in the Epitran package is morphophonemic.

One thousand eight hundred nouns in each language were selected for inclusion in either the training or test sets using the eligibility criteria described above. One thousand two hundred nouns in each language were selected for the training set, and 600 nouns in each language were selected for the test set. There was an equal distribution of nouns by class in each training and test set (see Table 2.1).

All models in this experiment were evaluated on a held-out test set. Monner et al.(2013) did not use a test set comprised of unseen nouns to evaluate the performance of their models. Instead, they evaluated the models with the same data that was used to train it. In personal correspondence with the first author of Monner et al. (2013), it was stated that the authors were primarily interested in how effectively the model could internalize patterns on words it had already seen. This is tantamount to evaluating how well the models memorized the dataset. In contrast with the methodology of Monner et al. (2013), this experiment deliberately used a held-out test set to evaluate the extent to which models learn generalizable patterns in the phonological representation of nouns.

2.2.3 Transforming Nouns into Articulatory Feature Vectors

Once the nouns for the training and test sets were identified, each noun was converted into a sequence of articulatory feature vectors and assigned to a unique gender class. The method used here to convert nouns into articulatory feature vectors conceptually follows the method used in Monner et al.(2013).

First, the orthographic form of each noun was converted to IPA segments

Table 2.1

Means (standard deviations) of orthographic length, IPA length, and lexical frequency for each gender class across the four languages of the gender assignment experiment.

	Train				*Test*			
	N	*Ortho*	*IPA*	*Freq*	*N*	*Ortho*	*IPA*	*Freq*
French								
f	600	9.23 (2.55)	8.69 (2.82)	0.85 (0.75)	300	9.00 (2.74)	8.58 (2.83)	0.88 (0.80)
m	600	7.98 (2.68)	7.31 (2.54)	0.94 (0.80)	300	7.89 (2.63)	7.09 (2.41)	0.98 (0.80)
German								
f	400	7.88 (2.76)	8.31 (3.12)	0.68 (0.66)	200	7.81 (2.71)	8.04 (3.02)	0.59 (0.61)
m	400	7.03 (2.33)	6.84 (2.44)	0.81 (0.68)	200	6.94 (2.53)	6.95 (2.82)	0.69 (0.65)
n	400	7.32 (2.72)	7.38 (2.87)	0.78 (0.73)	200	7.20 (2.69)	7.18 (2.89)	0.82 (0.79)
Russian								
f	400	7.62 (2.78)	9.10 (3.50)	0.59 (0.49)	200	7.72 (2.66)	9.28 (3.36)	0.59 (0.48)
m	400	7.30 (2.55)	8.39 (3.15)	0.55 (0.45)	200	7.10 (2.57)	8.12 (3.15)	0.62 (0.49)
n	400	9.25 (2.70)	10.5 (3.16)	0.57 (0.47)	200	9.40 (2.76)	10.6 (3.16)	0.58 (0.48)
Spanish								
f	600	9.25 (2.71)	9.23 (2.74)	0.63 (0.66)	300	9.21 (2.66)	9.19 (2.68)	0.61 (0.63)
m	600	8.18 (2.46)	8.12 (2.47)	0.75 (0.69)	300	8.16 (2.37)	8.09 (2.40)	0.69 (0.68)

representing individual phonemes. Then, each phonemic segment of the noun was mapped to a vector of 22 articulatory features. The transliteration of text to phonemic segments and the subsequent encoding of segments as feature vectors was aided by two computational tools developed at Carnegie Mellon University in Pittsburgh, Pennsylvania USA (Mortensen et al., 2016). Epitran, the Python package mentioned above, converts orthographic text into phonemic segments represented by the IPA, and *PanPhon*[5] maps each segment to a unique articulatory feature vector (see Table 2.2 for a description of each articulatory feature). The PanPhon package produces ternary feature vectors. These ternary feature vectors were converted into binary feature vectors by converting values with a -1 to a 0.

As an example, the Spanish noun, casa (house), is first converted to phonemic segments represented with IPA graphemes, like /kasa/. Next, each phoneme is

[5]https://github.com/dmort27/panphon

Table 2.2

Description of articulatory features used to encode phonemes.

Feature	Description
Syllabic	The segment is the nucleus of the syllable.
Sonorant	Produced with continuous, non-turbulent airflow in the vocal tract.
Consonantal	An audible constriction of the vocal tract.
Continuant	A nearly complete constriction of the vocal tract.
Delayed Release	A pause before the release of the vocal tract
Lateral	Produced by stopping airflow around the mid-section of the vocal tract.
Nasal	Produced by lowering the velum and releasing air through the nose.
Strident	A sound with high frequency frication, similar to white noise.
Voice	A sound produced with vocal fold vibration.
Spread Glottis	An opening of the vocal tract
Constricted Glottis	A closing of the vocal tract.
Anterior	Obstructing the front of the palateo-alveolar region of the mouth.
Coronal	Produced by raising the blade of the tongue from the neutral position.
Distributed	Produced by constricting a distance following the airflow.
Labial	Produced by constricting the lips.
High	Produced with a raised tongue body.
Low	Produced with a lowered tongue body.
Back	Produced with a retracted tongue body.
Round	Produced by narrowing the orifice of the lips.
Velaric	Produced by using the back of tongue to seal off air from the lungs.
Tense	Produced by elongating a vowel sound.
Long	Produced by elongating a consonant sound.

mapped to its corresponding articulatory feature vector to create a two-dimensional vector (e.g., 4 x 22; see Figure 2.2).

Articulatory features represent aspects of how humans physiologically produce speech sounds. The Spanish word casa can be represented using IPA phonemes as /kasa/. The /k/ sound is made by raising the body of the tongue to the roof of the mouth and moving the tongue body slightly to the back of the mouth to constrict the vocal tract. The release of the tongue then produces the intended sound. This is a verbal representation of what occurs within the human vocal tract. The vector representation of this sound can be encoded as a series of ones (1) and zeros (0), like in the following vector: [00100000000000001010000]. In this vector, the consonantal, high, and back articulatory features are turned on, while the other features are deactivated. This results in a multi-label encoded representation of a phoneme.

/k/ [00100000000000001010000]
/a/ [11010000100000000110010]
/s/ [00110000000011000000000]
/a/ [11010000100000000110010]

Figure 2.2. Articulatory feature vectors representing the phoneme segments of /kasa/.

Encoding phonemes with articulatory features provides more data to the training algorithm than would otherwise be available if the phonemes were one-hot encoded (i.e., all elements of the vector are set to zero except for one). One-hot encoding is a technique used to represent a non-numeric entity (i.e., categorical) as a numerical vector. The distributed representational nature of the densely encoded data often leads to better discriminating models that learn faster than their one-hot encoded counterparts. Although encoding phonemes with articulatory features is arguably more linguistically realistic than simply one-hot encoding phonemes, this representation is still a gross oversimplification of actual speech encountered by learners. Speech produced by humans is much more variable and often accompanied by other extraneous sounds not relevant to the linguistic task. These extraneous sounds can place additional strain on the cognitive system, ultimately distracting the speaker from attending to the signal of interest. The representation of speech as articulatory features ensures the models are provided with unbiased and ideal representations of the phonemes not sullied by extraneous acoustic noise.

2.2.4 Training Procedure

Recurrent neural networks (see Table 2.3 and Figure 2.3) representing language learning agents were trained to classify a noun by its gender class using a sequence of articulatory feature vectors representing phonemes. Table 2.3 provides details on the specific architecture and hyperparameters used to train the models in this experiment. The selection of hyperparameters for this experiment was largely influenced by previous studies, especially the gender assignment experiment by Monner et al. (2013). The learning rate was chosen so that learning could happen quickly and reach peak performance relatively early during the training phase. It was important that the models reached peak performance on the learning task during the training phase. The objective of this experiment was not to optimize the performance of a model on a particular outcome; instead, the objective was to show how performance is relative to linguistic and developmental factors. The selection of specific hyperparameters was not determined in a systematic fashion.

Previous studies trained models with only one input sequence in a sequential manner (see Monner et al., 2013); however, this approach leads to very long training times. In order to speed up the training of each model, mini-batches were created. The size of the mini-batch was important so that the model avoiding getting stuck in a local minimum during the training process. For each step during training, mini-batches of ten nouns were selected as input to the model. Each noun was represented as a sequence of vectors consisting of 22 articulatory features representing unique phonemes. The ten nouns were selected randomly from the list of weighted nouns

in the training set. After all vectors propagated through the model, the model produced ten output vectors representing the three possible gender classes (e.g., feminine, masculine, neuter) for each noun. Figure 2.4 illustrates how phonemes (x_t) at each time-step were passed into the network and then accumulated to generate a prediction (y) on the gender class of the input.

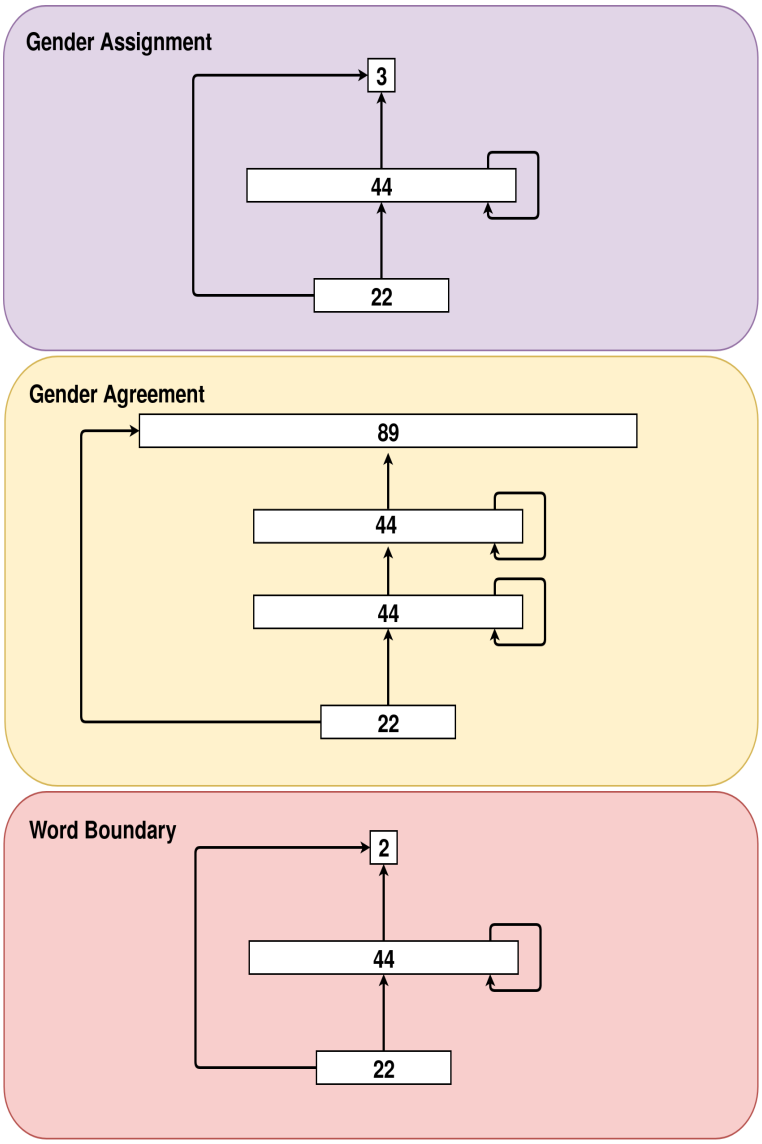

Figure 2.3. Diagram of the network architectures of the gender assignment, gender agreement, and word boundary experiments.

Table 2.3

Details of the neural network architectures and hyperparameters for the gender assignment, gender agreement, and word boundary experiments.

	Gender Assignment	*Gender Agreement*	*Word Boundary*
Architecture	LSTM	LSTM	LSTM
Input Layer Size	22	22	22
Hidden Layer Size	44	44	44
Output Layer Size	3	89	2
Number of Hidden Layers	1	2	1
Learning Rate	0.01	0.1	0.1
Momentum	0.8	0.8	0.8
Batch Size	10	10	10
Loss Criterion	Cross Entropy	Cross Entropy	Cross Entropy
Optimizer	SGD	SGD	SGD

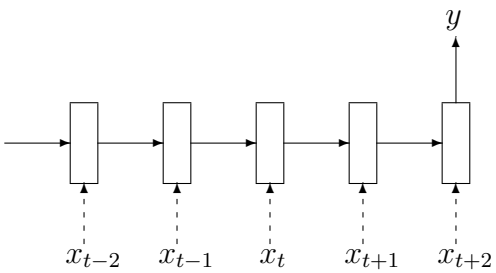

Figure 2.4. One-layer recurrent neural network that maps a sequence to one output (many-to-one).

Even though French and Spanish only have two gender classes, the output layer size was consistently held at three regardless of the number of gender classes possible given the language pair. The mean cross-entropy loss between the ten output vectors and the ten true target vectors was used during the backpropagation step of the training phase. The cross-entropy loss function (see Equation 2.8) measures the number of bits required to explain the difference between the estimated distribution ($\hat{y}$) and the true distribution (y). To perform this calculation, the activation values in the output layer are converted into a unit vector via the softmax function (see

Equation 2.9). The softmax function normalizes a vector so that the sum of all components adds up to 1. Therefore, each component represents a probability. The normalized vector is then compared to the true vector associated with the input. The loss value from the cross-entropy loss function was used to calculate the gradient for each parameter in the model. The stochastic gradient descent (SGD) learning method was followed throughout the training process. The open-source machine learning Python library *PyTorch* (version 1.1.0) was used to train all models (Paszke et al., 2017).

$$H(y, \hat{y}) = -\sum_{i=1}^{n} y_i \; log(\hat{y}_i) \tag{2.8}$$

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{2.9}$$

Model training was divided into two phases, a monolingual (i.e., entrenchment) phase and a bilingual (i.e., post-entrenchment) phase. The monolingual phase varied by the entrenchment level, while the bilingual phase always had a constant length of 2,000,000 nouns. The duration of the bilingual phase was kept constant to ensure each model would have enough time to reach ceiling performance in both languages regardless of the length of the entrenchment phase. Therefore, the total training time varied across the entrenchment levels.

Neural networks experience catastrophic forgetting when training data shifts distributional characteristics. In order to avoid catastrophic forgetting, an equal probability of exposure to either L1 or L2 linguistic input was ensured during the

bilingual phase. Although the interleaving of L1 and L2 input alleviates catastrophic forgetting in the network, it does not eliminate competition and inference between the two languages.

The monolingual phase represented the entrenchment factor, which manipulates the quantity of L1 input prior to the introduction of L2 input. There were six levels of L1 entrenchment (t): 0, 200,000, 400,000, 600,000, 800,000, 1,000,000. The first level of L1 entrenchment, t=0, represented a balanced (or native) bilingual network since both the L1 and L2 were present from the beginning of training. These models did not have any entrenchment. The last level of L1 entrenchment, t=1,000,000, represented a late L2 learner.

As in Monner et al. (2013), the monolingual and bilingual training phases were followed under four different memory development conditions: 1) No Growth, 2) Unit Growth, 3) Unit Replacement, 4) Connection Growth (see Figure 2.5). The goal of the memory development condition was to manipulate the architecture of the model in a way that conceptually mirrors memory development in humans. The No Growth condition represented a mature, fully developed language learner. The number of units in the hidden layer was not manipulated during training, and the connections between layers remained fully connected throughout the training phases. This condition represented a fully developed network with a constant amount of working memory and long-term memory capacity throughout the learning period.

The Unit Growth condition started with four units in the hidden layer and added one unit and any connections associated with that unit every 10,000 nouns throughout training until the network had the same number of units as found in the

No Growth condition (i.e., the network reaches 44 units after 400,000 nouns). This condition manipulated the development of both working memory capacity and long-term memory capacity during the initial stages of language learning. As described by Monner et al. (2013), the growth of the network is meant to represent the recruitment of new resources, like dendritic outgrowth (Uylings, 2006), not the creation of new memory resources.

The Unit Growth condition confounds the variables of working memory capacity and long-term memory capacity. As new units are added to the network, more information can be processed by the network, therefore increasing the capacity of the network. This is argued to be akin to a working memory capacity in that more information can be processed at a given instant. Adding a new unit not only increases processing capacity, it also leads to more connections to and from the new units. The addition of more connections increases the storage capacity of the network, which is similar to a long-term memory store. In order to isolate any potential effects of this factor, two additional memory conditions were added.

The Unit Replacement condition randomly reset a unit within the hidden layer and the associated connections every 10,000 nouns during the first 400,000 nouns. In this condition, the working memory and long-term memory capacities remain unchanged; however, the weights of the connections and the unit state were reset periodically (every 10,000 nouns). This condition separates the effects of starting small from the effects of adding new and untrained resources.

The final condition, Connection Growth, began training with a sparsely connected network. Only 10% of the connections were available initially. Every 10,000

nouns, 2.5% of these missing connections were added. This was done until a fully connected network was reached at 400,000 nouns. This condition had a constant working memory capacity throughout training but grew the long-term memory capacity at regular intervals throughout training. This condition was intended to help determine which aspect of memory development, working memory capacity or long-term memory stores, influences learning outcomes.

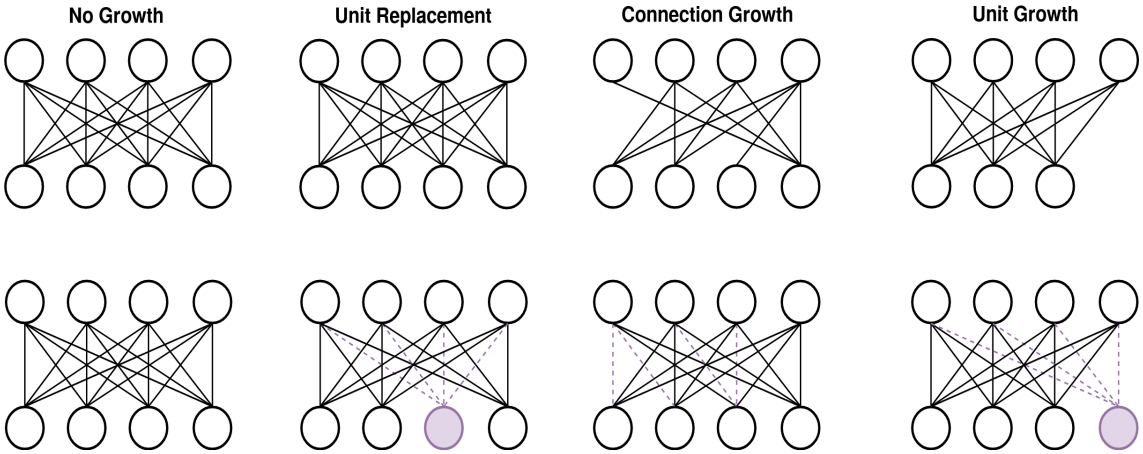

Figure 2.5. Visual depiction of the memory development factor levels.

Thirty networks were trained in each cell of the design matrix. Six levels of L1 entrenchment under four different memory development conditions yielded 720 models for each language pair. A total of 12 language pairs (L1 French - L2 German, L1 French - L2 Russian, L1 French - L2 Spanish, L1 German - L2 French, L1 German - L2 Russian, L1 German - L2 Spanish, L1 Russian - L2 French, L1 Russian - L2 German, L1 Russian - L2 Spanish, L1 Spanish - L2 French, L1 Spanish - L2 German, L1 Spanish - L2 Russian) led to 8,640 models across all conditions.

2.2.5 Evaluation Criteria

Each model was evaluated on the L1 and L2 test sets every 100,000 nouns. All formal analyses and reporting of descriptive statistics used the F1 score (Equation 2.13). The confusion matrix is a common way to illustrate the different ways performance can be evaluated on a classification task (see Table 2.4). Accuracy is perhaps the most common metric used; however, accuracy only takes into account true positive and true negatives (Equation 2.10). The F1 score is the harmonic mean of precision (Equation 2.11) and recall (Equation 2.12). Since it accounts better for false negatives and false positives, the F1 score was chosen over accuracy. The F1 score reported is the mean F1 score across all classes in the output layer.

Table 2.4

Confusion matrix.

		Target (y)		
		True	False	Total
Predicted ($\hat{y}$)	True	*True Positive* (TP)	*False Positive* (FP)	$TP + FP$
	False	*False Negative* (FN)	*True Negative* (TN)	$FN + TN$
	Total	$TP + FN$	$FP + TN$	N

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.13}$$

2.3 Results

2.3.1 Monolingual Baseline

Research looking into the causes of age-related declines in L2 outcomes often compares L2 learner results to those of native-speaking controls, which are often native bilingual speakers. In order to establish a monolingual baseline for each language and a corresponding naive L2 baseline for each language, 30 monolingual models were trained for each of the 12 language pairs under each of the four memory development conditions. The naive L2 baseline is determined by evaluating L2 performance in monolingual models prior to explicitly training on the L2. This led to 1,440 monolingual models trained on 1,000,000 nouns. Model performance was evaluated on the held-out test set every 100,000 nouns, so each model had ten data points in which model performance on the test set was measured. The mean and standard deviation of the final F1 score for each language and memory condition after being trained for 1,000,000 nouns is reported in Table 2.5. Figure 2.6 shows the distributions of the baseline performance of each language in the L1 and the L2 across all four memory conditions.

Separate mixed effects models were fit to the F1 score associated with the L1 (monolingual) and the L2 (naive L2 baseline) in order to identify differences between languages and between memory development conditions. The analysis of the naive

L2 baseline was conducted separately for each L2. Models were fit to the data using the *lme4* package (version 1.1-21; Bates, Mächler, Bolker, & Walker, 2015) within the R statistical platform (version 3.5.2; R Core Team, 2015).

In the monolingual model, contrast coding was used for the language (here, the L1) and memory development factors. The reference for the monolingual model corresponds to French in the No Growth condition. Therefore, all fixed effects across the other factors are relative to these reference levels. The performance of each model was evaluated on the L1 and the naive L2 every 100,000 nouns. These data points were collected at different moments in the training phase. Each model had ten evaluation runs for each language (repeated measurements of the dependent variable). Since the manipulation of experimental factors in this study was entirely between-subjects (i.e., between-models), the maximal random effects structure only included random intercepts for the grouping factors of model, which here represents the subjects in the experiment, and each evaluation run (the repeated measures grouping factor) (Barr, Levy, Scheepers, & Tily, 2013). This random effects structure was used for all mixed effects models.

In the monolingual mixed effects model (see Table 2.6 and Figure 2.7), there was a strong main effect of language. Relative to French, Spanish monolingual models outperformed other monolingual models. Russian models performed worse than French models and German models performed much worse than the other three monolingual language models.

Table 2.5

Mean (standard deviation) final F1 score in the gender assignment experiment for monolingual and naive L2 baselines of each language in each memory development condition. The monolingual (L1) values were gathered after 1,000,000 nouns were used to train the model on the L1. The naive L2 baseline values represent performance on an L2 when no training occurred for that L2.

	No Growth	*Connection Growth*	*Unit Growth*	*Unit Replacement*
French				
L1	0.75 (0.02)	0.76 (0.01)	0.75 (0.04)	0.75 (0.01)
L2	0.33 (0.12)	0.32 (0.08)	0.33 (0.09)	0.34 (0.08)
German				
L1	0.46 (0.05)	0.40 (0.02)	0.36 (0.06)	0.45 (0.01)
L2	0.29 (0.01)	0.27 (0.01)	0.27 (0.02)	0.28 (0.02)
Russian				
L1	0.67 (0.01)	0.56 (0.03)	0.64 (0.02)	0.69 (0.01)
L2	0.29 (0.02)	0.23 (0.04)	0.26 (0.03)	0.29 (0.05)
Spanish				
L1	0.94 (0.04)	0.82 (0.01)	0.74 (0.08)	0.93 (0.01)
L2	0.34 (0.20)	0.29 (0.12)	0.28 (0.20)	0.35 (0.14)

Table 2.6

Mixed effects model examining the effect of memory development and language on learning outcomes in monolingual models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.75	0.00	192.18*	0.00	0.02	0.00	0.01
Connection Growth	0.01	0.00	3.32*	-	-	-	-
Unit Growth	-0.02	0.00	-5.24*	-	-	-	-
Unit Replacement	-0.00	0.00	-0.09	-	-	-	-
L1 German	-0.29	0.00	-89.39*	-	-	-	-
L1 Russian	-0.07	0.00	-15.58*	-	-	-	-
L1 Spanish	0.19	0.00	49.93*	-	-	-	-
Connection Growth x German	-0.08	0.00	-16.07*	-	-	-	-
Unit Growth x L1 German	-0.08	0.00	-16.08*	-	-	-	-
Unit Replacement x L1 German	-0.01	0.00	-1.84	-	-	-	-
Connection Growth x L1 Russian	-0.12	0.01	-17.24*	-	-	-	-
Unit Growth x L1 Russian	-0.04	0.01	-5.50*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.01	2.07*	-	-	-	-
Connection Growth x L1 Spanish	-0.13	0.01	-24.73*	-	-	-	-
Unit Growth x L1 Spanish	-0.17	0.01	-31.24*	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.01	-1.67	-	-	-	-

Note: Mixed effects model formula: F1 ~ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

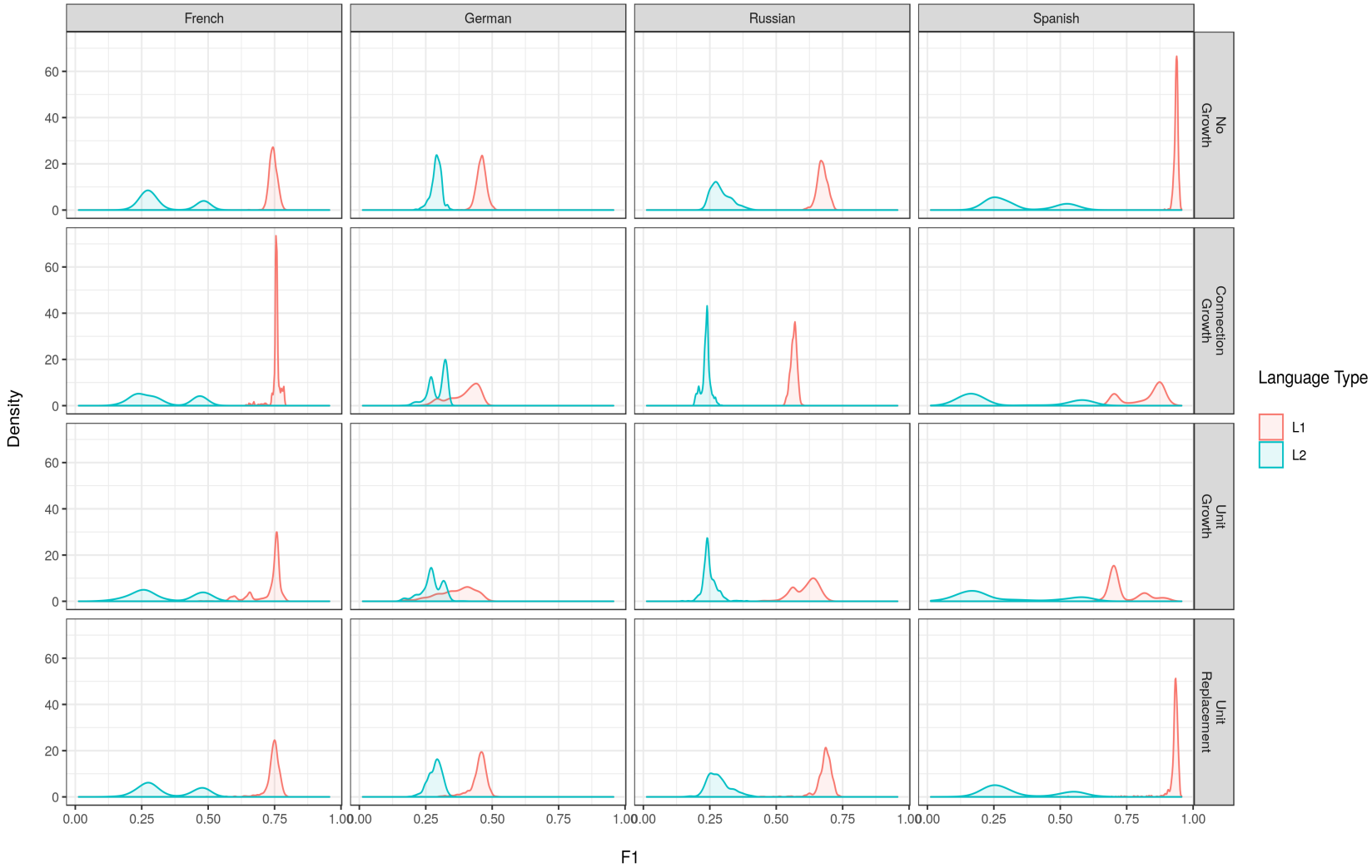

Figure 2.6. Density plots visualizing the distribution of F1 scores of monolingual and naive L2 models in the gender assignment experiment.

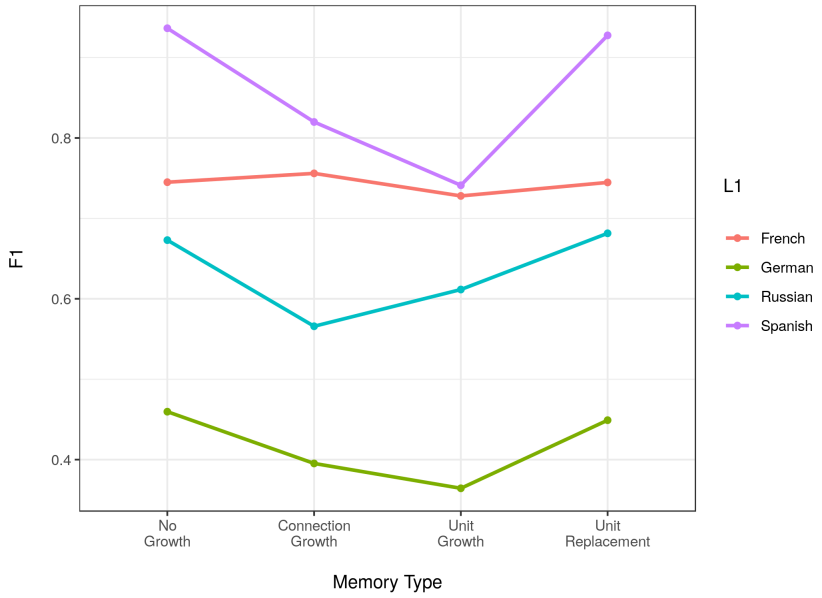

Figure 2.7. Plot visualizing the interaction between memory development condition and L1 in the monolingual models of the gender assignment experiment.

In relation to the No Growth condition, Unit Growth led to poorer outcomes across all languages (a main effect of memory development). The Unit Replacement condition did not have a substantial performance difference from the No Growth condition when French was the reference language. However, Unit Replacement did lead to a small improvement in outcome performance over the No Growth condition when the language was Russian. Relative to the No Growth condition, Connection Growth led to slightly better outcomes in French, whereas it led to considerably worse outcomes in German, Russian, and Spanish. The difference in effect sizes between the Connection Growth condition in French and the same condition in the other three languages was substantial.

Starting with a smaller working memory capacity (Unit Growth) did not lead to better outcomes than starting with a fully developed network. The similar results

between the Connection Growth and Unit Growth conditions suggests that long-term memory capacity is important for successful learning outcomes and that a lack of long-term resources during the early stages of development leads to poorer learning outcomes. French was the only language that slightly benefited from the starting with a sparsely connected network. This result indicates that starting with fewer long-term memory resources in French confers a learning benefit not seen in the other three languages.

A naive baseline in which the L2 was not explicitly trained was established for each L2. The purpose of examining the performance of networks that have not been trained on an L2 was to determine how a particular L1 may influence the learning outcomes of a particular L2. The L2 of each monolingual model was evaluated every 100,000 nouns. Therefore, each model had ten evaluation runs for the L2 (repeated measurements of the dependent variable). A mixed effects model was fit to each L2. The random effects structure included random intercepts for the grouping factors of model and evaluation run.

In the naive L2 French model (see Table 2.7 and Figure 2.8), compared to L1 German, both L1 Russian and L1 Spanish led to better L2 French outcomes. Spanish, in particular, led to a large boost in L2 French outcomes relative to the other languages (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth resulted in worse naive L2 French outcomes (a main effect of memory development). L2 French outcomes were negatively impacted when the L1 was either Russian or Spanish in the Connection Growth condition (an interaction between memory development and L1). In the Unit Growth condition,

L2 French outcomes were negatively impacted by L1 German and L1 Russian. In each L1, the Unit Replacement condition led to a slight decrease in L2 French outcomes relative to the No Growth condition.

In the naive L2 German model, L1 Spanish led to better L2 German outcomes than L1 French, while L1 Russian led to worse outcomes (a main effect of L1). This effect was small, however. There was a main effect of memory development condition; Connection Growth, Unit Growth, and Unit Replacement led to poorer L2 German outcomes when the L1 was French. L1 Russian in the Connection Growth condition led to a decrease in L2 German outcomes (an interaction between memory development and L1). When the L1 was Spanish, both the Connection Growth and Unit Growth conditions performed better than the No Growth and Unit Replacement conditions.

In the naive L2 Russian model, L1 German led to better L2 Russian outcomes than L1 French, and L1 Spanish led to worse outcomes than L1 French (a main effect of L1). There was a main effect of memory development condition; Connection Growth, Unit Growth and Unit Replacement led to poorer L2 Russian outcomes is all L1s (a main effect of L1 and an interaction between memory development and L1).

In the naive L2 Spanish model, compared to L1 French L1 German and L1 Russian led to led considerably worse L2 Spanish outcomes (a main effect of L1). The magnitude of this effect was greater in L1 German. Relative to the No Growth condition, the Connection Growth and Unit Replacement conditions positively impacted L2 Spanish when the L1 was French (a main effect of memory development).

L1 German and L1 Russian interacted with the Connection Growth and Unit Growth conditions to decrease L2 Spanish outcomes (an interaction between memory development and L1).

The naive L2 baselines performed at chance or worse for all L1 - L2 language pairs except for L1 Spanish - L2 German in the Connection Growth and Unit Growth conditions, L1 German - L2 Russian in the No Growth and Unit Replacement conditions and L1 French - L2 Spanish in all conditions. Any positive transfer between these languages was minimal, and only the L1 French - L2 Spanish language pair saw a consistently small positive transfer in all conditions. Next, bilingual models are evaluated using the same mixed effects modeling approach.

2.3.2 Bilingual Models

Eight thousand six hundred forty models were trained to predict the grammatical gender class of a noun represented as a sequence of articulatory feature vectors; 30 training runs per experimental condition (six entrenchment levels, four memory conditions, and 12 language pairs). The performance of each model was evaluated on the held-out test set (600 nouns) of each language every 100,000 nouns and at the end of the bilingual training phase. The F1 score was used to perform statistical analyses and report descriptive statistics. The mean and standard deviation of the final F1 score (after all training was completed) for each language, entrenchment level, and memory condition is reported in Table 2.8.

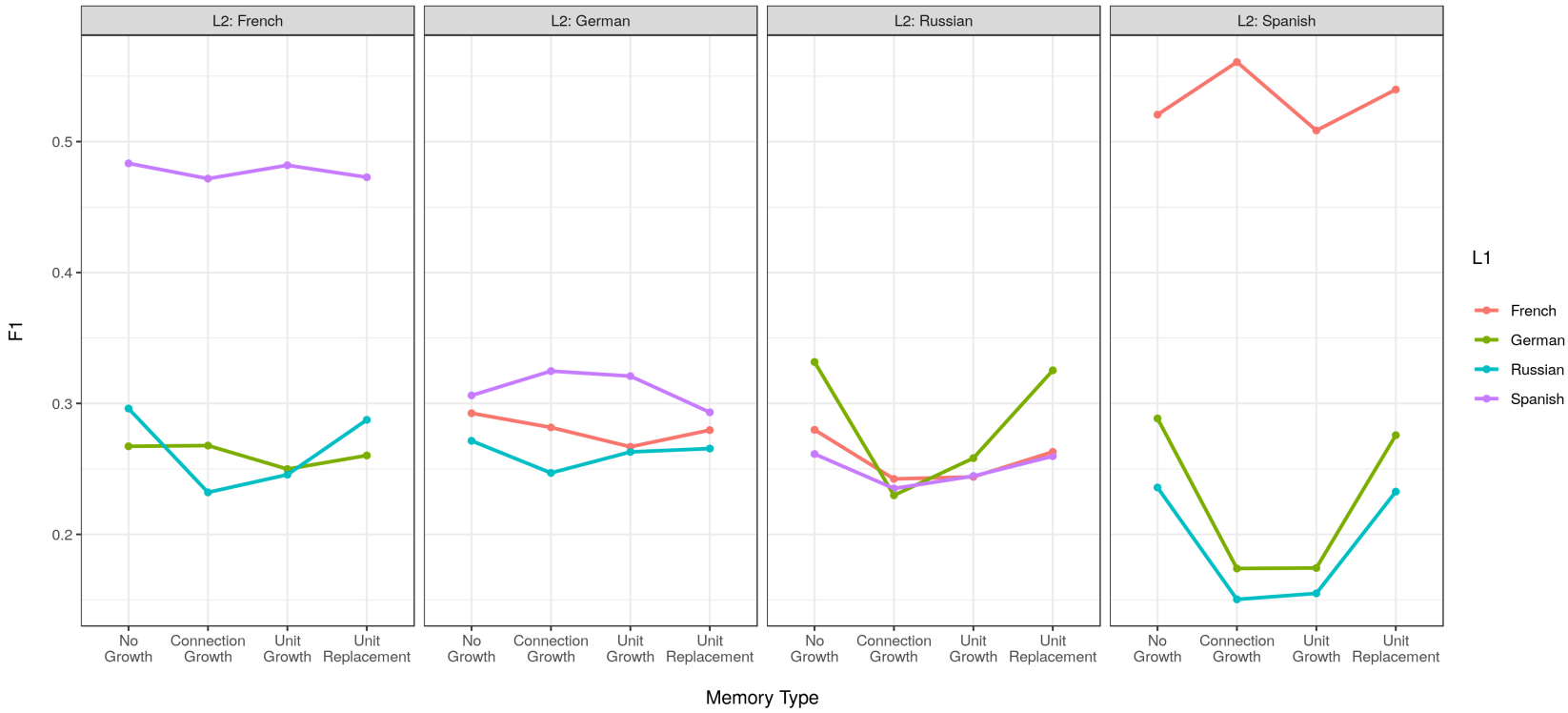

Figure 2.8. Plot visualizing the interaction between memory development condition and L1 for each naive L2 in the Monolingual models.

Table 2.7

Mixed effects models examining the effect of memory development and L1 on learning outcomes in naive L2 models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
French							
(Intercept)	0.27	0.00	95.30*	0.000	0.014	0.000	0.003
Connection Growth	0.00	0.00	0.13	-	-	-	-
Unit Growth	-0.02	0.00	-3.89*	-	-	-	-
Unit Replacement	-0.01	0.00	-1.59	-	-	-	-
L1 Russian	0.03	0.01	5.19*	-	-	-	-
L1 Spanish	0.22	0.00	43.32*	-	-	-	-
Connection Growth x L1 Russian	-0.06	0.01	-7.97*	-	-	-	-
Unit Growth x L1 Russian	-0.03	0.01	-4.02*	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.01	-0.20	-	-	-	-
Connection Growth x L1 Spanish	-0.01	0.01	-1.58	-	-	-	-
Unit Growth x L1 Spanish	0.02	0.01	2.11*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.01	-0.48	-	-	-	-
German							
(Intercept)	0.29	0.00	114.34*	0.000	0.011	0.000	0.004
Connection Growth	-0.01	0.00	-3.38*	-	-	-	-
Unit Growth	-0.03	0.00	-8.06*	-	-	-	-
Unit Replacement	-0.01	0.00	-4.02*	-	-	-	-
L1 Russian	-0.02	0.00	-4.95*	-	-	-	-
L1 Spanish	0.01	0.00	2.73*	-	-	-	-
Connection Growth x L1 Russian	-0.01	0.01	-2.26*	-	-	-	-
Unit Growth x L1 Russian	0.02	0.01	2.85*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.01	1.16	-	-	-	-
Connection Growth x L1 Spanish	0.03	0.01	4.97*	-	-	-	-
Unit Growth x L1 Spanish	0.04	0.01	5.59*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.01	0.00	-	-	-	-
Russian							
(Intercept)	0.28	0.00	89.88*	0.000	0.015	0.000	0.003
Connection Growth	-0.04	0.00	-9.02*	-	-	-	-
Unit Growth	-0.04	0.00	-8.64*	-	-	-	-
Unit Replacement	-0.02	0.00	-4.05*	-	-	-	-
L1 German	0.05	0.00	12.77*	-	-	-	-
L1 Spanish	-0.02	0.00	-4.56*	-	-	-	-
Connection Growth x L1 German	-0.06	0.01	-11.07*	-	-	-	-
Unit Growth x L1 German	-0.04	0.01	-6.45*	-	-	-	-
Unit Replacement x L1 German	0.01	0.01	1.81	-	-	-	-
Connection Growth x L1 Spanish	0.01	0.01	1.94	-	-	-	-
Unit Growth x L1 Spanish	0.02	0.01	3.31*	-	-	-	-
Unit Replacement x L1 Spanish	0.02	0.01	2.63*	-	-	-	-
Spanish							
(Intercept)	0.52	0.01	83.02*	0.000	0.018	0.000	0.015
Connection Growth	0.04	0.01	6.68*	-	-	-	-
Unit Growth	-0.01	0.01	-2.01*	-	-	-	-
Unit Replacement	0.02	0.01	3.19*	-	-	-	-
L1 German	-0.23	0.01	-39.20*	-	-	-	-
L1 Russian	-0.28	0.01	-46.91*	-	-	-	-
Connection Growth x L1 German	-0.15	0.01	-18.26*	-	-	-	-
Unit Growth x L1 German	-0.10	0.01	-12.04*	-	-	-	-
Unit Replacement x L1 German	-0.03	0.01	-3.82*	-	-	-	-
Connection Growth x L1 Russian	-0.13	0.01	-14.57*	-	-	-	-
Unit Growth x L1 Russian	-0.07	0.01	-8.01*	-	-	-	-
Unit Replacement x L1 Russian	-0.02	0.01	-2.59*	-	-	-	-

Note: Mixed effects model formula: F1 ~ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 2.8

Means (standard deviations) of final F1 scores for L1 and L2 outcomes of each language across entrenchment and memory development conditions in the gender assignment experiment. All of the values reported were calculated using F1 score obtain after all training was completed.

		No Growth		*Connection Growth*		*Unit Growth*		*Unit Replacement*	
		L1	*L2*	*L1*	*L2*	*L1*	*L2*	*L1*	*L2*
French									
	Early	0.57 (0.12)	0.53 (0.11)	0.57 (0.12)	0.54 (0.10)	0.56 (0.13)	0.53 (0.12)	0.57 (0.12)	0.55 (0.12)
	Middle	0.57 (0.12)	0.52 (0.11)	0.58 (0.12)	0.54 (0.10)	0.56 (0.13)	0.53 (0.12)	0.57 (0.12)	0.55 (0.12)
	Late	0.57 (0.12)	0.52 (0.10)	0.58 (0.12)	0.54 (0.09)	0.56 (0.13)	0.52 (0.12)	0.57 (0.12)	0.54 (0.12)
German									
	Early	0.46 (0.02)	0.46 (0.02)	0.38 (0.05)	0.37 (0.04)	0.35 (0.06)	0.35 (0.06)	0.45 (0.03)	0.45 (0.03)
	Middle	0.45 (0.02)	0.46 (0.02)	0.39 (0.04)	0.37 (0.04)	0.36 (0.06)	0.35 (0.05)	0.45 (0.02)	0.45 (0.02)
	Late	0.45 (0.02)	0.46 (0.02)	0.39 (0.04)	0.37 (0.04)	0.37 (0.06)	0.35 (0.05)	0.45 (0.02)	0.45 (0.03)
Russian									
	Early	0.67 (0.02)	0.67 (0.02)	0.53 (0.04)	0.53 (0.05)	0.57 (0.07)	0.51 (0.08)	0.67 (0.03)	0.67 (0.03)
	Middle	0.67 (0.02)	0.66 (0.02)	0.54 (0.04)	0.53 (0.05)	0.58 (0.06)	0.51 (0.08)	0.68 (0.02)	0.67 (0.02)
	Late	0.67 (0.02)	0.66 (0.02)	0.55 (0.03)	0.53 (0.05)	0.60 (0.05)	0.50 (0.07)	0.68 (0.02)	0.67 (0.02)
Spanish									
	Early	0.69 (0.14)	0.77 (0.16)	0.68 (0.15)	0.76 (0.13)	0.57 (0.14)	0.66 (0.11)	0.68 (0.13)	0.77 (0.16)
	Middle	0.69 (0.14)	0.76 (0.16)	0.71 (0.15)	0.76 (0.12)	0.58 (0.15)	0.66 (0.11)	0.68 (0.13)	0.76 (0.16)
	Late	0.68 (0.13)	0.76 (0.16)	0.71 (0.15)	0.76 (0.12)	0.56 (0.14)	0.66 (0.10)	0.67 (0.13)	0.76 (0.16)

There are two main components to the experimental design. One is to identify any effects due to L1 entrenchment and memory development, and the other is to understand how specific L1s influence L2 outcomes. Similar to the modeling approach taken in the naive L2 baseline analysis, for each L2 mixed effects models were fit to the fixed factors of entrenchment, memory development, and L1. The factor of entrenchment was simplified in these analyses by reducing the factor to three levels (*Early* (0 and 200,000), *Middle* (400,000 and 600,000), and *Late* (800,000 and 1,000,000) L2 learners). Contrast coding was used for the entrenchment, memory development, and L1 factors. The L2 of each model was evaluated every 100,000 nouns during the bilingual phase. Therefore, each model had 20 evaluation runs for the L2 (repeated measurements of the dependent variable). Models were fit to the data using the lme4 package within the R statistical platform. The random effects structure included random intercepts for the grouping factors of model and evaluation run. This random effects structure was used for all mixed effects models.

In the L2 French model (see Table 2.9 and Figure 2.9), L1 Spanish led to considerably better L2 French outcomes than L1 German (a main effect of L1). Compared to early L2 learners, starting an L2 at higher entrenchment levels led to worse L2 French learning outcomes (a main effect of entrenchment). Connection Growth, Unit Growth, and Unit Replacement led to an increase in L2 French outcomes when the L1 was German (a main effect of memory development). The opposite was true, however, when the L1 was either Russian or Spanish; in these languages, the Connection Growth and Unit Growth conditions led to poorer L2 French outcomes (an interaction between memory development and L1).

Table 2.9

Mixed effects model examining the effect of entrenchment, memory development, and L1 on learning outcomes in L2 French models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.48	0.00	314.89*	0.000	0.012	0.000	0.002
Middle L2 Learner	-0.01	0.00	-5.69*	-	-	-	-
Late L2 Learner	-0.01	0.00	-5.50*	-	-	-	-
Connection Growth	0.03	0.00	11.33*	-	-	-	-
Unit Growth	0.01	0.00	3.00*	-	-	-	-
Unit Replacement	0.01	0.00	3.00*	-	-	-	-
L1 Russian	0.00	0.00	0.42	-	-	-	-
L1 Spanish	0.27	0.00	99.29*	-	-	-	-
Middle L2 Learner x Connection Growth	0.01	0.00	3.26*	-	-	-	-
Late L2 Learner x Connection Growth	0.01	0.00	3.46*	-	-	-	-
Middle L2 Learner x Unit Growth	0.01	0.00	1.71	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.00	1.28	-	-	-	-
Middle L2 Learner x Unit Replacement	0.00	0.00	0.20	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.00	-1.34	-	-	-	-
Middle L2 Learner x L1 Russian	0.01	0.00	3.01*	-	-	-	-
Late L2 Learner x L1 Russian	0.01	0.00	2.62*	-	-	-	-
Middle L2 Learner x L1 Spanish	0.01	0.00	2.27*	-	-	-	-
Late L2 Learner x L1 Spanish	0.01	0.00	1.85	-	-	-	-
Connection Growth x L1 Russian	-0.03	0.00	-7.31*	-	-	-	-
Unit Growth x L1 Russian	-0.03	0.00	-7.66*	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.00	-0.38	-	-	-	-
Connection Growth x L1 Spanish	-0.03	0.00	-7.81*	-	-	-	-
Unit Growth x L1 Spanish	-0.04	0.00	-9.77*	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.00	-1.91	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	-0.01	0.01	-2.03*	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	-0.01	0.01	-1.65	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	-0.01	0.01	-1.18	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	-0.01	0.01	-0.97	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.00	0.01	-0.63	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	0.00	0.01	0.78	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	-0.01	0.01	-2.05*	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	-0.02	0.01	-2.91*	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	0.00	0.01	0.06	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.00	0.01	0.54	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.00	0.01	0.30	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.01	0.01	1.79	-	-	-	-

Note: Mixed effects model formula: F1 $\sim$ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

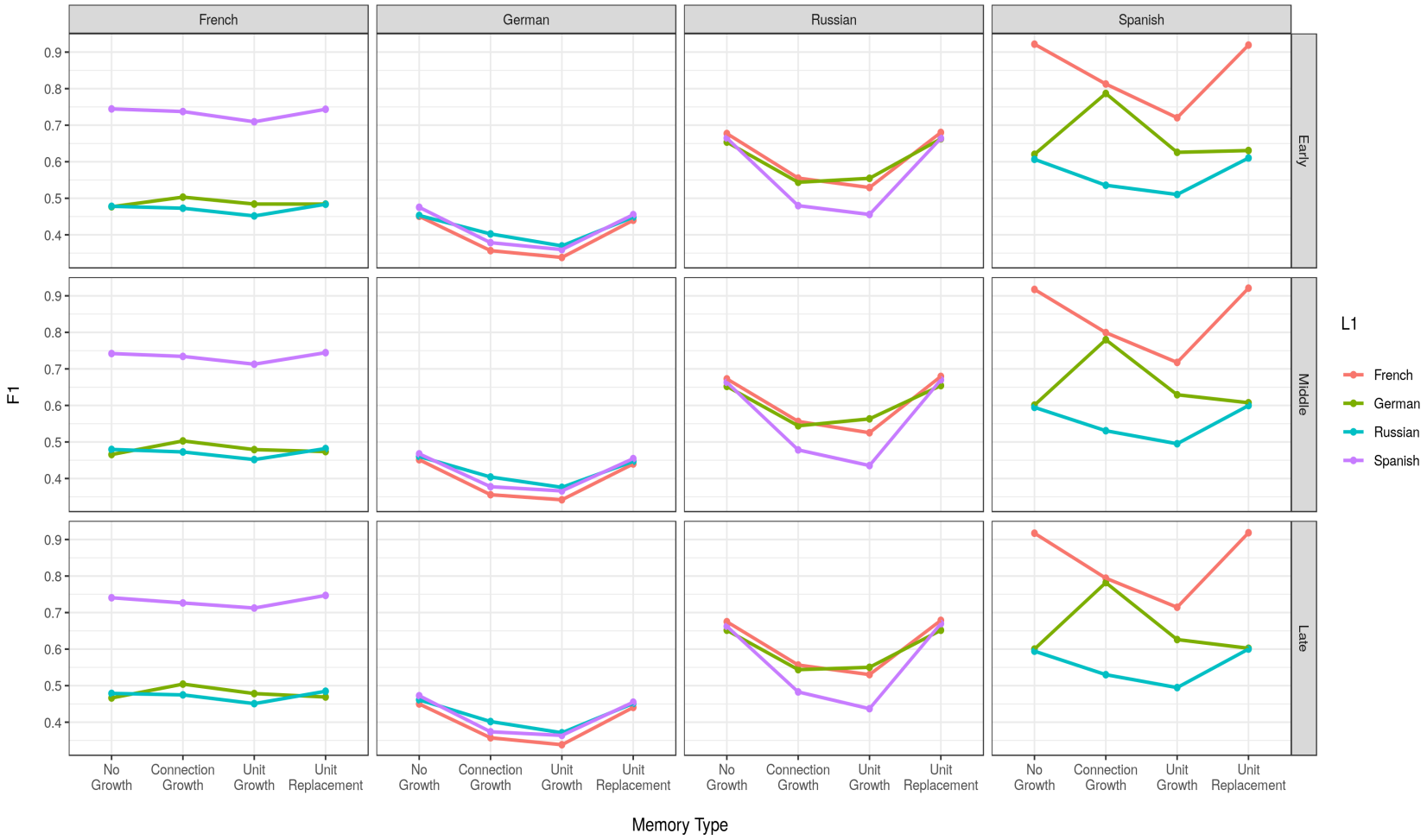

Figure 2.9. Plot visualizing the interaction between entrenchment, memory development, and L1 for each L2 in the gender assignment experiment.

In the L2 German model (see Table 2.10), L1 Spanish led to modest improvements in L2 German outcomes compared to L1 French (a main effect of L1). Compared to early L2 learners, starting an L2 at higher entrenchment levels led to worse L2 German outcomes (a main effect of entrenchment). However, this entrenchment effect was minimal. Connection Growth, Unit Growth, and Unit Replacement led to a decrease in L2 German outcomes across all languages (a main effect of memory development). The magnitude of the decrease was quite large for the Connection Growth and Unit Growth conditions and quite small for the Unit Replacement condition.

In the L2 Russian model (see Table 2.11), L1 German and L1 Spanish led to small reductions in L2 Russian outcomes compared to L1 French (a main effect of L1). Higher entrenchment levels led to slightly worse L2 Russian outcomes (a main effect of entrenchment). Connection Growth and Unit Growth led to a large decrease in L2 Russian outcomes across all languages (a main effect of memory development), especially when the L1 was Spanish (an interaction between memory development and L1).

In the L2 Spanish model (see Table 2.12), L1 German and L1 Russian led to large reductions in L2 Spanish outcomes compared to L1 French (a main effect of L1). Compared to low entrenchment levels, higher entrenchment levels led to slightly worse L2 Spanish outcomes (a main effect of entrenchment). Connection Growth and Unit Growth led to a large decrease in L2 Spanish outcomes when the L1 was either French or Russian (a main effect of memory development); however, when the L1 was German these conditions led to improvements in L2 Spanish performance

(an interaction between memory development and L1).

The probability of producing false negatives increased due to the multiple comparisons across the factors of interest. However, the mixed effects results were not adjusted with a Bonferroni correction because the data points were not independent and many of the patterns were consistent.

Table 2.10

Mixed effects model examining the effect of entrenchment, memory development, and L1 on learning outcomes in L2 German models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.45	0.00	151.35*	0.000	0.012	0.000	0.013
Middle L2 Learner	-0.01	0.00	-2.23*	-	-	-	-
Late L2 Learner	-0.01	0.00	-4.14*	-	-	-	-
Connection Growth	-0.09	0.00	-37.81*	-	-	-	-
Unit Growth	-0.11	0.00	-45.53*	-	-	-	-
Unit Replacement	-0.01	0.00	-4.36*	-	-	-	-
L1 Russian	0.00	0.00	0.78	-	-	-	-
L1 Spanish	0.02	0.00	6.49*	-	-	-	-
Middle L2 Learner x Connection Growth	0.00	0.00	-0.51	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.00	0.37	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.00	0.85	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.00	0.28	-	-	-	-
Middle L2 Learner x Unit Replacement	0.00	0.00	-0.25	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.00	0.37	-	-	-	-
Middle L2 Learner x L1 Russian	0.01	0.00	1.44	-	-	-	-
Late L2 Learner x L1 Russian	0.01	0.00	1.95	-	-	-	-
Middle L2 Learner x L1 Spanish	-0.01	0.01	-1.54	-	-	-	-
Late L2 Learner x L1 Spanish	0.00	0.01	-0.41	-	-	-	-
Connection Growth x L1 Russian	0.04	0.00	9.70*	-	-	-	-
Unit Growth x L1 Russian	0.03	0.00	6.57*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.00	1.38	-	-	-	-
Connection Growth x L1 Spanish	0.00	0.00	-0.67	-	-	-	-
Unit Growth x L1 Spanish	0.00	0.01	-0.59	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.00	-2.14*	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.00	0.01	-0.58	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	-0.01	0.01	-1.57	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.00	0.01	-0.64	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	-0.01	0.01	-1.20	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	-0.01	0.01	-1.24	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.01	0.01	-1.00	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.01	0.01	1.31	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.00	0.01	-0.47	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	0.01	0.01	1.42	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.01	0.01	0.78	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.01	0.01	1.25	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.00	0.01	0.20	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 2.11

Mixed effects model examining the effect of entrenchment, memory development, and L1 on learning outcomes in L2 Russian models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.68	0.00	183.87*	0.000	0.023	0.000	0.012
Middle L2 Learner	-0.01	0.00	-2.39*	-	-	-	-
Late L2 Learner	-0.01	0.00	-2.63*	-	-	-	-
Connection Growth	-0.12	0.00	-28.50*	-	-	-	-
Unit Growth	-0.15	0.00	-34.39*	-	-	-	-
Unit Replacement	0.00	0.00	0.69	-	-	-	-
L1 German	-0.02	0.00	-5.54*	-	-	-	-
L1 Spanish	-0.01	0.00	-2.96*	-	-	-	-
Middle L2 Learner x Connection Growth	0.01	0.01	0.91	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.01	0.56	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.01	0.06	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.01	0.49	-	-	-	-
Middle L2 Learner x Unit Replacement	0.00	0.01	0.64	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.01	0.12	-	-	-	-
Middle L2 Learner x L1 German	0.00	0.01	0.48	-	-	-	-
Late L2 Learner x L1 German	0.00	0.01	0.02	-	-	-	-
Middle L2 Learner x L1 Spanish	0.00	0.01	0.49	-	-	-	-
Late L2 Learner x L1 Spanish	0.00	0.01	-0.03	-	-	-	-
Connection Growth x L1 German	0.01	0.01	2.00*	-	-	-	-
Unit Growth x L1 German	0.05	0.01	8.07*	-	-	-	-
Unit Replacement x L1 German	0.01	0.01	1.01	-	-	-	-
Connection Growth x L1 Spanish	-0.06	0.01	-10.51*	-	-	-	-
Unit Growth x L1 Spanish	-0.06	0.01	-10.19*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.01	-0.58	-	-	-	-
Middle L2 Learner x Connection Growth x L1 German	0.00	0.01	-0.40	-	-	-	-
Late L2 Learner x Connection Growth x L1 German	0.00	0.01	-0.15	-	-	-	-
Middle L2 Learner x Unit Growth x L1 German	0.01	0.01	1.15	-	-	-	-
Late L2 Learner x Unit Growth x L1 German	-0.01	0.01	-0.62	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 German	-0.01	0.01	-1.25	-	-	-	-
Late L2 Learner x Unit Replacement x L1 German	-0.01	0.01	-1.13	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	-0.01	0.01	-0.64	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.00	0.01	0.23	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	-0.02	0.01	-2.24*	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	-0.02	0.01	-2.27*	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.00	0.01	0.41	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.01	0.01	0.84	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

2.4 Discussion

The experiment presented above investigated the impact specific languages can have on the learning of grammatical gender assignment when developmental

Table 2.12

Mixed effects model examining the effect of entrenchment, memory development, and L1 on learning outcomes in L2 Spanish models in the gender assignment experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.93	0.00	250.35*	0.000	0.019	0.000	0.012
Middle L2 Learner	-0.01	0.00	-2.24*	-	-	-	-
Late L2 Learner	-0.01	0.00	-3.46*	-	-	-	-
Connection Growth	-0.11	0.00	-26.48	-	-	-	-
Unit Growth	-0.20	0.00	-48.92*	-	-	-	-
Unit Replacement	0.00	0.00	-0.61	-	-	-	-
L1 German	-0.30	0.00	-74.53*	-	-	-	-
L1 Russian	-0.32	0.00	-76.08*	-	-	-	-
Middle L2 Learner x Connection Growth	-0.01	0.01	-1.57	-	-	-	-
Late L2 Learner x Connection Growth	-0.01	0.01	-2.42*	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.01	0.26	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.01	-0.26	-	-	-	-
Middle L2 Learner x Unit Replacement	0.01	0.01	1.03	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.01	0.61	-	-	-	-
Middle L2 Learner x L1 German	-0.02	0.01	-2.76*	-	-	-	-
Late L2 Learner x L1 German	-0.02	0.01	-2.80*	-	-	-	-
Middle L2 Learner x L1 Russian	-0.01	0.01	-1.07	-	-	-	-
Late L2 Learner x L1 Russian	-0.01	0.01	-1.02	-	-	-	-
Connection Growth x L1 German	0.28	0.01	47.90*	-	-	-	-
Unit Growth x L1 German	0.21	0.01	35.64*	-	-	-	-
Unit Replacement x L1 German	0.01	0.01	2.17*	-	-	-	-
Connection Growth x L1 Russian	0.04	0.01	6.38*	-	-	-	-
Unit Growth x L1 Russian	0.10	0.01	17.28*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.01	0.92	-	-	-	-
Middle L2 Learner x Connection Growth x L1 German	0.02	0.01	2.77*	-	-	-	-
Late L2 Learner x Connection Growth x L1 German	0.03	0.01	3.44*	-	-	-	-
Middle L2 Learner x Unit Growth x L1 German	0.02	0.01	2.66*	-	-	-	-
Late L2 Learner x Unit Growth x L1 German	0.02	0.01	2.79*	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 German	-0.01	0.01	-1.11	-	-	-	-
Late L2 Learner x Unit Replacement x L1 German	-0.01	0.01	-1.47	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.02	0.01	1.96	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.02	0.01	2.51*	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.00	0.01	0.00	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.01	0.06	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.00	0.01	-0.49	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	0.00	0.01	-0.10	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

factors like linguistic entrenchment and memory development vary. Previous work by Monner et al. (2013) mostly found support for the entrenchment and less-is-more hypotheses but failed to find entrenchment or memory development effects in L1 French - L2 Spanish. The authors of that study suggested that the formal

cue regularity of the language may have led to the null effects in the L1 French - L2 Spanish language pair. This experiment extends that research by investigating two additional languages, German and Russian. Language models represented as recurrent neural networks were trained to map a gender class to a noun represented as a sequence of phonemes. Contrary to Monner et al. (2013), which evaluated the model performance on the data used to train the network, this experiment used a held-out test set for model evaluation.

An analysis of monolingual performance for each memory development condition indicated large baseline performance differences between the languages. Spanish performed the best, followed by French, Russian, and then German. This pattern suggests that formal cue transparency and the number of gender classes influence gender system learning outcomes. The results in the monolingual models also show that starting small did not lead to better learning outcomes. In French, however, starting small actually led to increases in learning outcomes. The naive L2 baselines performed at chance or worse than chance for all L1 - L2 language pairs except for L1 Spanish - L2 German, L1 German - L2 Russian, and L1 French - L2 Spanish. The low, but consistent above chance performance of naive L2 Spanish models when L1 was French suggests that L1 French provided some positive transfer across the gender classes.

L2 outcomes were affected by the number of gender classes shared between the L1 and the L2. If the L2 had fewer gender classes than the L1, the L2 outcomes were negatively affected (L1 German - L2 French, L1 Russian - L2 French, L1 German - L2 Spanish, L1 Russian - L2 Spanish). This finding is contrary to results

obtained with human L2 learners; learning an L2 that is less complex than the L1 generally leads to better results than if the L1 had the same number of classes as the L2 (Schepens, Van der Slik, & Van Hout, 2013, 2014). It is possible that the phoneme sequences across the two languages did not provide enough information in order to activate unique patterns for the two languages. If the model could not separate the activation patterns of one language from those of another, the model would be unable to identify when a particular gender class was not an option. This would lead to an increase in misclassifications. In other words, the model is treating all input as though it belongs to one language. Knowledge that it is possible to have two languages that may differ with regards to the number of possible gender classes is not represented in the network. The language models represent knowledge of the statistical regularities of the phonemic sequences of two languages within the same network. This knowledge is entirely statistical and only includes phonological information. The number of shared phonemes between the two languages and any corresponding n-phones may lead to greater levels of interference. Phoneme sequences and the articulatory features represented within them may not have sufficient information for the network to learn the language-specific patterns needed to classify nouns correctly. A different neural network architecture may provide the appropriate apparatus to store this information. Recent research by Santoro, Bartunov, Botvinick, Wierstra, and Lillicrap (2016) using what they call memory-augmented neural networks may be the kind of model architecture that leads to results that more closely mirror those found with human learners.

Except for L1 German - L2 French and L1 German - L2 Spanish in the Con-

nection Growth and Unit Growth conditions, starting with fewer memory resources (Connection Growth and Unit Growth) led to worse L2 learning outcomes. There was a weak entrenchment effect found in all of the L2 models. Entrenchment negatively influenced L2 learning outcomes, but only slightly. The magnitude of the effect was minimal compared to other fixed effects. This result is consistent with those for the L1 Spanish - L2 French language pair of the gender assignment experiment in Monner et al. (2013), but not with the results of the other language pair.

Chapter 3: Gender Agreement

Grammatical gender helps determine the form of other words during sentence construction (Corbett, 2007). For L2 learners, this fundamental component of successful language use is often difficult to acquire, an effect even seen in advanced L2 speakers who began to learn their L2 during early childhood (Morgan-Short, Sanz, Steinhauer, & Ullman, 2010).

In Monner et al.(2013), the gender agreement experiment found that outcomes on the L2 were negatively impacted by the linguistic entrenchment factor. The results also indicated that starting with fewer working memory resources (i.e., Unit Growth condition) leads to a muting of these entrenchment effects. Since the results in their study were evaluated on the same data used to train the models, it is unknown how well language models would generalize this knowledge to unseen data. Also, the data used to train the models was limited in scope to gender agreement between nouns and adjectives. The models were only trained on determiner-noun-adjective or determiner-adjective-noun word sets. This type of linguistic data is not typically encountered in isolation by human language learners. Instead, sequences tend to be longer, more complex, and carry more contextual information.

In order to understand better how the factors of entrenchment and memory

development impact L2 learning outcomes on a more difficult task like the gender agreement task, a more diverse and complex dataset is used to train the language models to predict the next phoneme given a context of previously seen phonemes. This task is more difficult than the gender assignment task presented previously. Similar to the gender assignment experiment, linguistic entrenchment is expected to have an overall negative effect on L2 learning outcomes. Also, starting the learning process with a smaller working memory capacity is not expected to mute the negative effects of entrenchment, nor is it expected to lead to better L2 learning outcomes (Rohde & Plaut, 1999; Monner et al., 2013). It is expected that models with an L2 gender system that is high in formal cue regularity (e.g., Spanish) will have better L2 learning outcomes than those with lower levels of formal cue regularity (e.g., French). Specifically, Spanish, the language with the most formal cue regularity, is expected to outperform both French and Russian, two languages which also have high levels of formal cue regularity.

English, French, Russian, and Spanish corpora were used to train neural networks to predict the next phoneme in a sequence of phonemes. English does not use gender to classify nouns systematically. However, gender (feminine or masculine) is visible in the pronouns when the object is informed by biological traits, like when a noun is referring to a person or animal. Most objects are otherwise neutral, except in the case of certain culturally determined instances (e.g., the use of *she* for a ship or boat). Unlike French, Russian, and Spanish, English does not have formal cues that inform and determine the class of the noun. The inclusion of English in this experiment addresses whether an L1 without formal principles governing gram-

matical gender agreement between nouns and adjectives will impact the learning of gender agreement in an L2 that does have a high degree of formal cue regularity with regards to gender class assignment. In other words, does the absence of this linguistic system in an L1 influence its acquisition in an L2?

3.1 Method

In the present experiment, recurrent neural networks using an LSTM architecture (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002) were trained to predict the next phoneme in a sequence of phonemes. These types of models are often referred to as sequence models. Table 2.3 provides details on the specific architecture and hyperparameters used to train the models in this experiment. At every step of the sequence, the model predicts the next phoneme. Figure 3.1 illustrates how phonemes (x_t) at each time-step were passed into the network and at each time-step generated a prediction (x_{t+1}) for the next phoneme in the sequence. Four different languages were paired (English, French, Russian, Spanish), creating a total of 12 unique language pairs. Following the experimental design of the gender assignment experiment, the same two developmental variables were manipulated: linguistic entrenchment and memory development.

Language models are often comprised of words (lexical items), parts-of-speech, characters, or phonemes. For example, in a language model with words as the linguistic unit, the next word in the context "Her teacher wrote on the ...." would produce a vector of probabilities across the vocabulary (i.e., all words present in

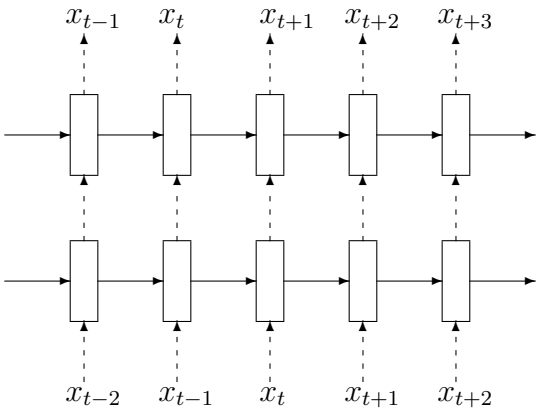

Figure 3.1. Two-layer recurrent neural network that maps each phoneme (x_t) in a sequence to the next phoneme in the sequence (x_{t+1}) (many-to-many).

the training corpus). Depending on this linguistic context and other contexts the model previously encountered, the model would predict the best word to follow the sequence, which could be: "chalkboard", "whiteboard", "paper", "wall", "canvas". However, it could also select non-intuitive words like "carrot", "wind", "cloud", or "cat". The probability distribution depends upon the data used to train the model. If training data contains instances of people writing on carrots, then the likelihood that the word "carrot" is selected for sentences like the one above will increase.

In this experiment, the language models used phonemes as the linguistic unit. For each trial during training, a sequence of vectors consisting of 22 articulatory features that represent a unique phoneme was presented as input to the network. At each time-step (x_t) of the sequence, the model produced an output vector representing the phoneme at the next time-step (x_{t+1}). For example, at each time-step in the sequence of phonemes for /mikasaessukasa/ (my house is your house), the

model would produce an output vector predicting which phoneme is likely to follow the current phoneme (see Figure 3.2).

[01100010100100100000000] → /i/
[11010000100000001000010] → /k/
[00100000000000001010000] → /a/
[11010000100000000110010] → /s/
[00110000000110000000000] → /a/
[11010000100000000110010] → /e/
[11010000100000000000010] → /s/
[00110000000110000000000] → /s/
[00110000000110000000000] → /u/
[11010000100000011011010] → /k/
[00100000000000001010000] → /a/
[11010000100000000110010] → /s/
[00110000000110000000000] → /a/

Figure 3.2. Articulatory feature vectors of input phonemes mapped to target phonemes for /mikasaessukasa/.

The model learned solely from encoded phonological features of the phonemes in the phrases; therefore, only patterns associated with phonological features are potentially being learned by the models. The models were evaluated on two tasks, one intrinsic and the other extrinsic. The intrinsic task was identical to the learning task used to train the models; how well does the model predict the next phoneme. The extrinsic task still required the model to predict the next phoneme, but the type of input in the extrinsic task was specific to grammatical gender agreement. In the extrinsic task, models were asked to predict the final two phonemes of a noun-adjective pair. The noun and the root of the adjective were provided to the model. The model then predicted the next two phonemes in the sequence.

3.1.1 Phrases

All of the data used to train and test the models in this experiment came from the UNPC (Ziemski et al., 2016). Language-specific part-of-speech tagging models using a recurrent neural network architecture (RNNTagger; Schmid, 2019) were applied to each of the aligned English, French, Russian and Spanish phrases of the UNPC corpus. The UNPC contains aligned corpora for all six official UN languages (Chinese, English, French, Modern Standard Arabic, Russian, Spanish). In this experiment, only four of the six official UN languages were used. English, French, Russian, and Spanish were chosen over Modern Standard Arabic and Chinese, the other two languages, for several reasons. Modern Standard Arabic is a formalized, literary language. It is mainly spoken in formal settings, or when speakers do not speak a common dialect. Chinese is a more difficult language to parse than the other languages. For example, word boundaries are more difficult to determine in written Chinese. For these reasons, Modern Standard Arabic and Chinese were not included. German was not used in this experiment because it is not included in the UNPC corpus.

As a first pass across the raw UNPC data, content within and including a pair of parentheses or brackets (e.g., (...), {...}, [...], <...>) was removed. Often, content within parentheses or brackets is dissimilar from other content in that it is either an incomplete phrase, an acronym, a citation, a numerical value, or some other class of atypical spoken language. Simply put, content within parentheses or brackets is not fully integrated into the syntax of the main clause.

All phrases containing an Arabic numeral (i.e., number) were excluded. Numbers can represent various concepts that are not conducive to being represented as speech. Speech with large precise numbers is also not common and often does not provide additional linguistic information. Content at the beginning of the phrase that was not linguistic, but merely organizational, like table formats and list identifiers (e.g., A., B., I., II. etc.), was removed from the phrase. Special characters and symbols (e.g., %, $, #, etc.) were also removed from the phrase. These preprocessing steps led to a much smaller subset of the UNPC.

Language-specific part-of-speech tagging models using a recurrent neural network architecture (RNNTagger; Schmid, 2019) were applied to the subset of the UNPC corpora in English, French, Russian and Spanish. Each part-of-speech model produced a sequence of part-of-speech tags and lemmas for each word in the phrase. Lexical frequency dictionaries were used to select eligible phrases for the training and test sets (Brysbaert et al., 2011; Cuetos et al., 2011; New et al., 2004; Sharoff, 2002). If all words from the phrase were present in the associated lexical frequency dictionary, it was set aside for possible inclusion in either the training or test sets.

Each phrase had a mean lexical frequency. This was calculated by summing the lexical frequency of each word in the phrase and then dividing by the number of words in the phrase. The mean lexical frequency of each phrase was used to influence the selection of phrases during model training. Phrases with a higher frequency were more likely to be selected as input during the training phase.

All of the phrases where each word was present in the lexical frequency dictionary were aligned across the four languages (English, French, Russian, Spanish).

Of these aligned phrases, 102,000 were randomly selected without replacement. An aligned phrase refers to a phrase and its translation equivalent in the other languages. The train and test sets of each language contained the same semantic content. That is, the translation equivalents were in each corresponding language set. Once the training and test sets were created, each phrase was converted into a sequence of articulatory feature vectors. The orthographic text of each noun was transliterated into phonemic segments represented by IPA symbols using the *Epitran* Python package (Mortensen et al., 2016). Each phonemic segment was then mapped to a set of 22 articulatory features using the *PanPhon* Python package (Mortensen et al., 2016).

An aligned phrase was only eligible for selection if it did not contain an Arabic numeral, had a sequence length between 20 and 100 IPA symbols, and each word in the phrase was present in the appropriate lexical dictionary (Brysbaert et al., 2011; Cuetos et al., 2011; New et al., 2004; Sharoff, 2002). From the 102,000 aligned phrases selected, 100,000 were randomly assigned to the training set, and the remaining 2,000 were assigned to the test set (see Table 3.1 for a description of the datasets).

Russian had the most number of unique phonemes (n=56), followed by English (n=48), French (n=47), and then Spanish (n=40). Across all of the languages, there were 89 unique phonemes represented. Figure 3.3 shows which phonemes the languages share. Russian is the most different from the other three. The most similar language pair in terms of phoneme overlap is English and French, followed by French and Spanish.

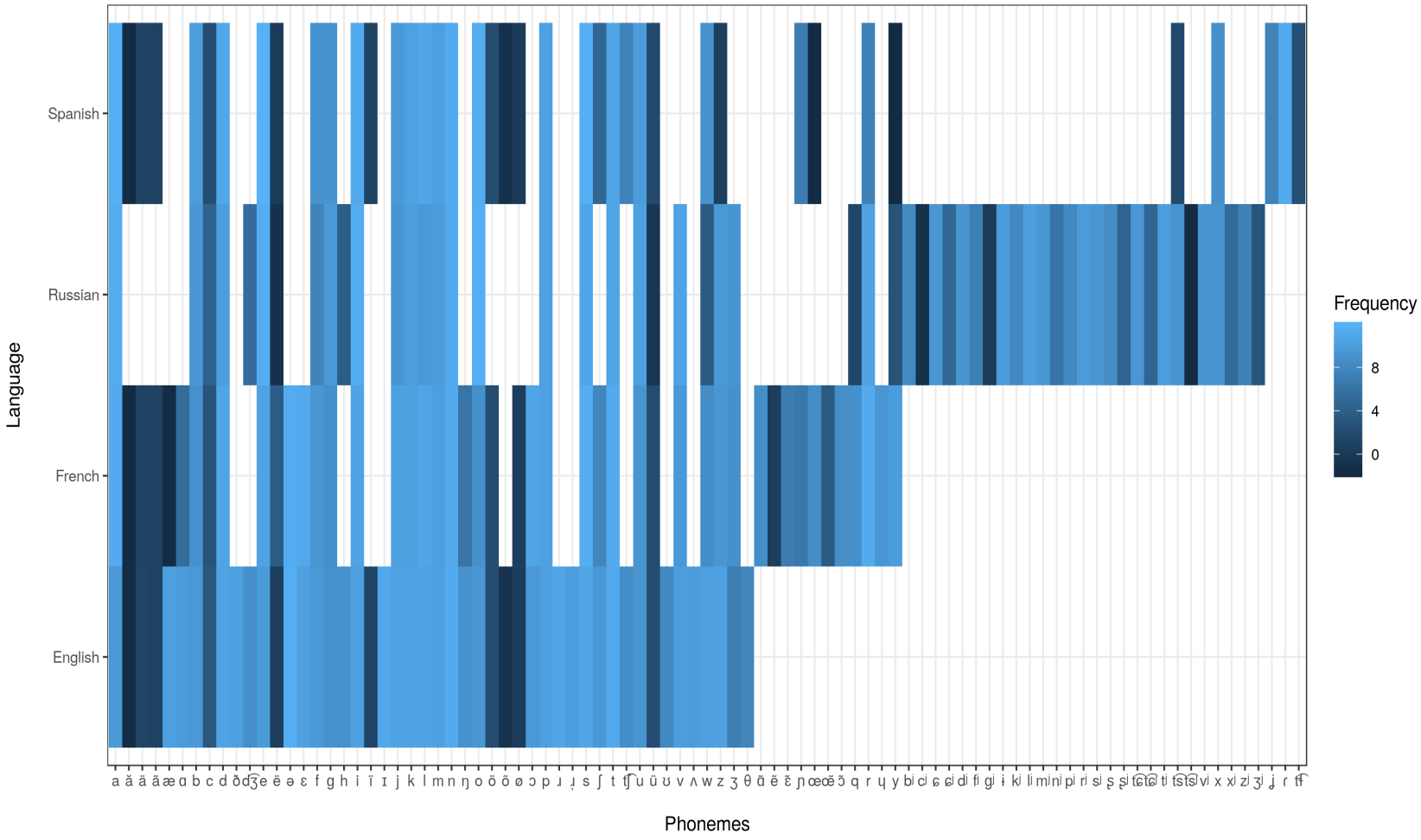

Figure 3.3. Frequency of phonemes across English, French, Russian, and Spanish.

Table 3.1

Means (standard deviations) of orthographic length, IPA length, and mean lexical frequency across the four languages of the gender agreement experiment.

	Train (100,000 *phrases*)			*Test* (2,000 *phrases*)		
	Ortho	*IPA*	*Freq*	*Ortho*	*IPA*	*Freq*
English	61.2 (22.1)	47.0 (16.4)	1.48 (1.18)	61.0 (22.2)	46.8 (16.4)	1.48 (1.18)
French	68.3 (24.3)	53.7 (19.0)	1.22 (1.10)	68.2 (24.3)	53.6 (19.0)	1.25 (1.09)
Russian	64.5 (22.9)	56.3 (19.6)	1.03 (0.993)	64.2 (22.5)	56.1 (19.4)	1.05 (0.995)
Spanish	68.6 (23.8)	57.8 (19.5)	1.98 (1.07)	68.4 (23.7)	57.7 (19.5)	2.00 (1.05)

3.1.2 Cross-Linguistic Similiarity across Phonemic Sequences

The way in which languages differ in terms of their phonemic inventory and the statistical patterns of phonemic sequences is potentially an influential variable. Certain language pairs may lead to a positive transfer of knowledge related to the statistics of phonemic sequences, while others may lead to greater levels of catastrophic interference, either due to incongruency between phonemic inventories or statistical inconsistencies.

In order to understand how the phonemic similarity between languages may influence L2 learning outcomes, an analysis between the aligned phrases in the training set was accomplished. The phonemic sequences of the aligned phrases for all four languages were used to calculate a similarity metric between each unique language pair, regardless of its order (e.g., L1 or L2) in the language learning process. A total of six similarity scores were created for the following language pairs: English-French, English-Russian, English-Spanish, French-Russian, French-Spanish,

Russian-Spanish. This metric was used to assess how similarity between the L1 and the L2 influenced L2 learning outcomes.

The generation of this similarity metric required aligned corpora, which is why this approach was not taken in the gender assignment experiment. The first step in this process was to combine the training sets for English, French, Russian, and Spanish into a large corpus of sequences. The Sequence Graph Transform (SGT) algorithm (Ranjan, Ebrahimi, & Paynabar, 2016) was used to extract short- and long-term sequence features in order to embed them into a finite-dimensional feature space. The SGT algorithm produced an embedding vector of 7,921 dimensions for each phrase. The mean Euclidean distance (see Equation 3.1) between the aligned phrases of each unique language pair was calculated. This process produced six similarity scores (English-French = -0.66, English-Russian = -1.38, English-Spanish = 0.22, French-Russian = -0.61, French-Spanish = 1.57, Russian-Spanish = 0.87). These scores are standardized, so the mean is 0.0 and the standard deviation is 1.0. French and Spanish were by far the most similar languages, while English and Russian were the least similar. English - French and French - Russian pairs had roughly the same amout of similarity in the negative direction, while Russian - Spanish had about half of the similarity as the French-Spanish pair.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3.1}$$

3.1.3 Training Procedure

Recurrent neural networks (see Table 2.3) representing language learning agents were trained to predict the next phoneme from a context of previously seen phonemes. The selection of hyperparameters for this experiment was largely influenced by previous studies, especially the gender agreement experiment by Monner et al. (2013). The learning rate was chosen so that the learning task could reach peak performance relatively early during the training phase. It was important that the models reached peak performance on the learning task during the training phase. Since the learning task in this experiment was more difficult than the gender assignment task, two hidden layers were used in the network architecture. Similar to the previous experiment, the objective of this experiment was not to optimize the performance of a model on a particular outcome. The objective was to show how performance is relative to linguistic and developmental factors. The selection of specific hyperparameters was not determined in a systematic fashion.

In order to speed up the training of each model, mini-batches were created. For each step during training, mini-batches of ten phrases were selected as input to the model. Each phrase was represented as a sequence of vectors consisting of 22 articulatory features representing unique phonemes. The ten phrases were selected randomly from the list of weighted phrases in the training set. The network produced ten output vectors representing the different possible phonemes (n=89) at every step of each phrase. The mean cross-entropy loss between the ten output vectors and the ten true target vectors was used during the backpropagation step of the training

phase. As described in the gender assignment experiment, the cross-entropy loss function (see Equation 2.8) measures the number of bits required to explain the difference between the estimated distribution ($\hat{y}$) and the true distribution (y). To perform this calculation, the activation values in the output layer are converted into a unit vector via the softmax function (see Equation 2.9). The softmax function normalizes a vector so that the sum of all components adds up to 1. Therefore, each component represents a probability. The normalized vector is then compared to the true vector associated with the input. This loss value was used to calculate the gradient for each parameter in the model. The SGD learning method was followed throughout the training process. The open-source machine learning library PyTorch (version 1.1.0) was used to train all models (Paszke et al., 2017).

Identical to the gender assignment experiment, model training was divided into two phases, a monolingual (i.e., entrenchment) phase and a bilingual (i.e., post-entrenchment) phase. The monolingual phase varied by the entrenchment level, while the bilingual phase always had a constant length of 2,000,000 nouns. The duration of the bilingual phase was kept constant to ensure each model would have enough time to reach ceiling performance in both languages regardless of the length of the entrenchment phase. Therefore, the total training time varied across the entrenchment levels.

In order to avoid catastrophic forgetting in the models due to a distributional shift in input characteristics, an equal probability of exposure to either L1 or L2 linguistic input was ensured during the bilingual phase. The interleaving approach addresses catastrophic forgetting in the network; however, it does not eliminate

competition and inference between the two languages.

The monolingual phase represented the entrenchment factor, which manipulates the quantity of L1 input prior to the introduction of L2 input. There were 6 levels of L1 entrenchment (t): 0, 200,000, 400,000, 600,000, 800,000, 1,000,000. The first level of L1 entrenchment, t=0, represented a balanced (or native) bilingual network since both L1 and L2 were present from the beginning of training. The last level of L1 entrenchment, t=1,000,000, represented a late L2 learner.

As in the gender assignment experiment, the monolingual and bilingual training phases were followed under four different memory development conditions: 1) No Growth, 2) Unit Growth, 3) Unit Replacement, 4) Connection Growth (see Figure 2.5). The goal of the memory development factor was to manipulate the architecture of the model in order to mirror memory development in humans. A detailed description of the memory development factor is provided in the methods section for the gender assignment experiment.

Ten networks were trained in each cell of the design matrix. Six levels of L1 entrenchment under four different memory development conditions yielded 240 models for each language pair. A total of 12 language pairs (L1 English - L2 French, L1 English - L2 Russian, L1 English - L2 Spanish, L1 French - L2 English, L1 French - L2 Russian, L1 French - L2 Spanish, L1 Russian - L2 English, L1 Russian - L2 French, L1 Russian - L2 Spanish, L1 Spanish - L2 English, L1 Spanish - L2 French, L1 Spanish - L2 Russian) led to 2,880 models across all conditions.

3.1.4 Evaluation Criteria

Each model was evaluated on the L1 and L2 test sets every 100,000 phrases. This experiment had two evaluation tasks, one intrinsic and one extrinsic. These tasks are described below. All formal analyses and reporting of descriptive statistics used the F1 score (Equation 2.13).

3.1.4.1 Intrinsic Task

In machine learning applications, an intrinsic evaluation task is one that is based upon a task identical to that performed during training. The model has been optimized to solve this problem by minimizing the error calculated during training through the stochastic gradient descent method. Therefore, the intrinsic task in this experiment is simply the prediction of the next phoneme in a sequence. For this experiment, the test set used in the intrinsic evaluation task consisted of 2,000 phrases with sequence lengths between 20 and 100 (see Table 3.1 for descriptive statistics of the test sets). Performance on the task was measured using the top-k approach (here, $k = 4$). If the unit corresponding to the correct phoneme was one of the top-k units in the output vector, the prediction was marked as correct.

3.1.4.2 Extrinsic Task

A similar version of the extrinsic task reported in Monner et al.(2013) was used in the present study. This task required that the model produce the last two phonemes of a noun-adjective pair. A noun and the neutral portion of a paired

adjective were presented to the model. The model was tasked to produce the last two phonemes of the noun-adjective pair. The goal was to complete the adjective using the gender-appropriate phonemes. As in the intrinsic task, performance on the extrinsic task was measured using the top-k approach ($k = 4$). If the unit corresponding to the correct phoneme was one of the top-k units in the output vector, the prediction was marked as correct. The objective was to evaluate how well models are able to use formal cues in non-adjacent dependent structures. Noun-adjective pairs are common in French and Spanish, but not in Russian. The typical order is reversed in Russian (i.e., adjective-noun). Since the noun-adjective word order is seldom encountered in Russian, Russian noun-adjective pairs were not included in the extrinsic task. Since grammatical gender in nouns is not present in English, model performance on English was not evaluated with an extrinsic task. See Table 3.2 for descriptive statistics related to the extrinsic evaluation test set.

Table 3.2

Means (standard deviations) of orthographic length, IPA length, and mean lexical frequency for each gender class across the four languages of the extrinsic task dataset in the gender agreement experiment.

	N	*Ortho*	*IPA*	*Freq*
French				
f	900	19.2 (4.08)	17.9 (3.96)	1.17 (0.541)
m	900	17.2 (3.83)	16.0 (3.64)	1.28 (0.560)
Russian				
f	600	18.4 (3.73)	21.1 (4.65)	1.04 (0.404)
m	600	18.1 (3.75)	20.6 (4.72)	1.01 (0.403)
n	600	20.3 (3.98)	23.1 (4.82)	0.995 (0.424)
Spanish				
f	900	19.5 (3.90)	19.5 (3.91)	0.842 (0.508)
m	900	17.6 (3.32)	17.6 (3.31)	0.987 (0.524)

3.2 Results

3.2.1 Monolingual Baseline

Ten monolingual models were trained for each of the 12 language pairs under each of the four memory development conditions in order to establish a monolingual baseline for each language and a corresponding naive L2 baseline for each language. The naive L2 baseline is determined by evaluating L2 performance in monolingual models prior to explicitly training on the L2. This led to 480 monolingual models trained on 1,000,000 phrases. Model performance was evaluated on the test set every 100,000 phrases, so each model had ten data points in which model performance on the test set was measured. The mean and standard deviation of the final F1 score for each language and memory condition for both the intrinsic and extrinsic tasks is reported in Table 3.3. Figures 3.4 and 3.5 show the baseline performance in the intrinsic and extrinsic tasks for each language in the L1 and the L2 across all four memory conditions.

Separate mixed effects models were fit to the F1 score associated with the L1 (monolingual) and the L2 (naive L2 baseline) in order to identify differences between languages and memory development conditions. The analysis of the naive L2 baseline was conducted separately for each L2. Separate models are fit for the intrinsic and extrinsic evaluation tasks. All models were fit to the data using the lme4 package (version 1.1-21; Bates et al., 2015) within the R statistical platform (version 3.5.2; R Core Team, 2015).

Table 3.3

Mean (standard deviation) final F1 score in the gender agreement experiment for monolingual and naive L2 baselines of each language in each memory development condition. The monolingual (L1) values were gathered after 1,000,000 nouns were used to train the model on the L1. The naive L2 baseline values represent performance on an L2 when no training occurred for that L2. Results are provided for the intrinsic task and the extrinsic task.

		No Growth	*Connection Growth*	*Unit Growth*	*Unit Replacement*
Intrinsic					
	English				
	L1	0.64 (0.08)	0.35 (0.01)	0.36 (0.02)	0.55 (0.09)
	L2	0.09 (0.01)	0.10 (0.02)	0.10 (0.02)	0.08 (0.02)
	French				
	L1	0.48 (0.11)	0.29 (0.01)	0.29 (0.02)	0.41 (0.08)
	L2	0.11 (0.02)	0.11 (0.01)	0.12 (0.01)	0.10 (0.02)
	Russian				
	L1	0.52 (0.10)	0.24 (0.01)	0.24 (0.01)	0.46 (0.08)
	L2	0.07 (0.02)	0.07 (0.01)	0.08 (0.02)	0.07 (0.02)
	Spanish				
	L1	0.62 (0.04)	0.34 (0.01)	0.34 (0.01)	0.54 (0.07)
	L2	0.10 (0.03)	0.10 (0.03)	0.11 (0.03)	0.10 (0.03)
Extrinsic					
	French				
	L1	0.51 (0.13)	0.32 (0.01)	0.32 (0.02)	0.42 (0.10)
	L2	0.10 (0.02)	0.11 (0.03)	0.13 (0.02)	0.10 (0.02)
	Russian				
	L1	0.21 (0.01)	0.24 (0.02)	0.22 (0.03)	0.20 (0.02)
	L2	0.05 (0.02)	0.06 (0.02)	0.08 (0.03)	0.05 (0.02)
	Spanish				
	L1	0.61 (0.04)	0.45 (0.02)	0.44 (0.03)	0.53 (0.07)
	L2	0.09 (0.03)	0.10 (0.04)	0.11 (0.04)	0.09 (0.04)

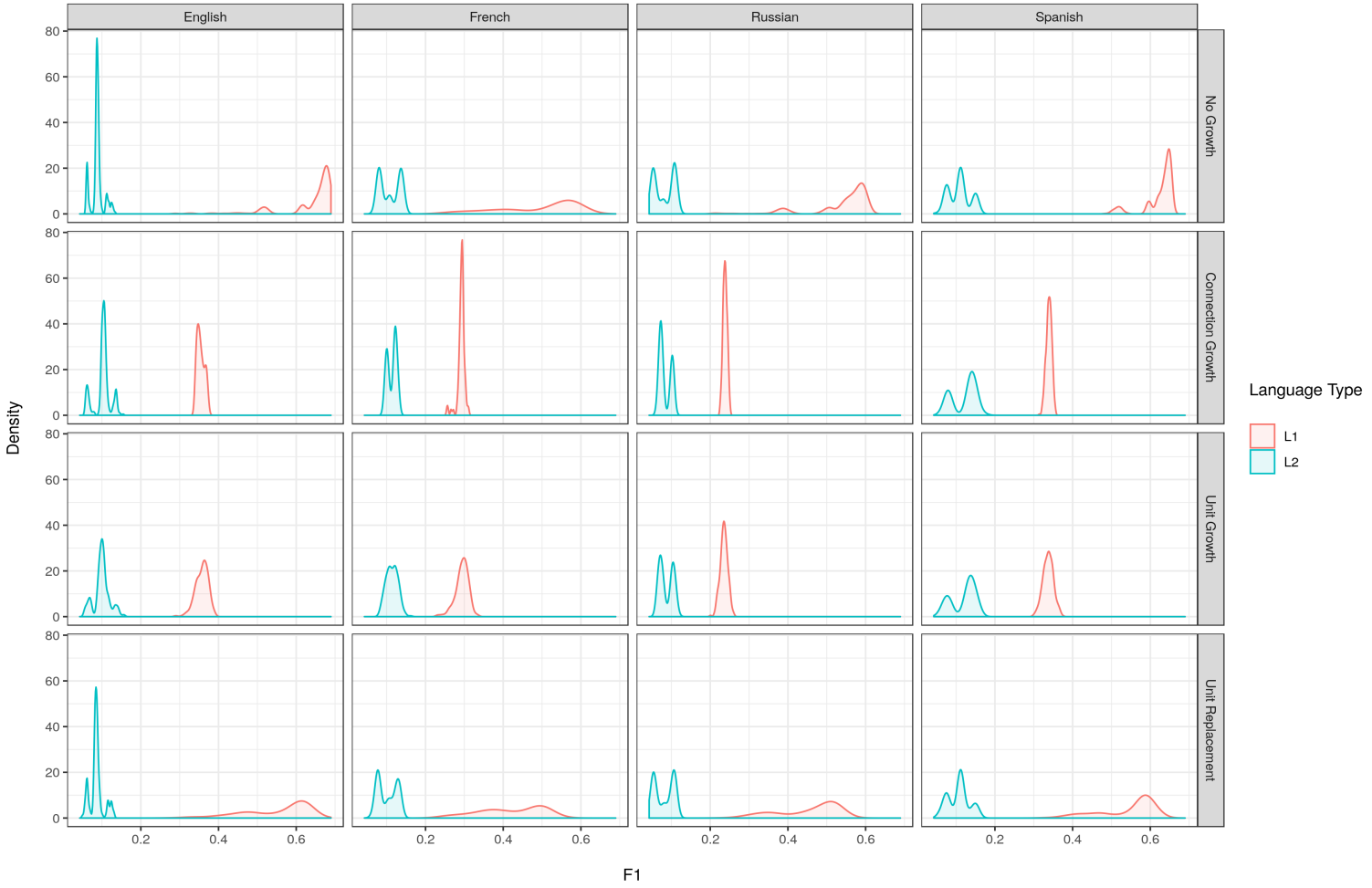

Figure 3.4. Density plots visualizing the distribution of F1 scores of monolingual and naive L2 models in the intrinsic task of the gender agreement experiment.

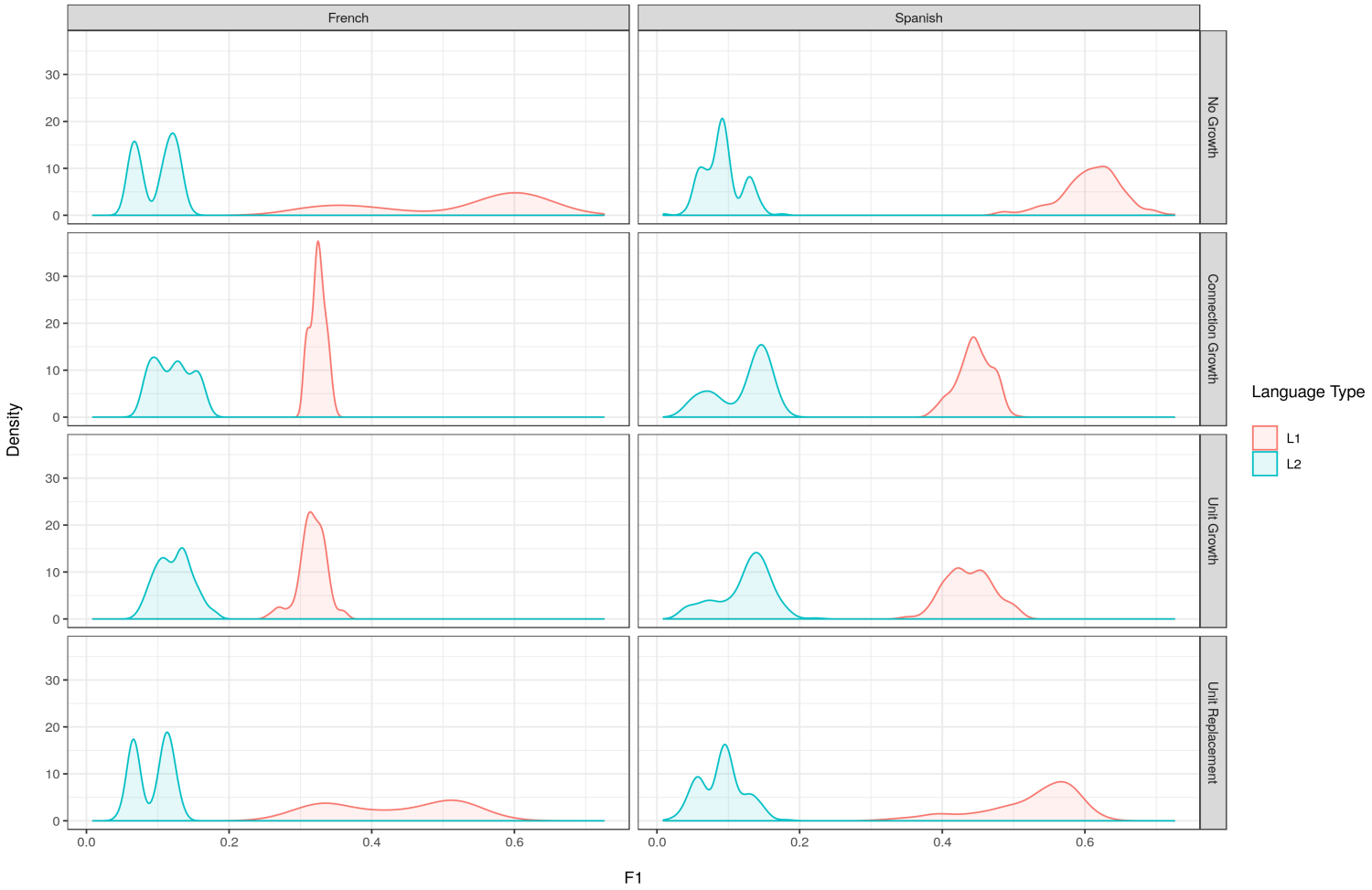

Figure 3.5. Density plots visualizing the distribution of F1 scores of monolingual and naive L2 models in the extrinsic task of the gender agreement experiment.

In the monolingual models, contrast coding was used for the language (here, the L1) and memory development factors. The reference for both the intrinsic and extrinsic monolingual models corresponds to French in the No Growth condition. All fixed effects across the other factors are relative to these reference levels. The performance of each model was evaluated on the L1 and the naive L2 every 100,000 nouns. These data points were collected at different moments in the training phase. Therefore, each model had ten evaluation runs for each language (repeated measurements of the dependent variable). The maximal random effects structure included random intercepts for the grouping factors of model and each evaluation run (the repeated measures grouping factor) (Barr et al., 2013). This random effects structure was used for all mixed effects models.

In the monolingual mixed effects model of the intrinsic task (see Table 3.4 and Figure 3.6), there was a main effect of language. In the No Growth condition, English, Russian, and Spanish monolingual models consistently outperformed French monolingual models. English and Spanish performed nearly identical in the No Growth and Unit Replacement conditions. The Connection Growth, Unit Growth, and Unit Replacement conditions led to poorer outcomes across all languages (a main effect of memory development). The Connection Growth and Unit Growth conditions had an effect size roughly two times greater than the Unit Replacement condition. In Russian, the negative effect of Connection Growth and Unit Growth was greater than it was in the other languages (an interaction between memory development and language).

The results of the monolingual mixed effects model of the extrinsic task were

very similar to those in the intrinsic task model (see Table 3.5 and Figure 3.6). In the No Growth condition for French monolingual models, Spanish models had better performance outcomes than French (a main effect of language). The Connection Growth, Unit Growth, and Unit Replacement conditions led to poorer outcomes in French and Spanish (a main effect of memory condition).

Table 3.4

Mixed effects model examining the effect of memory development and language on learning outcomes in monolingual models in the intrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.48	0.01	32.09*	0.001	0.034	0.001	0.03
Connection Growth	-0.19	0.02	-10.28*	-	-	-	-
Unit Growth	-0.19	0.02	-11.45*	-	-	-	-
Unit Replacement	-0.07	0.02	-4.14*	-	-	-	-
L1 English	0.16	0.01	10.85*	-	-	-	-
L1 Russian	0.04	0.02	2.65*	-	-	-	-
L1 Spanish	0.14	0.01	10.80*	-	-	-	-
Connection Growth x L1 English	-0.09	0.02	-4.18*	-	-	-	-
Unit Growth x L1 English	-0.09	0.02	-4.49*	-	-	-	-
Unit Replacement x L1 English	-0.02	0.02	-0.90	-	-	-	-
Connection Growth x L1 Russian	-0.10	0.02	-3.98*	-	-	-	-
Unit Growth x L1 Russian	-0.10	0.02	-4.39*	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.02	0.02	-	-	-	-
Connection Growth x L1 Spanish	-0.10	0.02	-4.61*	-	-	-	-
Unit Growth x L1 Spanish	-0.10	0.02	-5.25*	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.02	-0.71	-	-	-	-

Note: Mixed effects model formula: F1 $\sim$ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Starting with a smaller working memory capacity (Unit Growth) did not lead to better outcomes than starting with a fully developed network, like in the No Growth or Unit Replacement conditions. Connection Growth and Unit Growth conditions had very similar impacts on model performance in both tasks. This result suggests that long-term memory capacity is important for successful learning outcomes and that a lack of long-term resources during the early stages of develop-

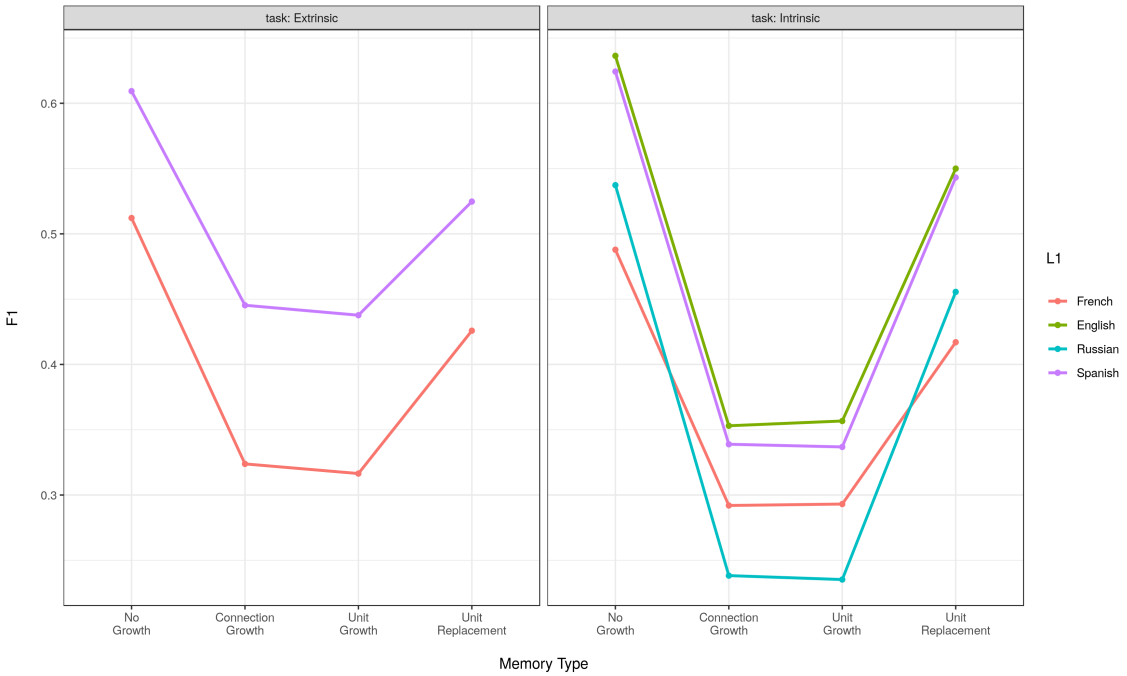

Figure 3.6. Plot visualizing the interaction between memory development condition and L1 in the monolingual models of the gender agreement experiment.

Table 3.5

Mixed effects model examining the effect of memory development and language on learning outcomes in monolingual models in the extrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.50	0.01	41.09*	0.001	0.029	0.000	0.019
Connection Growth	-0.18	0.02	-10.62*	-	-	-	-
Unit Growth	-0.19	0.02	-12.39*	-	-	-	-
Unit Replacement	-0.08	0.02	-5.37*	-	-	-	-
L1 Russian	-0.29	0.02	-19.07*	-	-	-	-
L1 Spanish	0.10	0.01	8.30*	-	-	-	-
Connection Growth x L1 Russian	0.20	0.02	8.93*	-	-	-	-
Unit Growth x L1 Russian	0.19	0.02	8.98*	-	-	-	-
Unit Replacement x L1 Russian	0.06	0.02	2.92*	-	-	-	-
Connection Growth x L1 Spanish	0.02	0.02	1.02	-	-	-	-
Unit Growth x L1 Spanish	0.02	0.02	1.16	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.02	0.08	-	-	-	-

Note: Mixed effects model formula: F1 $\sim$ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

ment leads to poorer learning outcomes.

A naive baseline in which the L2 was not explicitly trained was established for each L2 for both the intrinsic and extrinsic tasks. The purpose of examining the performance of networks that have not been trained on an L2 was to determine how a particular L1 may influence the learning outcomes of a particular L2. The L2 of each monolingual model was evaluated every 100,000 nouns. This resulted in ten evaluation runs for the L2 in each model (repeated measurements of the dependent variable). For each evaluation ask, a mixed effects model was fit to each L2. The random effects structure included random intercepts for the grouping factors of model and evaluation run. First, the naive L2 performance on the intrinsic task is reported.

In the naive L2 English model in the intrinsic task (see Table 3.6 and Figure 3.7), L1 French and L1 Spanish led to better naive L2 English outcomes than L1 Russian (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth led to better naive L2 English outcomes (a main effect of memory development).

In the naive L2 French model in the intrinsic task (see Table 3.6 and Figure 3.7), L1 Spanish led to better naive L2 French outcomes while L1 Russian led to worse outcomes compared to L1 English (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth led to better naive L2 French outcomes (a main effect of memory development). This effect was only seen when the L1 was either Russian or English (an interaction between memory development and L1). When the L1 was Spanish in the Connection Growth and

Unit Growth conditions, naive L2 French outcomes decreased.

In the naive L2 Russian model in the intrinsic task (see Table 3.6 and Figure 3.7), L1 Spanish led to better naive L2 Russian outcomes while L1 English led to worse outcomes compared to L1 French (a main effect of L1). When the L1 was English in the Connection Growth and Unit Growth conditions, naive L2 Russian outcomes increased (an interaction between memory development and L1).

In the naive L2 Spanish model in the intrinsic task (see Table 3.6 and Figure 3.7), L1 English and Russian led to worse naive L2 Spanish outcomes relative to L1 French (a main effect of L1). When the L1 was Russian in the Connection Growth and Unit Growth conditions, naive L2 Spanish outcomes increased (an interaction between memory development and L1).

In the naive L2 French model in the extrinsic task, L1 Spanish led to better naive L2 French outcomes than either L1 English or L1 Russian (a main effect of L1). In the naive L2 Spanish model in the extrinsic task, a main effect of L1 was also obtained. L1 French led to better L2 Spanish outcomes than either L1 English or L1 Russian. In all of the naive L2 baseline models in the extrinsic task, there was a main effect of memory development. The Connection Growth and Unit Growth conditions led to better outcomes in the naive L2.

Table 3.6

Mixed effects models examining the effect of memory development and L1 on learning outcomes in naive L2 models in the intrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
English							
(Intercept)	0.12	0.00	69.95*	-	-	-	-
Connection Growth	0.02	0.00	6.96*	-	-	-	-
Unit Growth	0.01	0.00	6.11*	-	-	-	-
Unit Replacement	0.00	0.00	0.62	-	-	-	-
L1 Russian	-0.06	0.00	-27.29*	-	-	-	-
L1 Spanish	-0.03	0.00	-17.95*	-	-	-	-
Connection Growth x L1 Russian	-0.02	0.00	-5.11*	-	-	-	-
Unit Growth x L1 Russian	-0.01	0.00	-3.33*	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.00	-1.36	-	-	-	-
Connection Growth x L1 Spanish	0.00	0.00	0.16	-	-	-	-
Unit Growth x L1 Spanish	0.00	0.00	-0.55	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.00	-1.74	-	-	-	-
French							
(Intercept)	0.11	0.00	90.04*	0.000	0.001	0.000	0.001
Connection Growth	-0.01	0.00	-10.96*	-	-	-	-
Unit Growth	0.01	0.00	7.28*	-	-	-	-
Unit Replacement	0.00	0.00	-1.99	-	-	-	-
L1 Russian	-0.03	0.00	-19.32*	-	-	-	-
L1 Spanish	0.03	0.00	21.06*	-	-	-	-
Connection Growth x L1 Russian	0.04	0.00	21.08*	-	-	-	-
Unit Growth x L1 Russian	0.01	0.00	4.33*	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.00	0.09	-	-	-	-
Connection Growth x L1 Spanish	-0.03	0.00	-12.56*	-	-	-	-
Unit Growth x L1 Spanish	-0.02	0.00	-11.77*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.00	-2.20*	-	-	-	-
Russian							
(Intercept)	0.05	0.00	76.40*	0.000	0.001	0.000	0.000
Connection Growth	0.02	0.00	19.40*	-	-	-	-
Unit Growth	0.02	0.00	18.12*	-	-	-	-
Unit Replacement	0.00	0.00	0.61	-	-	-	-
L1 French	0.03	0.00	18.64*	-	-	-	-
L1 Spanish	0.05	0.00	56.39*	-	-	-	-
Connection Growth x L1 French	-0.02	0.00	-12.03*	-	-	-	-
Unit Growth x L1 French	-0.02	0.00	-10.70*	-	-	-	-
Unit Replacement x L1 French	0.00	0.00	-0.28	-	-	-	-
Connection Growth x L1 Spanish	-0.02	0.00	-16.97*	-	-	-	-
Unit Growth x L1 Spanish	-0.02	0.00	-15.36*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.00	-1.99	-	-	-	-
Spanish							
(Intercept)	0.32	0.02	13.33*	0.000	0.000	0.000	0.000
Connection Growth	-0.14	0.04	-3.43*	-	-	-	-
Unit Growth	-0.10	0.03	-2.89*	-	-	-	-
Unit Replacement	-0.04	0.03	-1.20	-	-	-	-
L1 English	0.02	0.03	0.68	-	-	-	-
L1 Russian	0.01	0.04	0.35	-	-	-	-
Connection Growth x L1 English	0.02	0.05	0.41	-	-	-	-
Unit Growth x L1 English	-0.03	0.04	-0.59	-	-	-	-
Unit Replacement x L1 English	0.00	0.04	0.11	-	-	-	-
Connection Growth x L1 Russian	0.02	0.05	0.41	-	-	-	-
Unit Growth x L1 Russian	-0.05	0.05	-0.92	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.05	0.10	-	-	-	-

Note: Mixed effects model formula: F1 ~ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

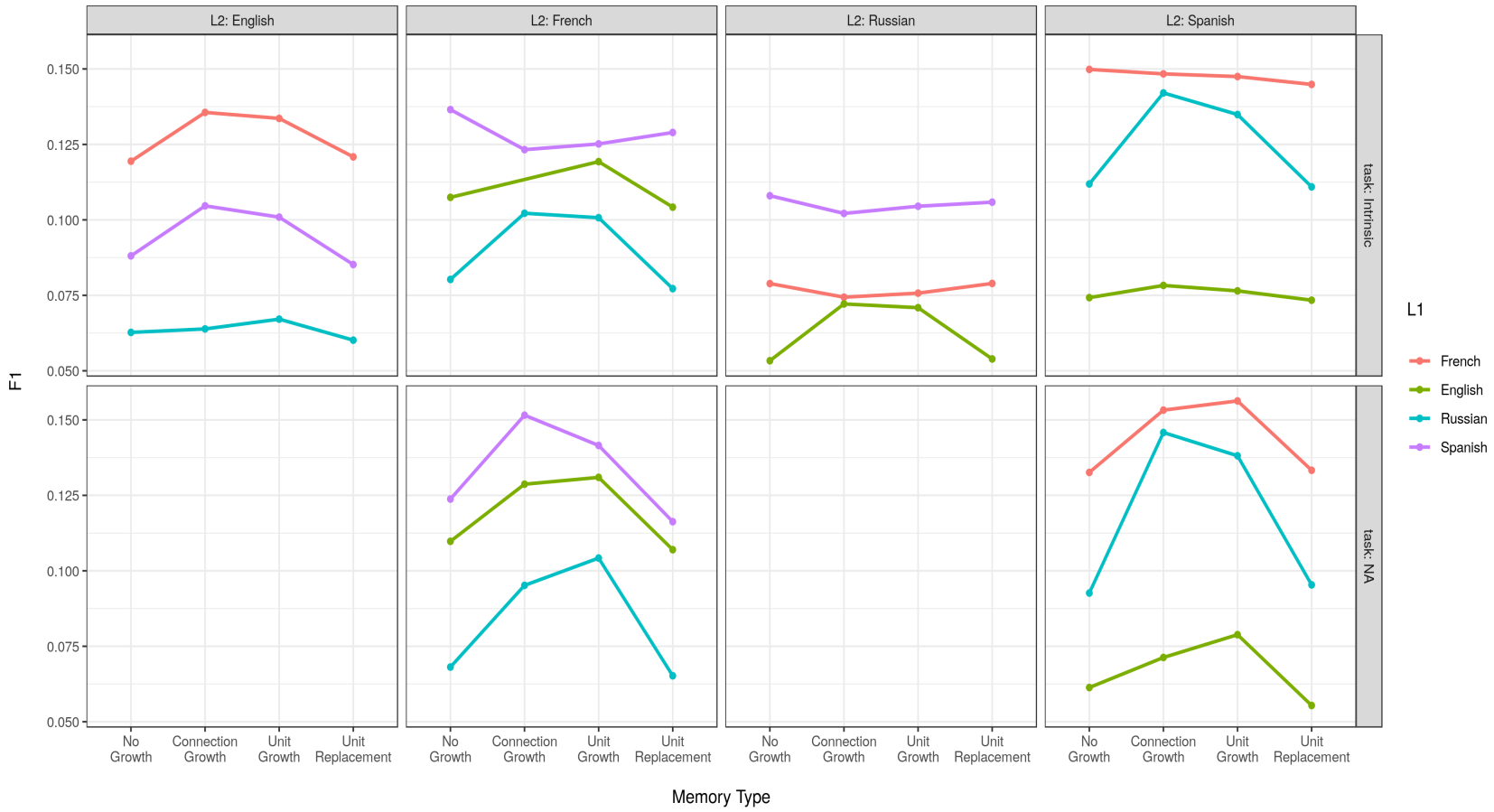

Figure 3.7. Plot visualizing the interaction between memory development condition and L1 in the naive L2 models of the gender agreement experiment.

Table 3.7

Mixed effects models examining the effect of memory development and L1 on learning outcomes in naive L2 models in the extrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
French							
(Intercept)	0.11	0.02	5.17*	0.000	0.000	0.000	0.000
Connection Growth	-0.06	0.03	-1.85	-	-	-	-
Unit Growth	0.02	0.03	0.71	-	-	-	-
Unit Replacement	-0.00	0.03	-0.09	-	-	-	-
L1 Russian	0.03	0.03	1.19	-	-	-	-
L1 Spanish	0.25	0.03	9.74*	-	-	-	-
Connection Growth x L1 Russian	0.09	0.04	2.08*	-	-	-	-
Unit Growth x L1 Russian	0.00	0.04	0.04	-	-	-	-
Unit Replacement x L1 Russian	-0.01	0.04	-0.37	-	-	-	-
Connection Growth x L1 Spanish	0.09	0.04	2.08*	-	-	-	-
Unit Growth x L1 Spanish	-0.09	0.04	-2.32*	-	-	-	-
Unit Replacement x L1 Spanish	-0.04	0.04	-1.10	-	-	-	-
Spanish							
(Intercept)	0.06	0.01	4.13*	0.000	0.000	0.000	0.000
Connection Growth	0.01	0.02	0.45	-	-	-	-
Unit Growth	0.02	0.02	0.84	-	-	-	-
Unit Replacement	-0.01	0.02	-0.28	-	-	-	-
L1 French	0.26	0.02	12.38*	-	-	-	-
L1 Russian	0.09	0.02	4.04*	-	-	-	-
Connection Growth x L1 French	0.03	0.03	0.76	-	-	-	-
Unit Growth x L1 French	-0.10	0.03	-3.43*	-	-	-	-
Unit Replacement x L1 French	-0.04	0.03	-1.25	-	-	-	-
Connection Growth x L1 Russian	0.03	0.03	0.76	-	-	-	-
Unit Growth x L1 Russian	-0.00	0.03	-0.00	-	-	-	-
Unit Replacement x L1 Russian	-0.00	0.03	-0.07	-	-	-	-

Note: Mixed effects model formula: F1 ~ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

3.2.2 Bilingual Models

Two thousand eight hundred eighty models were trained to predict the next phoneme given a sequence of phonemes represented as a sequence of articulatory feature vectors; ten training runs per experimental condition (six entrenchment levels, four memory conditions, and 12 language pairs). The performance of each model was evaluated on the intrinsic test set (2,000 phrases) and extrinsic test set

(1,200 noun-adjective pairs; but not in English) of each language every 100,000 phrases and at the end of the bilingual training phase. The F1 score was used to perform statistical analyses and report descriptive statistics. The mean and standard deviation of the final F1 score for each language, entrenchment level, and memory condition is reported in Table 3.8. Figure 3.8 shows the distributions of L2 performance across all four memory conditions in the intrinsic task.

As in the gender assignment experiment, the factor of entrenchment was simplified in the analysis by reducing the factor to three levels (Early, Middle, and Late L2 learners). To understand how the experimental factors influenced L2 learning outcomes, mixed effects models were fit to the F1 score associated with the L2. Separate analyses were performed on the two test sets. Mixed effects models were fit to the fixed factors of entrenchment, memory development, and L1 in each L2. The intrinsic test set included data from all four languages, while the extrinsic test set only included noun-adjective pairs for French, Russian, and Spanish.

3.2.2.1 Analysis of the Intrinsic Evaluation Task

A mixed effect model was fit to the fixed factors of entrenchment, memory development, and L1 in each of the four languages. Contrast coding was used for the entrenchment, memory development, and L1 factors. The L2 of each model was evaluated every 100,000 nouns during the bilingual phase. Therefore, each model had 20 evaluation runs for the L2 (repeated measurements of the dependent variable). Models were fit to the data using the *lme4* package (version 1.1-21; Bates

et al., 2015) within the R statistical platform (version 3.5.2; R Core Team, 2015). The random effects structure included random intercepts for the grouping factors of model and evaluation run. This random effects structure was used for all mixed effects models.

In the L2 English model (see Table 3.9 and Figure 3.8), L1 Spanish led to better L2 English outcomes than L1 French or L1 Russian in the No Growth condition for early L2 learners (a main effect of L1). Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 English outcomes. Connection Growth, Unit Growth, and Unit Replacement negatively impacted L2 English outcomes in each L1 (a main effect of memory development).

In the L2 French model (see Table 3.10 and Figure 3.8), L1 Spanish led to better L2 French outcomes than L1 English or L1 Russian in the No Growth condition for early L2 learners (a main effect of L1). Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 French outcomes. Connection Growth, Unit Growth, and Unit Replacement negatively impacted L2 French outcomes in each L1 (a main effect of memory development).

In the L2 Russian model (see Table 3.11 and Figure 3.8), L1 Spanish led to better L2 Russian outcomes than L1 English or L1 French in the No Growth condition for early L2 learners (a main effect of L1). L1 French led to slightly worse L2 Russian outcomes than L1 English. Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 Russian outcomes. Connection Growth, Unit Growth, and Unit Replacement negatively impacted L2 Russian outcomes in each L1 (a main effect of memory development).

Table 3.8

Intrinsic and extrinsic task means (standard deviations) of final F1 scores for L2 outcomes of each language across entrenchment and memory development conditions in the gender agreement experiment.

		No Growth		*Connection Growth*		*Unit Growth*		*Unit Replacement*	
		L1	*L2*	*L1*	*L2*	*L1*	*L2*	*L1*	*L2*
Intrinsic									
English									
	Early	0.52 (0.08)	0.55 (0.08)	0.31 (0.02)	0.31 (0.02)	0.29 (0.03)	0.30 (0.03)	0.44 (0.08)	0.46 (0.09)
	Middle	0.54 (0.07)	0.54 (0.08)	0.31 (0.01)	0.31 (0.02)	0.29 (0.03)	0.30 (0.04)	0.46 (0.07)	0.46 (0.07)
	Late	0.55 (0.06)	0.55 (0.07)	0.31 (0.02)	0.31 (0.02)	0.30 (0.03)	0.30 (0.03)	0.47 (0.07)	0.46 (0.06)
French									
	Early	0.40 (0.08)	0.43 (0.06)	0.25 (0.01)	0.25 (0.01)	0.24 (0.01)	0.24 (0.02)	0.34 (0.07)	0.36 (0.06)
	Middle	0.42 (0.07)	0.43 (0.06)	0.25 (0.01)	0.25 (0.01)	0.24 (0.01)	0.24 (0.02)	0.36 (0.06)	0.36 (0.06)
	Late	0.45 (0.06)	0.43 (0.06)	0.25 (0.01)	0.25 (0.01)	0.24 (0.02)	0.24 (0.02)	0.38 (0.06)	0.35 (0.05)
Russian									
	Early	0.45 (0.08)	0.42 (0.09)	0.17 (0.01)	0.17 (0.01)	0.17 (0.02)	0.17 (0.02)	0.35 (0.08)	0.34 (0.09)
	Middle	0.47 (0.07)	0.42 (0.08)	0.17 (0.01)	0.17 (0.01)	0.17 (0.02)	0.17 (0.02)	0.38 (0.08)	0.33 (0.07)
	Late	0.48 (0.06)	0.42 (0.06)	0.17 (0.01)	0.17 (0.01)	0.17 (0.02)	0.17 (0.02)	0.40 (0.06)	0.34 (0.07)
Spanish									
	Early	0.42 (0.04)	0.41 (0.06)	0.26 (0.01)	0.26 (0.02)	0.23 (0.03)	0.25 (0.03)	0.36 (0.06)	0.35 (0.06)
	Middle	0.44 (0.04)	0.40 (0.05)	0.26 (0.01)	0.25 (0.01)	0.24 (0.03)	0.24 (0.03)	0.38 (0.04)	0.34 (0.05)
	Late	0.44 (0.04)	0.40 (0.05)	0.26 (0.01)	0.25 (0.02)	0.24 (0.03)	0.24 (0.03)	0.39 (0.04)	0.34 (0.05)
Extrinsic									
French									
	Early	0.44 (0.10)	0.49 (0.08)	0.28 (0.01)	0.28 (0.02)	0.28 (0.02)	0.28 (0.03)	0.37 (0.08)	0.39 (0.08)
	Middle	0.46 (0.10)	0.49 (0.06)	0.28 (0.01)	0.28 (0.02)	0.28 (0.02)	0.27 (0.03)	0.38 (0.08)	0.39 (0.06)
	Late	0.50 (0.07)	0.49 (0.06)	0.28 (0.01)	0.28 (0.01)	0.28 (0.02)	0.27 (0.03)	0.40 (0.07)	0.38 (0.06)
Spanish									
	Early	0.56 (0.06)	0.53 (0.07)	0.35 (0.04)	0.35 (0.04)	0.33 (0.06)	0.33 (0.06)	0.48 (0.08)	0.46 (0.09)
	Middle	0.55 (0.06)	0.53 (0.06)	0.35 (0.04)	0.35 (0.03)	0.31 (0.07)	0.32 (0.05)	0.49 (0.07)	0.47 (0.07)
	Late	0.56 (0.06)	0.53 (0.06)	0.36 (0.03)	0.34 (0.04)	0.31 (0.06)	0.32 (0.05)	0.49 (0.07)	0.46 (0.07)

Table 3.9

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 English outcomes in the intrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.50	0.01	46.01*	0.001	0.031	0.001	0.031
Middle L2 Learner	-0.01	0.01	-0.97	-	-	-	-
Late L2 Learner	-0.02	0.02	-1.14	-	-	-	-
Connection Growth	-0.19	0.01	-13.60*	-	-	-	-
Unit Growth	-0.20	0.01	-15.51*	-	-	-	-
Unit Replacement	-0.07	0.01	-5.43*	-	-	-	-
L1 Russian	-0.01	0.01	-1.00	-	-	-	-
L1 Spanish	0.10	0.01	9.30*	-	-	-	-
Middle L2 Learner x Connection Growth	-0.01	0.02	-0.30	-	-	-	-
Late L2 Learner x Connection Growth	-0.01	0.02	-0.63	-	-	-	-
Middle L2 Learner x Unit Growth	-0.01	0.02	-0.29	-	-	-	-
Late L2 Learner x Unit Growth	-0.01	0.02	-0.31	-	-	-	-
Middle L2 Learner x Unit Replacement	-0.01	0.02	-0.66	-	-	-	-
Late L2 Learner x Unit Replacement	-0.02	0.02	-0.95	-	-	-	-
Middle L2 Learner x L1 Russian	0.00	0.02	-0.05	-	-	-	-
Late L2 Learner x L1 Russian	-0.01	0.02	-0.32	-	-	-	-
Middle L2 Learner x L1 Spanish	-0.01	0.02	-0.43	-	-	-	-
Late L2 Learner x L1 Spanish	-0.02	0.02	-0.99	-	-	-	-
Connection Growth x L1 Russian	0.01	0.02	0.68	-	-	-	-
Unit Growth x L1 Russian	0.01	0.02	0.31	-	-	-	-
Unit Replacement x L1 Russian	-0.01	0.02	-0.67	-	-	-	-
Connection Growth x L1 Spanish	-0.09	0.02	-5.77*	-	-	-	-
Unit Growth x L1 Spanish	-0.07	0.01	-4.84*	-	-	-	-
Unit Replacement x L1 Spanish	-0.02	0.02	-1.56	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.00	0.03	0.10	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.01	0.03	0.39	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.00	0.03	0.14	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.03	0.14	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.01	0.03	0.33	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	0.03	0.03	1.03	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.01	0.02	0.40	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.02	0.03	0.93	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	0.00	0.02	-0.07	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.01	0.02	0.32	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.02	0.02	0.63	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.03	0.02	1.17	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

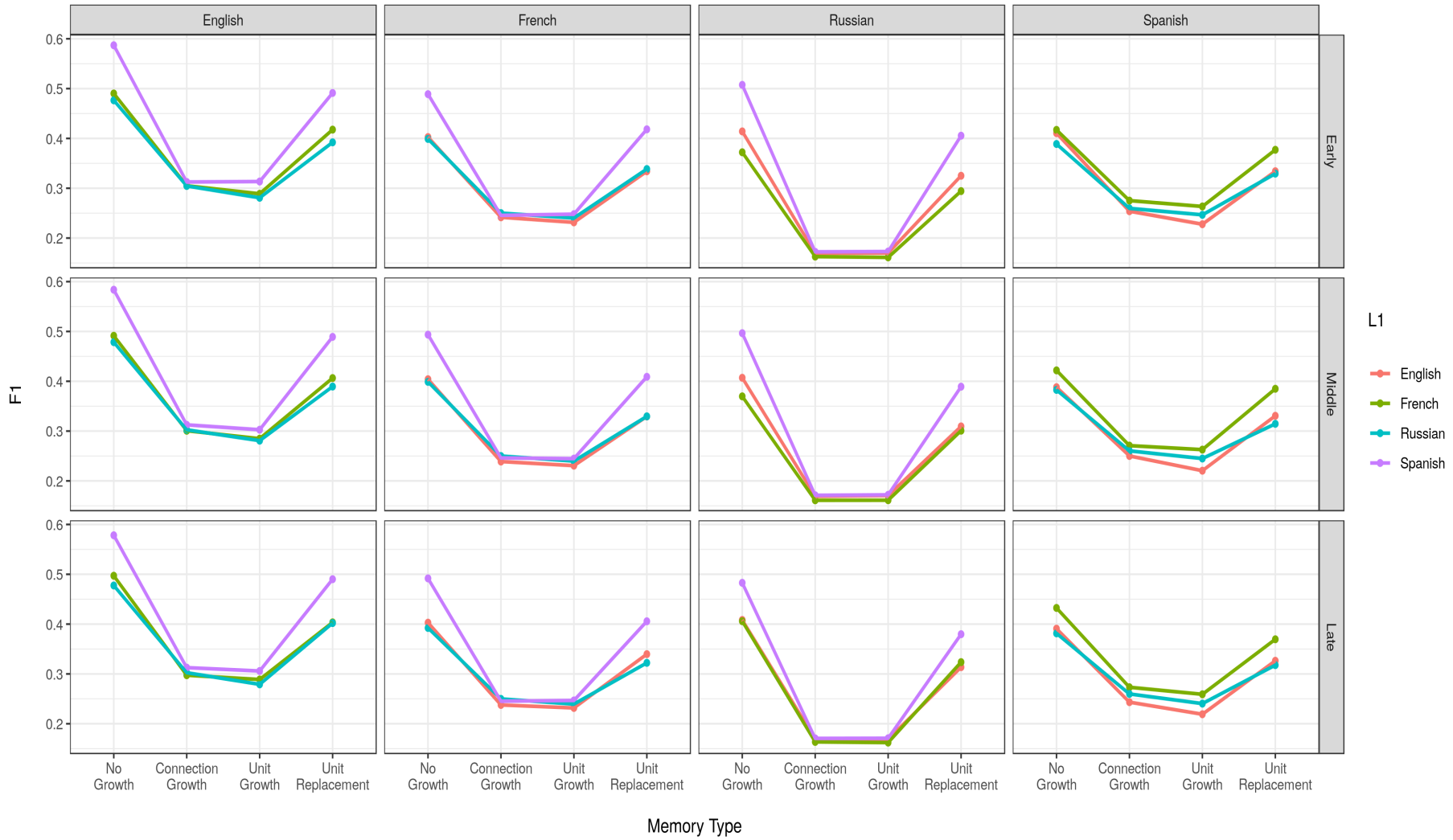

Figure 3.8. Plot visualizing the interaction between entrenchment, memory development, and L1 for each L2 in the intrinsic task of the gender agreement experiment.

Table 3.10

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 French outcomes in the intrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.41	0.01	54.85*	0.000	0.022	0.000	0.022
Middle L2 Learner	-0.01	0.01	-0.98	-	-	-	-
Late L2 Learner	-0.02	0.01	-1.66	-	-	-	-
Connection Growth	-0.16	0.01	-12.43*	-	-	-	-
Unit Growth	-0.17	0.01	-18.78*	-	-	-	-
Unit Replacement	-0.07	0.01	-7.55*	-	-	-	-
L1 Russian	0.00	0.01	-0.38	-	-	-	-
L1 Spanish	0.08	0.01	8.86*	-	-	-	-
Middle L2 Learner x Connection Growth	-0.01	0.02	-0.28	-	-	-	-
Late L2 Learner x Connection Growth	-0.01	0.03	-0.18	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.01	-0.15	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.01	0.02	-	-	-	-
Middle L2 Learner x Unit Replacement	-0.01	0.01	-0.44	-	-	-	-
Late L2 Learner x Unit Replacement	0.01	0.01	0.41	-	-	-	-
Middle L2 Learner x L1 Russian	0.00	0.01	0.07	-	-	-	-
Late L2 Learner x L1 Russian	0.00	0.02	-0.28	-	-	-	-
Middle L2 Learner x L1 Spanish	0.01	0.01	0.39	-	-	-	-
Late L2 Learner x L1 Spanish	0.01	0.01	0.40	-	-	-	-
Connection Growth x L1 Russian	0.01	0.02	0.75	-	-	-	-
Unit Growth x L1 Russian	0.01	0.01	0.91	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.01	-0.07	-	-	-	-
Connection Growth x L1 Spanish	-0.08	0.02	-4.19*	-	-	-	-
Unit Growth x L1 Spanish	-0.07	0.01	-5.00*	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.01	0.20	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.01	0.03	0.19	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.01	0.03	0.30	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.00	0.02	0.00	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.02	0.16	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.00	0.02	0.12	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.02	0.02	-0.76	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.00	0.03	0.04	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.00	0.04	-0.03	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	-0.01	0.02	-0.34	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	-0.01	0.02	-0.28	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.00	0.02	-0.09	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	-0.03	0.02	-1.51	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

In the L2 Spanish model (see Table 3.12 and Figure 3.8), there was no noticeable difference in L2 Spanish outcomes across the different L1 in the No Growth condition for early L2 learners. Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 Spanish outcomes. Connection

Table 3.11

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 Russian outcomes in the intrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.42	0.01	40.32*	0.001	0.038	0.001	0.034
Middle L2 Learner	-0.02	0.01	-1.97	-	-	-	-
Late L2 Learner	-0.03	0.01	-2.72*	-	-	-	-
Connection Growth	-0.24	0.01	-20.14*	-	-	-	-
Unit Growth	-0.24	0.01	-20.34*	-	-	-	-
Unit Replacement	-0.09	0.01	-7.35*	-	-	-	-
L1 French	-0.04	0.01	-2.83*	-	-	-	-
L1 Spanish	0.09	0.01	6.55*	-	-	-	-
Middle L2 Learner x Connection Growth	0.01	0.02	0.40	-	-	-	-
Late L2 Learner x Connection Growth	0.01	0.02	0.33	-	-	-	-
Middle L2 Learner x Unit Growth	0.01	0.02	0.46	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.02	0.27	-	-	-	-
Middle L2 Learner x Unit Replacement	-0.01	0.02	-0.48	-	-	-	-
Late L2 Learner x Unit Replacement	-0.01	0.02	-0.31	-	-	-	-
Middle L2 Learner x L1 French	0.01	0.02	0.23	-	-	-	-
Late L2 Learner x L1 French	0.04	0.02	1.69	-	-	-	-
Middle L2 Learner x L1 Spanish	-0.01	0.02	-0.37	-	-	-	-
Late L2 Learner x L1 Spanish	-0.03	0.02	-1.20	-	-	-	-
Connection Growth x L1 French	0.04	0.02	1.67	-	-	-	-
Unit Growth x L1 French	0.03	0.02	1.58	-	-	-	-
Unit Replacement x L1 French	0.01	0.02	0.54	-	-	-	-
Connection Growth x L1 Spanish	-0.09	0.03	-3.60*	-	-	-	-
Unit Growth x L1 Spanish	-0.09	0.02	-4.56*	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.02	-0.72	-	-	-	-
Middle L2 Learner x Connection Growth x L1 French	-0.01	0.03	-0.18	-	-	-	-
Late L2 Learner x Connection Growth x L1 French	-0.04	0.03	-1.16	-	-	-	-
Middle L2 Learner x Unit Growth x L1 French	-0.01	0.03	-0.18	-	-	-	-
Late L2 Learner x Unit Growth x L1 French	-0.04	0.03	-1.13	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 French	0.02	0.03	0.55	-	-	-	-
Late L2 Learner x Unit Replacement x L1 French	0.00	0.03	0.03	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.01	0.04	0.20	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.03	0.04	0.62	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	0.01	0.03	0.31	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.03	0.03	0.86	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.01	0.03	0.47	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.02	0.03	0.67	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Growth, Unit Growth, and Unit Replacement negatively impacted L2 Spanish outcomes in each L1 (a main effect of memory development), but the magnitude of the impact was much greater in the Connection Growth and Unit Growth conditions.

3.2.2.2 Analysis of the Extrinsic Evaluation Task

In the L2 French model (see Table 3.13 and Figure 3.9), L1 Spanish led to better L2 French outcomes than L1 English or L1 Russian (a main effect of L1). Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 French outcomes. Connection Growth, Unit Growth, and Unit Replacement negatively impacted L2 French outcomes in each L1 (a main effect of memory development).

In the L2 Spanish model (see Table 3.14 and Figure 3.9), there was no noticeable difference in L2 Spanish outcomes across the different L1 in the No Growth condition. Compared to early L2 learners, starting an L2 at higher entrenchment levels did not impact L2 Spanish outcomes. Connection Growth, Unit Growth, and Unit Replacement negatively impacted L2 Spanish outcomes in each L1 (a main effect of memory development).

The probability of producing false negatives increased due to the multiple comparisons across the factors of interest. However, the mixed effects results were not adjusted with a Bonferroni correction because the data points were not independent, and many of the patterns were consistent.

3.3 Discussion

The experiment presented above investigated the impact a specific L1 can have on learning to predict phonemes in a sequence of L2 phonemes when developmental factors like linguistic entrenchment and memory development vary. Previous work

by Monner et al. (2013) using similar methods found support for the entrenchment and less-is-more hypotheses. This experiment extends that research along several dimensions. First, two additional languages, English and Russian, were included. Second, more realistic and complex linguistic data was used. Third, a held-out test set was used to evaluate model performance.

Table 3.12

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 Spanish outcomes in the intrinsic task of the gender agreement experiment.

				Random Effects			
	Fixed Effects			*By models*		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.42	0.01	41.93*	0.001	0.031	0.000	0.020
Middle L2 Learner	-0.03	0.01	-2.40*	-	-	-	-
Late L2 Learner	-0.04	0.01	-2.62*	-	-	-	-
Connection Growth	-0.16	0.01	-11.09*	-	-	-	-
Unit Growth	-0.18	0.01	-14.01*	-	-	-	-
Unit Replacement	-0.08	0.01	-5.71*	-	-	-	-
L1 French	-0.01	0.01	-0.52	-	-	-	-
L1 Russian	-0.02	0.02	-1.19	-	-	-	-
Middle L2 Learner x Connection Growth	0.02	0.02	0.89	-	-	-	-
Late L2 Learner x Connection Growth	0.01	0.02	0.43	-	-	-	-
Middle L2 Learner x Unit Growth	0.01	0.02	0.78	-	-	-	-
Late L2 Learner x Unit Growth	0.01	0.02	0.57	-	-	-	-
Middle L2 Learner x Unit Replacement	0.02	0.02	0.94	-	-	-	-
Late L2 Learner x Unit Replacement	0.01	0.02	0.60	-	-	-	-
Middle L2 Learner x L1 French	0.03	0.02	1.55	-	-	-	-
Late L2 Learner x L1 French	0.05	0.02	2.17*	-	-	-	-
Middle L2 Learner x L1 Russian	0.01	0.02	0.50	-	-	-	-
Late L2 Learner x L1 Russian	0.01	0.02	0.27	-	-	-	-
Connection Growth x L1 French	0.03	0.03	0.99	-	-	-	-
Unit Growth x L1 French	0.04	0.02	2.25	-	-	-	-
Unit Replacement x L1 French	0.04	0.02	1.99	-	-	-	-
Connection Growth x L1 Russian	0.02	0.02	0.97	-	-	-	-
Unit Growth x L1 Russian	0.04	0.02	1.75	-	-	-	-
Unit Replacement x L1 Russian	0.02	0.02	0.98	-	-	-	-
Middle L2 Learner x Connection Growth x L1 French	0.00	0.02	0.13	-	-	-	-
Late L2 Learner x Connection Growth x L1 French	0.00	0.02	0.33	-	-	-	-
Middle L2 Learner x Unit Growth x L1 French	-0.02	0.03	-0.85	-	-	-	-
Late L2 Learner x Unit Growth x L1 French	-0.04	0.03	-1.37	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 French	-0.02	0.03	-0.70	-	-	-	-
Late L2 Learner x Unit Replacement x L1 French	-0.04	0.03	-1.30	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	-0.01	0.04	-0.20	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.00	0.04	0.13	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	-0.01	0.04	-0.38	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.04	-0.03	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	-0.04	0.04	-1.09	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.02	0.04	-0.44	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 3.13

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 French outcomes in the extrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.47	0.01	45.92*	0.001	0.010	0.001	0.013
Middle L2 Learner	-0.01	0.01	-0.97	-	-	-	-
Late L2 Learner	-0.02	0.01	-1.11	-	-	-	-
Connection Growth	-0.20	0.02	-11.08*	-	-	-	-
Unit Growth	-0.20	0.01	-15.90*	-	-	-	-
Unit Replacement	-0.10	0.01	-7.77*	-	-	-	-
L1 Russian	0.03	0.01	1.97	-	-	-	-
L1 Spanish	0.06	0.01	4.47*	-	-	-	-
Middle L2 Learner x Connection Growth	0.00	0.03	-0.14	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.04	-0.12	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.02	-0.09	-	-	-	-
Late L2 Learner x Unit Growth	-0.01	0.02	-0.48	-	-	-	-
Middle L2 Learner x Unit Replacement	-0.01	0.02	-0.42	-	-	-	-
Late L2 Learner x Unit Replacement	0.01	0.02	0.54	-	-	-	-
Middle L2 Learner x L1 Russian	0.00	0.02	0.08	-	-	-	-
Late L2 Learner x L1 Russian	0.00	0.02	0.07	-	-	-	-
Middle L2 Learner x L1 Spanish	0.01	0.02	0.33	-	-	-	-
Late L2 Learner x L1 Spanish	-0.01	0.02	-0.42	-	-	-	-
Connection Growth x L1 Russian	-0.01	0.02	-0.24	-	-	-	-
Unit Growth x L1 Russian	-0.01	0.02	-0.47	-	-	-	-
Unit Replacement x L1 Russian	-0.02	0.02	-0.90	-	-	-	-
Connection Growth x L1 Spanish	-0.03	0.03	-1.01	-	-	-	-
Unit Growth x L1 Spanish	-0.03	0.02	-1.41	-	-	-	-
Unit Replacement x L1 Spanish	0.01	0.02	0.41	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.00	0.04	0.07	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.00	0.04	-0.03	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.00	0.03	-0.03	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.03	0.12	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.00	0.03	0.16	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.03	0.03	-1.05	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.00	0.04	-0.10	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.01	0.05	0.13	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	-0.01	0.03	-0.33	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.01	0.03	0.43	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.00	0.03	0.05	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	-0.03	0.03	-1.08	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

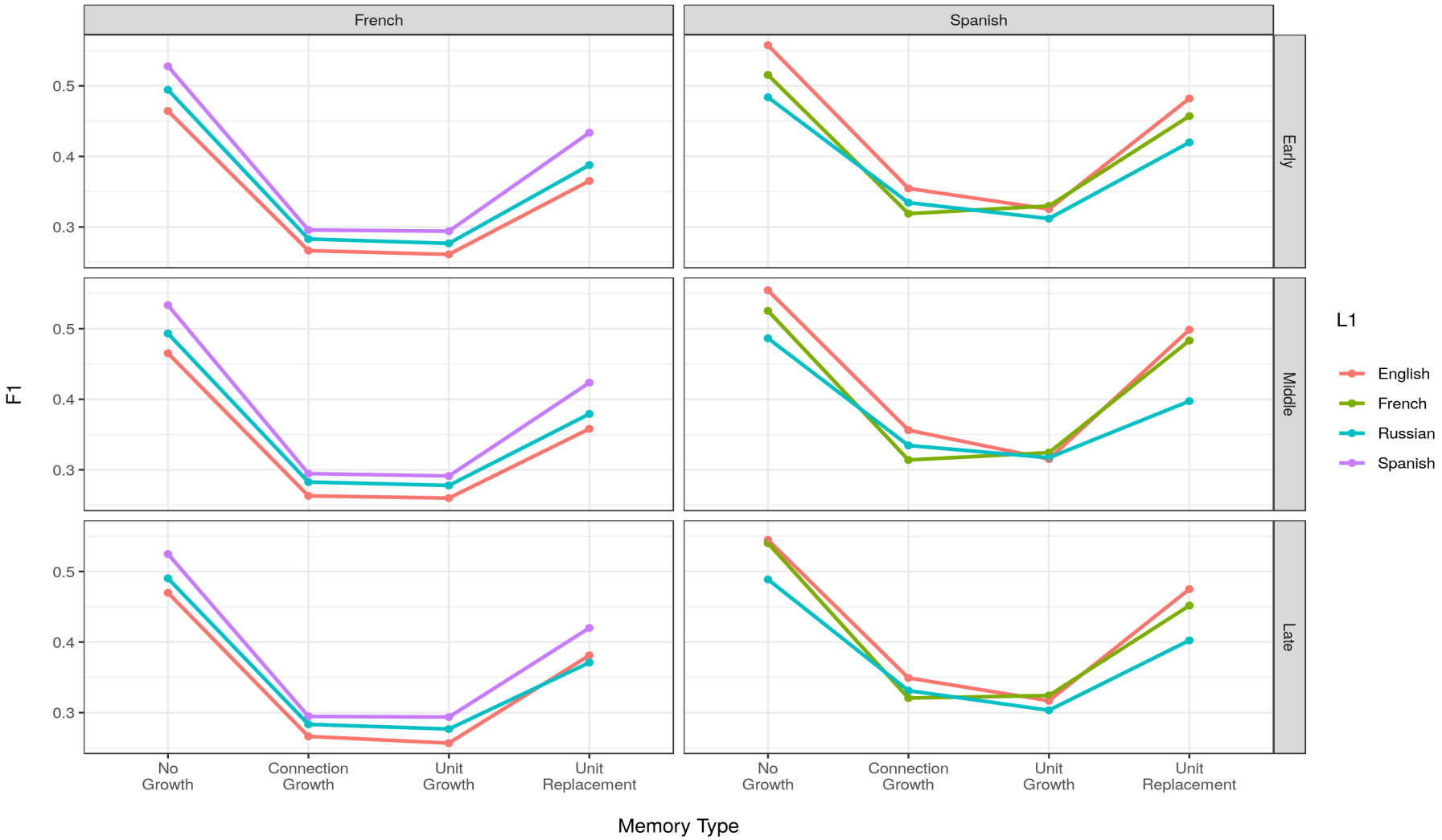

Figure 3.9. Plot visualizing the interaction between entrenchment, memory development, and L1 for each L2 in the extrinsic task of the gender agreement experiment.

An analysis of monolingual performance for each memory development condition indicated baseline performance differences between the languages. The learning task in this experiment did not reach peak performance in the No Growth and Unit Replacement conditions in the monolingual models. The task in this study was considerably harder than the gender assignment task. For the intrinsic task, English and Spanish performed the best, followed by French and Russian. In the extrinsic task, Spanish performed better than French. On both tasks, Connection Growth and Unit Growth led to much worse learning outcomes compared to either the No Growth or Unit Replacement conditions. Unit Replacement also consistently led to worse performance relative to No Growth.

Performance in the L2 models followed very similar patterns. L1 Spanish consistently led to better outcomes on L2 English, L2 French, and L2 Russian in the No Growth and early L2 learner conditions. In each L2 model, Connection Growth and Unit Growth led to a large decrease in L2 outcomes related to the No Growth condition. Unit Replacement also tended to perform worse than the No Growth condition, but just slightly.

Starting small consistently led to poorer L1 and L2 learning outcomes. This result may be due to the importance of fresh resources during the early stages of learning. Both the Connection Growth and Unit Growth conditions started with very small networks in terms of the number of parameters to be trained. If there are not a sufficient number of parameters available initially to learn the patterns found in the training data, the models may experience learning difficulties. Unlike the models trained in the previous experiment, entrenchment did not consistently

affect L2 outcomes. This finding diverges from previous research by Monner et al. (2013).

Table 3.14

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 Spanish outcomes in the extrinsic task of the gender agreement experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.56	0.01	56.94	0.001	0.030	0.001	0.023
Middle L2 Learner	-0.01	0.01	-1.10	-	-	-	-
Late L2 Learner	-0.03	0.01	-2.37	-	-	-	-
Connection Growth	-0.20	0.01	-14.81	-	-	-	-
Unit Growth	-0.23	0.01	-18.33	-	-	-	-
Unit Replacement	-0.08	0.01	-5.80	-	-	-	-
L1 French	-0.05	0.01	-3.95	-	-	-	-
L1 Russian	-0.08	0.02	-4.93	-	-	-	-
Middle L2 Learner x Connection Growth	0.00	0.02	0.20	-	-	-	-
Late L2 Learner x Connection Growth	0.01	0.02	0.34	-	-	-	-
Middle L2 Learner x Unit Growth	-0.01	0.02	-0.39	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.02	0.22	-	-	-	-
Middle L2 Learner x Unit Replacement	0.02	0.02	0.98	-	-	-	-
Late L2 Learner x Unit Replacement	0.01	0.02	0.28	-	-	-	-
Middle L2 Learner x L1 French	0.01	0.02	0.75	-	-	-	-
Late L2 Learner x L1 French	0.03	0.02	1.53	-	-	-	-
Middle L2 Learner x L1 Russian	0.01	0.02	0.60	-	-	-	-
Late L2 Learner x L1 Russian	0.01	0.02	0.50	-	-	-	-
Connection Growth x L1 French	0.02	0.03	0.58	-	-	-	-
Unit Growth x L1 French	0.05	0.02	2.89	-	-	-	-
Unit Replacement x L1 French	0.01	0.02	0.56	-	-	-	-
Connection Growth x L1 Russian	0.06	0.02	2.45	-	-	-	-
Unit Growth x L1 Russian	0.07	0.02	3.06	-	-	-	-
Unit Replacement x L1 Russian	0.02	0.02	0.92	-	-	-	-
Middle L2 Learner x Connection Growth x L1 French	0.02	0.02	0.34	-	-	-	-
Late L2 Learner x Connection Growth x L1 French	0.02	0.02	0.31	-	-	-	-
Middle L2 Learner x Unit Growth x L1 French	-0.01	0.03	-0.22	-	-	-	-
Late L2 Learner x Unit Growth x L1 French	-0.03	0.03	-0.97	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 French	-0.01	0.03	-0.32	-	-	-	-
Late L2 Learner x Unit Replacement x L1 French	-0.01	0.03	-0.49	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	-0.02	0.04	-0.60	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	-0.02	0.04	-0.47	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	-0.01	0.03	-0.31	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.00	0.03	-0.13	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	-0.05	0.03	-1.45	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.02	0.03	-0.51	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Chapter 4: Word Boundary Detection

Parsing the speech stream into meaningful lexical items is a fundamental task for all language learners, especially early learners. Speech is typically the only form of linguistic input to which young children can attend. This problem is relatively difficult since the acoustic signal does not regularly indicate which sounds correspond to the beginning or end of a word. The experiment presented here manipulates the factors of linguistic entrenchment and memory development while training a model to identify the word boundaries in a sequence of phonemes.

This task was chosen for several reasons. One, word boundary identification is a fundamental skill early learners must master before developing more complex linguistic capabilities. Two, this skill continues to develop over time as learners are exposed to more varied input. Third, the identification of word boundaries is often the subject of study in research addressing the impact statistical regularities have on child and adult learners (for a review, see Newport, 2016). Researchers have utilized computational methods to explore the nature of word segmentation (Elman, 1993; Goldwater, Griffiths, & Johnson, 2009). Finally, the binary nature of word boundary identification makes it possible to control the number of possible outputs produced by a neural network. A diverse set of languages can be modeled while maintaining

the same output space. This was not possible in the gender assignment and gender agreement experiments. In the gender assignment experiment, the output space was dependent upon the number of gender classes in the language, while the output space in the gender agreement experiment was dependent upon the phonemes present in a particular language. In these experiments, not all languages had an input mapping for each unit in the output layer.

Linguistic entrenchment is expected to have an overall negative effect on L2 learning outcomes. Starting the learning process with a smaller working memory capacity is not expected to mute the negative effects of entrenchment, nor is it expected to lead to better L2 learning outcomes (for supporting results, see Rohde & Plaut, 1999; Monner et al., 2013).

4.1 Method

Recurrent neural networks using an LSTM architecture (Hochreiter & Schmidhuber, 1997; Gers et al., 1999, 2002) were trained to predict word boundaries in a sequence of phonemes. At every step of the sequence, the model predicts whether there is a word boundary or not (see Table 2.3 for details regarding the architecture and hyperparameters of the network and Figure 3.1 for a graphical representation of the task). Just like the other two experiments, the model learned solely from encoded phonological features of the phonemes in the phrases; therefore, only patterns associated with phonological features are potentially being learned by the models. For each trial during training, a sequence of vectors consisting of 22 articulatory fea-

tures that represent a unique phoneme was presented as input to the network. For example, at each time-step in the sequence of phonemes for /mikasaessukasa/ (my house is your house), the model would produce an output vector predicting whether the current phoneme is a word boundary (see Figure 4.1). Four different languages were paired (English, French, Russian, Spanish), creating a total of 12 unique language pairs. Linguistic entrenchment and memory development are manipulated using the same methods described in the previous two experiments.

The statistical regularities of the phonemes in relation to word boundaries are being learned. In previous studies, the word boundaries were implicitly determined using transitional probabilities Elman (1990). Here, the same type of input is being used to learn boundaries through an explicit error signal.

/m/	[01100010100100100000000]	→	[0]
/i/	[11010000100000010000010]	→	[1]
/k/	[00100000000000001010000]	→	[0]
/a/	[11010000100000000110010]	→	[0]
/s/	[00110000000110000000000]	→	[0]
/a/	[11010000100000000110010]	→	[1]
/e/	[11010000100000000000010]	→	[0]
/s/	[00110000000110000000000]	→	[1]
/s/	[00110000000110000000000]	→	[0]
/u/	[11010000100000011011010]	→	[1]
/k/	[00100000000000001010000]	→	[0]
/a/	[11010000100000000110010]	→	[0]
/s/	[00110000000110000000000]	→	[0]
/a/	[11010000100000000110010]	→	[1]

Figure 4.1. Articulatory feature vectors of input phonemes mapped to word boundaries for /mikasaessukasa/.

4.1.1 Phrases

The word boundary identification experiment used the training and test sets from the grammatical gender agreement experiment. The only difference between the training and test sets of the two experiments was the target vector, which in this experiment was binary (e.g., boundary versus non-boundary). A target value was associated with every phoneme in the sequence. If the phoneme corresponded to the index location of the end of a word (as determined by the text of the corpus), the target value was assigned to the boundary class (1); otherwise, the phoneme was assigned to the non-boundary class (0). The class distribution between boundary and non-boundary was skewed in favor of non-boundary. As in the gender agreement experiment, 100,000 aligned phrases were used for the training sets, and 2,000 aligned phrases were used for the test sets (see Table 4.1 for statistics on the training and test sets distribution of target classes).

Table 4.1

Means (standard deviations) of boundary and non-boundary occurences in the training and test sets for each language in the word boundary experiment.

	Train			*Test*		
	N	*Boundaries*	*Non − boundaries*	*N*	*Boundaries*	*Non − boundaries*
English	100,000	8.85 (4.07)	47.0 (16.4)	2,000	8.81 (4.13)	46.8 (16.4)
French	100,000	10.0 (4.34)	53.7 (19.0)	2,000	9.99 (4.3)	53.6 (19.0)
Russian	100,000	7.66 (3.44)	56.3 (19.6)	2,000	7.61 (3.32)	56.1 (19.3)
Spanish	100,000	10.1 (4.25)	57.8 (19.5)	2,000	9.98 (4.22)	57.7 (19.5)

4.1.2 Cross-Linguistic Similiarity across Phonemic Sequences

The same similarity scores from the gender agreement experiment were used in this study. Again, these scores were derived by creating embedding vectors of the sequences in the training sets with the SGT algorithm (Ranjan et al., 2016). The Euclidean distance (see Equation 3.1) between each language pair of each aligned phrase was calculated. The mean Euclidean distance between the language pairs was the similarity score, which was subsequently transformed into a z-score. The six similarity scores are the following: English-French = -0.66, English-Russian = -1.38, English-Spanish = 0.22, French-Russian = -0.61, French-Spanish = 1.57, Russian-Spanish = 0.87.

4.1.3 Training Procedure

Recurrent neural networks (see Table 2.3) representing language learning agents were trained to predict the presence of a word boundary from a context of previously seen phonemes. The selection of hyperparameters for this experiment was similar to those used in the gender assignment and gender agreement experiments. The learning rate was chosen so that the learning task could reach peak performance relatively early during the training phase. It was important that the models reached peak performance on the learning task during the training phase. For each step during training, mini-batches of ten phrases were selected as input to the model. Each phrase was represented as a sequence of vectors consisting of 22 articulatory features representing unique phonemes. The ten phrases were selected randomly

from the list of weighted phrases in the training set. The network produced ten output vectors predicting either boundary or non-boundary for the given phoneme (n=2) at every step of each phrase. The mean cross-entropy loss between the ten output vectors and the ten true target vectors was used during the backpropagation step of the training phase. This loss value was used to calculate the gradient for each parameter in the model. The SGD learning method was followed throughout the training process. The open-source machine learning library PyTorch (version 1.1.0) was used to train all models (Paszke et al., 2017).

Identical to the previous two experiments, model training was divided into two phases, a monolingual (i.e., entrenchment) phase and a bilingual (i.e., post-entrenchment) phase. The monolingual phase varied by the entrenchment level, while the bilingual phase always had a constant length of 2,000,000 nouns. The duration of the bilingual phase was kept constant to ensure each model would have enough time to reach ceiling performance in both languages regardless of the length of the entrenchment phase. Therefore, the total training time varied across the entrenchment levels.

In order to avoid catastrophic forgetting in the models due to a distributional shift in input characteristics, an equal probability of exposure to either L1 or L2 linguistic input was ensured during the bilingual phase. The interleaving approach addresses catastrophic forgetting in the network; however, it does not eliminate competition and inference between the two languages.

The monolingual phase represented the entrenchment factor, which manipulates the quantity of L1 input prior to the introduction of L2 input. There were

6 levels of L1 entrenchment (t): 0, 200,000, 400,000, 600,000, 800,000, 1,000,000. The first level of L1 entrenchment, t=0, represented a balanced (or native) bilingual network since both L1 and L2 were present from the beginning of training. The last level of L1 entrenchment, t=1,000,000, represented a late L2 learner.

The monolingual and bilingual training phases were followed under four different memory development conditions: 1) No Growth, 2) Unit Growth, 3) Unit Replacement, 4) Connection Growth (see Figure 2.5). The goal of the memory development condition was to manipulate the architecture of the model in order to mirror memory development in humans. A detailed description of this condition is provided in the methods section for the gender assignment experiment.

Ten networks were trained in each cell of the design matrix. Six levels of L1 entrenchment under four different memory development conditions yielded 240 models for each language pair. A total of 12 language pairs (L1 English - L2 French, L1 English - L2 Russian, L1 English - L2 Spanish, L1 French - L2 English, L1 French - L2 Russian, L1 French - L2 Spanish, L1 Russian - L2 English, L1 Russian - L2 French, L1 Russian - L2 Spanish, L1 Spanish - L2 English, L1 Spanish - L2 French, L1 Spanish - L2 Russian) led to 2,880 models across all conditions.

4.1.4 Evaluation Criteria

Each model was evaluated on the L1 and L2 test sets every 100,000 nouns. All formal analyses and reporting of descriptive statistics used the F1 score (Equation 2.13).

4.2 Results

4.2.1 Monolingual Baseline

Ten monolingual models were trained for each of the 12 language pairs under each of the four memory development conditions in order to establish a monolingual baseline for each language and a corresponding naive L2 baseline for each language. This led to 480 monolingual models trained on 1,000,000 phrases. Model performance was evaluated on the test set every 100,000 phrases, so each model had ten data points in which model performance on the test set was measured. The mean and standard deviation of the final F1 score for each language and memory condition is reported in Table 4.2. Figure 4.2 shows the baseline performance for each language in the L1 and the L2 across all four memory conditions.

Separate mixed effects models were fit to the F1 score associated with the L1 (monolingual) and the L2 (naive L2 baseline) in order to identify differences between languages and memory development conditions. The analysis of the naive L2 baseline was conducted separately for each L2. All models were fit to the data using the lme4 package (version 1.1-21; Bates et al., 2015) within the R statistical platform (version 3.5.2; R Core Team, 2015).

In the monolingual models, contrast coding was used for the language (here, the L1) and memory development factors. The reference for the monolingual models corresponds to French in the No Growth condition. All fixed effects across the other factors are relative to the reference level. The performance of each model was

evaluated on the L1 and the naive L2 every 100,000 nouns. These data points were collected at different moments in the training phase. Therefore, each model had ten evaluation runs for each language (repeated measurements of the dependent variable). The random effects structure for all models included random intercepts for the grouping factors of model and evaluation run (the repeated measure grouping factor).

In the monolingual mixed effects model (see Table 4.3 and Figure 4.3), there was a main effect of language. Relative to French in the No Growth condition, Russian and Spanish monolingual models consistently performed worse while English models did not differ from French models. In relation to the No Growth condition, the Connection Growth, Unit Growth, and Unit Replacement conditions led to poorer outcomes across all languages (a main effect of memory condition). For all languages, the negative impact on L1 outcomes was even more pronounced in the Connection Growth and Unit Growth conditions. English in the Unit Growth condition did not impact learning outcomes as much as the other languages (an interaction between memory development and language).

Table 4.2

Mean (standard deviation) final F1 score in the word boundary experiment for monolingual and naive L2 baselines of each language in each memory development condition. The monolingual (L1) values were gathered after 1,000,000 nouns were used to train the model on the L1. The naive L2 baseline values represent performance on an L2 when no training occurred for that L2.

	No Growth	*Connection Growth*	*Unit Growth*	*Unit Replacement*
English				
*L*1	0.88 (0.02)	0.58 (0.02)	0.59 (0.06)	0.83 (0.05)
*L*2	0.53 (0.02)	0.49 (0.03)	0.49 (0.04)	0.52 (0.02)
French				
*L*1	0.89 (0.01)	0.66 (0.01)	0.65 (0.05)	0.83 (0.04)
*L*2	0.54 (0.04)	0.50 (0.03)	0.50 (0.05)	0.55 (0.04)
Russian				
*L*1	0.83 (0.02)	0.58 (0.02)	0.55 (0.07)	0.77 (0.05)
*L*2	0.50 (0.02)	0.40 (0.06)	0.42 (0.05)	0.49 (0.03)
Spanish				
*L*1	0.85 (0.02)	0.56 (0.02)	0.54 (0.04)	0.78 (0.05)
*L*2	0.54 (0.04)	0.50 (0.06)	0.48 (0.06)	0.54 (0.04)

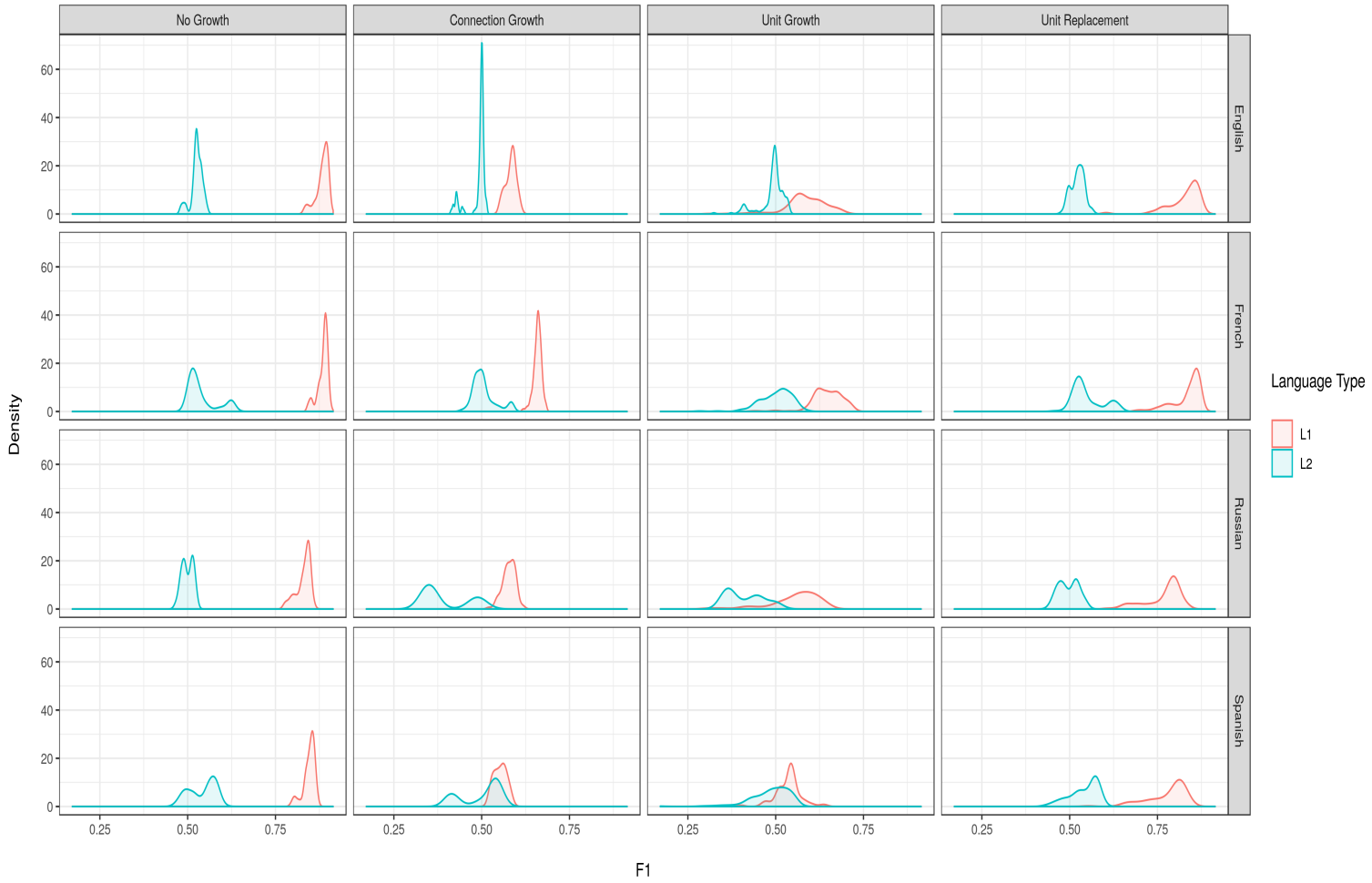

Figure 4.2. Density plots visualizing the distribution of F1 scores of monolingual and naive L2 models in the word boundary experiment.

Table 4.3

Mixed effects model examining the effect of memory development and language on learning outcomes in monolingual models in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.89	0.01	154.20*	0.000	0.006	0.000	0.016
Connection Growth	-0.23	0.00	-61.24*	-	-	-	-
Unit Growth	-0.24	0.00	-67.24*	-	-	-	-
Unit Replacement	-0.05	0.00	-14.25*	-	-	-	-
L1 English	-0.00	0.00	-0.39	-	-	-	-
L1 Russian	-0.06	0.00	-12.79*	-	-	-	-
L1 Spanish	-0.04	0.00	-7.80*	-	-	-	-
Connection Growth x L1 English	-0.07	0.01	-11.71*	-	-	-	-
Unit Growth x L1 English	-0.06	0.01	-9.39*	-	-	-	-
Unit Replacement x L1 English	-0.00	0.01	-0.70	-	-	-	-
Connection Growth x L1 Russian	-0.03	0.01	-3.90*	-	-	-	-
Unit Growth x L1 Russian	-0.04	0.01	-5.81*	-	-	-	-
Unit Replacement x L1 Russian	-0.01	0.01	-1.74	-	-	-	-
Connection Growth x L1 Spanish	-0.06	0.01	-6.72*	-	-	-	-
Unit Growth x L1 Spanish	-0.07	0.01	-9.90*	-	-	-	-
Unit Replacement x L1 Spanish	-0.02	0.01	-2.24*	-	-	-	-

Note: Mixed effects model formula: F1 $\sim$ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Starting with a smaller working memory capacity (Unit Growth) did not lead to better outcomes than starting with a fully developed network, like in the No Growth or Unit Replacement conditions. Connection Growth and Unit Growth conditions had very similar impacts on model performance in both tasks. This result suggests that long-term memory capacity is important for successful learning outcomes and that a lack of long-term resources during the early stages of development leads to poorer learning outcomes. The performance on this task was quite high in the No Groth condition for each of the four languages. English and French consistently performed better than Russian and Spanish across all memory development conditions.

A naive baseline in which the L2 was not explicitly trained was established

for each L2. The purpose of examining the performance of networks that have not been trained on an L2 was to determine how a particular L1 may influence the learning outcomes of a particular L2. The L2 of each monolingual model was evaluated every 100,000 nouns. This resulted in ten evaluation runs for the L2 in each model (repeated measurements of the dependent variable). For each evaluation ask, a mixed effects model was fit to each L2. The random effects structure included random intercepts for the grouping factors of model and evaluation run.

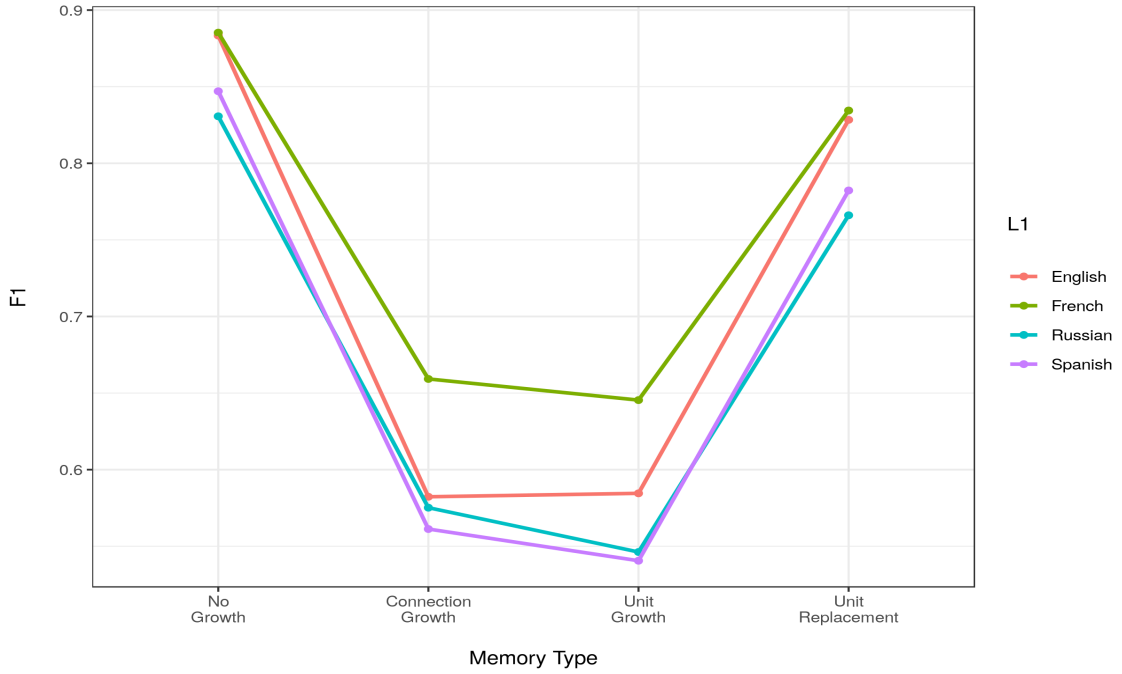

Figure 4.3. Plot visualizing the interaction between memory development condition and L1 in the monolingual models of the word boundary experiment.

In the naive L2 English model (see Table 4.4 and Figure 4.4), L1 French and L1 Russian led to better naive L2 English outcomes than L1 Spanish in the No Growth condition (a main effect of L1). Compared to the No Growth condition with an L1 as French, both Connection Growth and Unit Growth led to worse naive

L2 English outcomes (a main effect of memory development). This pattern also occurred when the L1 was Russian but did not occur when the L1 was Spanish (an interaction between memory development and L1).

In the naive L2 French model (see Table 4.4 and Figure 4.4), L1 Spanish led to better naive L2 French outcomes than L1 English and L1 Russian in the No Growth condition (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth led to worse naive L2 French outcomes in L1 English, L1 Russian, and L1 Spanish (a main effect of memory development).

In the naive L2 Russian model (see Table 4.4 and Figure 4.4), L1 French and L1 Spanish led to better naive L2 Russian outcomes than L1 English in the No Growth condition (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth led to worse naive L2 Russian outcomes in each L1 (a main effect of memory development).

In the naive L2 Spanish model (see Table 4.4 and Figure 4.4), L1 French and L1 Russian led to better naive L2 Spanish outcomes than L1 English in the No Growth condition (a main effect of L1). Compared to the No Growth condition, both Connection Growth and Unit Growth led to worse naive L2 Spanish outcomes in each L1 (a main effect of memory development).

Across all of the native L2 models (see Table 4.4), Connection Growth and Unit Growth negatively impacted L2 outcomes. The L1 did impact naive L2 outcomes, a result which suggests that certain L1s may provide positive transfer on the task of word boundary identification.

4.2.2 Bilingual Models

Two thousand eight hundred eighty models were trained to predict the presence of a word boundary given a sequence of phonemes represented as a sequence of articulatory feature vectors; ten training runs per experimental condition (six entrenchment levels, four memory conditions, and 12 language pairs). The performance of each model was evaluated on the test set (2,000 phrases) of each language every 100,000 phrases and at the end of the bilingual training phase. The F1 score was used to perform statistical analyses and report descriptive statistics. The mean and standard deviation of the final F1 score for each language, entrenchment level, and memory condition is reported in Table 4.5.

As in the other two experiments, the factor of entrenchment was simplified in the analysis by reducing the factor to three levels (Early, Middle, and Late L2 learners). To understand how the experimental factors influenced L2 learning outcomes, mixed effects models were fit to the F1 score associated with the L2. Mixed effects models were fit to the fixed factors of entrenchment, memory development, and L1 in each L2.

A mixed effect model was fit to the fixed factors of entrenchment, memory development, and L1 in each of the four languages. Contrast coding was used for the entrenchment, memory development, and L1 factors. The L2 of each model was evaluated every 100,000 nouns during the bilingual phase. Therefore, each model had 20 evaluation runs for the L2 (repeated measurements of the dependent variable). Models were fit to the data using the *lme4* package (version 1.1-21; Bates

et al., 2015) within the R statistical platform (version 3.5.2; R Core Team, 2015). The random effects structure included random intercepts for the grouping factors of model and evaluation run. This random effects structure was used for all mixed effects models.

The pattern of results of the mixed effects models for each L2 were identical (see Tables 4.6, 4.7, 4.8, 4.9 and Figure 4.5). The L1 did not influence L2 learning outcomes in the No Growth and Unit Replacement conditions. In the L2 French, L2 Russian, and L2 Spanish models, L1 English consistently had a larger negative impact in the Connection Growth and Unit Growth conditions. The Unit Replacement condition had a consistently negative impact on L2 outcomes. The Connection Growth and Unit Growth conditions also negatively impacted L2 outcomes relative to the No Growth condition, and the magnitude of this impact was large.

The probability of producing false negatives increased due to the multiple comparisons across the factors of interest. However, the mixed effects results were not adjusted with a Bonferroni correction because the data points were not independent and many of the patterns were consistent.

Table 4.4

Mixed effects models examining the effect of memory development and L1 on learning outcomes in naive L2 models in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
English							
(Intercept)	0.71	0.01	89.83*	0.000	0.000	0.000	0.000
Connection Growth	-0.13	0.01	-10.83*	-	-	-	-
Unit Growth	-0.13	0.01	-11.45*	-	-	-	-
Unit Replacement	-0.03	0.01	-2.38*	-	-	-	-
L1 Russian	-0.02	0.02	-1.11	-	-	-	-
L1 Spanish	-0.04	0.02	-1.89	-	-	-	-
Connection Growth x L1 Russian	-0.05	0.03	-1.58	-	-	-	-
Unit Growth x L1 Russian	-0.07	0.03	-2.08*	-	-	-	-
Unit Replacement x L1 Russian	-0.02	0.03	-0.51	-	-	-	-
Connection Growth x L1 Spanish	-0.02	0.04	-0.59	-	-	-	-
Unit Growth x L1 Spanish	-0.02	0.03	-0.82	-	-	-	-
Unit Replacement x L1 Spanish	-0.01	0.03	-0.33	-	-	-	-
French							
(Intercept)	0.70	0.01	60.36*	0.000	0.000	0.000	0.000
Connection Growth	-0.16	0.02	-7.43*	-	-	-	-
Unit Growth	-0.16	0.02	-9.16*	-	-	-	-
Unit Replacement	-0.03	0.02	-1.97	-	-	-	-
L1 Russian	-0.02	0.01	-1.63	-	-	-	-
L1 Spanish	0.03	0.02	1.63	-	-	-	-
Connection Growth x L1 Russian	0.02	0.03	0.71	-	-	-	-
Unit Growth x L1 Russian	0.00	0.02	0.04	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.02	0.28	-	-	-	-
Connection Growth x L1 Spanish	-0.00	0.04	-0.04	-	-	-	-
Unit Growth x L1 Spanish	-0.03	0.03	-1.18	-	-	-	-
Unit Replacement x L1 Spanish	-0.00	0.03	-0.02	-	-	-	-
Russian							
(Intercept)	0.68	0.01	46.26*	0.000	0.000	0.000	0.000
Connection Growth	-0.22	0.02	-10.30*	-	-	-	-
Unit Growth	-0.20	0.02	-9.45*	-	-	-	-
Unit Replacement	-0.03	0.02	-1.41	-	-	-	-
L1 French	0.01	0.03	0.51	-	-	-	-
L1 Spanish	-0.00	0.02	-0.04	-	-	-	-
Connection Growth x L1 French	0.09	0.04	2.30*	-	-	-	-
Unit Growth x L1 French	0.06	0.04	1.57	-	-	-	-
Unit Replacement x L1 French	0.01	0.04	0.32	-	-	-	-
Connection Growth x L1 Spanish	0.05	0.04	1.18	-	-	-	-
Unit Growth x L1 Spanish	0.01	0.03	0.18	-	-	-	-
Unit Replacement x L1 Spanish	-0.00	0.03	-0.08	-	-	-	-
Spanish							
(Intercept)	0.69	0.01	54.11*	0.000	0.000	0.000	0.000
Connection Growth	-0.19	0.02	-9.85*	-	-	-	-
Unit Growth	-0.17	0.02	-9.50*	-	-	-	-
Unit Replacement	-0.03	0.02	-1.45	-	-	-	-
L1 French	0.04	0.02	2.78*	-	-	-	-
L1 Russian	-0.01	0.02	-0.56	-	-	-	-
Connection Growth x L1 French	0.06	0.02	2.53*	-	-	-	-
Unit Growth x L1 French	0.02	0.02	0.82	-	-	-	-
Unit Replacement x L1 French	0.00	0.02	0.14	-	-	-	-
Connection Growth x L1 Russian	0.04	0.04	1.19	-	-	-	-
Unit Growth x L1 Russian	-0.01	0.03	-0.24	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.03	0.10	-	-	-	-

Note: Mixed effects model formula: F1 ~ (memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

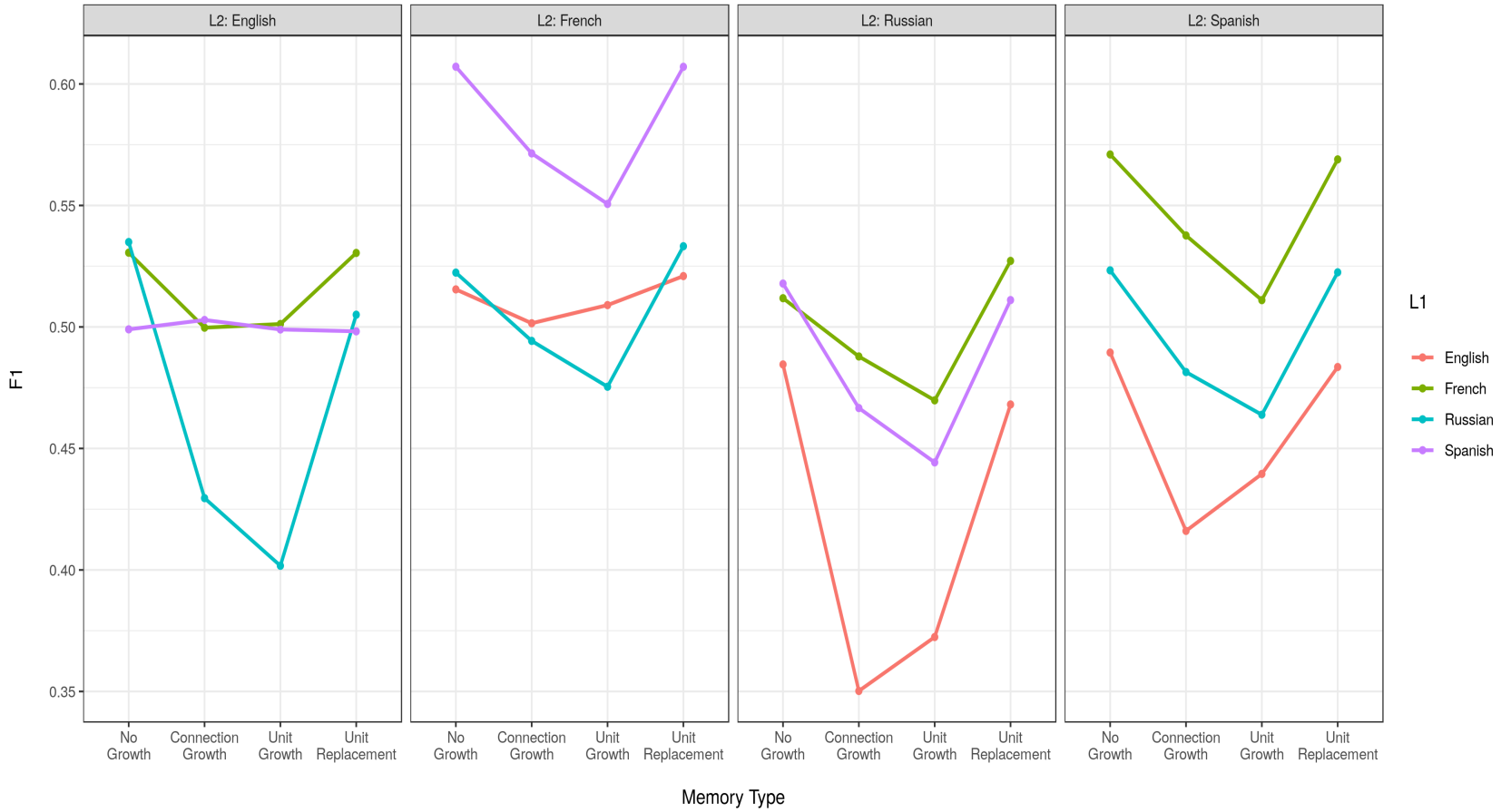

Figure 4.4. Plot visualizing the interaction between memory development condition and L1 in the naive L2 models of the word boundary experiment.

Table 4.5

Means (standard deviations) of final F1 scores for L2 outcomes of each language across entrenchment and memory development conditions in the word boundary experiment.

		No Growth		*Connection Growth*		*Unit Growth*		*Unit Replacement*	
		*L*1	*L*2	*L*1	*L*2	*L*1	*L*2	*L*1	*L*2
English									
	Early	0.87 (0.02)	0.86 (0.03)	0.56 (0.02)	0.57 (0.02)	0.56 (0.06)	0.57 (0.05)	0.81 (0.04)	0.81 (0.04)
	Middle	0.88 (0.01)	0.87 (0.02)	0.57 (0.02)	0.57 (0.02)	0.58 (0.05)	0.56 (0.06)	0.83 (0.02)	0.82 (0.03)
	Late	0.88 (0.01)	0.87 (0.02)	0.57 (0.02)	0.57 (0.01)	0.57 (0.06)	0.57 (0.06)	0.84 (0.02)	0.82 (0.03)
French									
	Early	0.87 (0.02)	0.87 (0.02)	0.62 (0.04)	0.64 (0.04)	0.61 (0.06)	0.62 (0.06)	0.82 (0.04)	0.82 (0.03)
	Middle	0.88 (0.01)	0.87 (0.02)	0.63 (0.04)	0.64 (0.04)	0.63 (0.06)	0.62 (0.06)	0.85 (0.02)	0.82 (0.02)
	Late	0.89 (0.01)	0.87 (0.02)	0.62 (0.04)	0.64 (0.04)	0.62 (0.06)	0.62 (0.07)	0.85 (0.01)	0.82 (0.02)
Russian									
	Early	0.82 (0.02)	0.82 (0.02)	0.53 (0.03)	0.51 (0.05)	0.53 (0.07)	0.49 (0.07)	0.77 (0.04)	0.75 (0.05)
	Middle	0.83 (0.01)	0.82 (0.02)	0.54 (0.03)	0.51 (0.04)	0.53 (0.07)	0.51 (0.06)	0.78 (0.02)	0.77 (0.03)
	Late	0.83 (0.01)	0.82 (0.02)	0.54 (0.03)	0.50 (0.04)	0.52 (0.07)	0.48 (0.08)	0.78 (0.02)	0.76 (0.03)
Spanish									
	Early	0.83 (0.02)	0.83 (0.02)	0.54 (0.03)	0.53 (0.03)	0.51 (0.07)	0.52 (0.07)	0.78 (0.04)	0.78 (0.04)
	Middle	0.85 (0.01)	0.83 (0.02)	0.54 (0.04)	0.52 (0.04)	0.52 (0.08)	0.52 (0.07)	0.80 (0.02)	0.78 (0.03)
	Late	0.85 (0.01)	0.83 (0.02)	0.52 (0.03)	0.53 (0.04)	0.51 (0.08)	0.53 (0.07)	0.80 (0.02)	0.78 (0.02)

Table 4.6

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 English outcomes in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.87	0.00	184.12*	0.000	0.012	0.000	0.018
Middle L2 Learner	0.00	0.01	-0.16	-	-	-	-
Late L2 Learner	-0.01	0.00	-2.40*	-	-	-	-
Connection Growth	-0.29	0.00	-58.63*	-	-	-	-
Unit Growth	-0.29	0.00	-59.28*	-	-	-	-
Unit Replacement	-0.06	0.00	-11.78*	-	-	-	-
L1 Russian	0.00	0.01	-0.36	-	-	-	-
L1 Spanish	0.00	0.01	-0.49	-	-	-	-
Middle L2 Learner x Connection Growth	-0.01	0.01	-0.88	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.01	-0.51	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.01	-0.34	-	-	-	-
Late L2 Learner x Unit Growth	0.00	0.01	0.13	-	-	-	-
Middle L2 Learner x Unit Replacement	0.00	0.01	0.46	-	-	-	-
Late L2 Learner x Unit Replacement	0.01	0.01	0.93	-	-	-	-
Middle L2 Learner x L1 Russian	-0.01	0.01	-0.47	-	-	-	-
Late L2 Learner x L1 Russian	0.00	0.01	-0.26	-	-	-	-
Middle L2 Learner x L1 Spanish	0.00	0.01	-0.43	-	-	-	-
Late L2 Learner x L1 Spanish	0.00	0.01	-0.14	-	-	-	-
Connection Growth x L1 Russian	-0.02	0.01	-1.69	-	-	-	-
Unit Growth x L1 Russian	0.02	0.01	2.02*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.01	0.87	-	-	-	-
Connection Growth x L1 Spanish	0.00	0.01	-0.15	-	-	-	-
Unit Growth x L1 Spanish	0.00	0.01	-0.15	-	-	-	-
Unit Replacement x L1 Spanish	0.00	0.01	-0.06	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.01	0.02	0.49	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.01	0.02	0.59	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	-0.03	0.02	-1.96	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	-0.06	0.02	-3.85*	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	-0.01	0.02	-0.82	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.01	0.02	-0.57	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.00	0.02	0.18	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.00	0.02	-0.08	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	-0.01	0.02	-0.52	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.00	0.01	-0.05	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.01	0.01	0.45	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.01	0.01	0.42	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 4.7

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 French outcomes in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.87	0.00	197.62*	0.000	0.008	0.000	0.015
Middle L2 Learner	0.00	0.00	-1.00	-	-	-	-
Late L2 Learner	-0.01	0.00	-1.57	-	-	-	-
Connection Growth	-0.29	0.01	-51.76*	-	-	-	-
Unit Growth	-0.29	0.00	-59.47*	-	-	-	-
Unit Replacement	-0.05	0.00	-11.16*	-	-	-	-
L1 Russian	0.00	0.00	0.48	-	-	-	-
L1 Spanish	0.00	0.01	0.88	-	-	-	-
Middle L2 Learner x Connection Growth	0.00	0.01	-0.25	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.01	-0.33	-	-	-	-
Middle L2 Learner x Unit Growth	0.00	0.01	-0.66	-	-	-	-
Late L2 Learner x Unit Growth	-0.01	0.01	-1.76	-	-	-	-
Middle L2 Learner x Unit Replacement	0.00	0.01	-0.02	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.01	-0.04	-	-	-	-
Middle L2 Learner x L1 Russian	-0.01	0.01	-0.86	-	-	-	-
Late L2 Learner x L1 Russian	-0.01	0.01	-1.57	-	-	-	-
Middle L2 Learner x L1 Spanish	0.00	0.01	-0.03	-	-	-	-
Late L2 Learner x L1 Spanish	-0.01	0.01	-0.71	-	-	-	-
Connection Growth x L1 Russian	0.08	0.01	12.06*	-	-	-	-
Unit Growth x L1 Russian	0.05	0.01	8.69*	-	-	-	-
Unit Replacement x L1 Russian	0.01	0.01	1.77	-	-	-	-
Connection Growth x L1 Spanish	0.06	0.01	5.77*	-	-	-	-
Unit Growth x L1 Spanish	0.05	0.01	6.04*	-	-	-	-
Unit Replacement x L1 Spanish	0.01	0.01	1.56	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.01	0.01	0.67	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.01	0.01	0.98	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	0.01	0.01	1.30	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	0.02	0.01	2.10*	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.00	0.01	-0.35	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	-0.01	0.01	-0.60	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.00	0.01	-0.05	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.01	0.02	0.62	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	0.01	0.01	0.89	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	0.03	0.01	2.21*	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	0.00	0.01	-0.03	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	0.00	0.01	0.38	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 4.8

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 Russian outcomes in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.82	0.01	132.70*	0.000	0.017	0.000	0.018
Middle L2 Learner	-0.01	0.01	-0.96	-	-	-	-
Late L2 Learner	-0.01	0.01	-1.71	-	-	-	-
Connection Growth	-0.34	0.01	-43.87*	-	-	-	-
Unit Growth	-0.35	0.01	-45.91*	-	-	-	-
Unit Replacement	-0.07	0.01	-8.94*	-	-	-	-
L1 French	0.01	0.01	0.70	-	-	-	-
L1 Spanish	-0.01	0.01	-0.75	-	-	-	-
Middle L2 Learner x Connection Growth	0.00	0.01	0.25	-	-	-	-
Late L2 Learner x Connection Growth	0.00	0.01	0.41	-	-	-	-
Middle L2 Learner x Unit Growth	0.03	0.01	2.92*	-	-	-	-
Late L2 Learner x Unit Growth	0.01	0.01	0.94	-	-	-	-
Middle L2 Learner x Unit Replacement	0.01	0.01	1.08	-	-	-	-
Late L2 Learner x Unit Replacement	0.02	0.01	1.70	-	-	-	-
Middle L2 Learner x L1 French	0.00	0.01	-0.31	-	-	-	-
Late L2 Learner x L1 French	-0.01	0.01	-0.52	-	-	-	-
Middle L2 Learner x L1 Spanish	0.00	0.01	0.13	-	-	-	-
Late L2 Learner x L1 Spanish	0.00	0.01	0.27	-	-	-	-
Connection Growth x L1 French	0.07	0.01	5.33*	-	-	-	-
Unit Growth x L1 French	0.04	0.01	2.85*	-	-	-	-
Unit Replacement x L1 French	0.01	0.01	0.76	-	-	-	-
Connection Growth x L1 Spanish	0.06	0.01	4.31*	-	-	-	-
Unit Growth x L1 Spanish	0.05	0.01	4.27*	-	-	-	-
Unit Replacement x L1 Spanish	0.01	0.01	0.66	-	-	-	-
Middle L2 Learner x Connection Growth x L1 French	-0.01	0.02	-0.39	-	-	-	-
Late L2 Learner x Connection Growth x L1 French	-0.01	0.02	-0.26	-	-	-	-
Middle L2 Learner x Unit Growth x L1 French	-0.03	0.02	-1.66	-	-	-	-
Late L2 Learner x Unit Growth x L1 French	0.01	0.02	0.35	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 French	0.00	0.02	-0.10	-	-	-	-
Late L2 Learner x Unit Replacement x L1 French	-0.01	0.02	-0.42	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Spanish	0.00	0.02	0.16	-	-	-	-
Late L2 Learner x Connection Growth x L1 Spanish	0.00	0.02	-0.11	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Spanish	-0.03	0.02	-1.57	-	-	-	-
Late L2 Learner x Unit Growth x L1 Spanish	-0.03	0.02	-1.82	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Spanish	-0.01	0.02	-0.37	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Spanish	-0.02	0.02	-1.25	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

Table 4.9

Mixed effects model examining the effect of entrenchment, memory development, and L1 on L2 Spanish outcomes in the word boundary experiment.

	Fixed Effects			*Random Effects*			
				By models		*By runs*	
Parameters	*Estimate*	*SE*	*t*	*Var*	*SD*	*Var*	*SD*
(Intercept)	0.84	0.01	123.89	0.000	0.015	0.000	0.022
Middle L2 Learner	-0.01	0.01	-1.00	-	-	-	-
Late L2 Learner	-0.01	0.01	-1.81	-	-	-	-
Connection Growth	-0.33	0.01	-42.10	-	-	-	-
Unit Growth	-0.33	0.01	-43.07	-	-	-	-
Unit Replacement	-0.06	0.01	-7.69	-	-	-	-
L1 French	0.01	0.01	1.38	-	-	-	-
L1 Russian	0.00	0.01	-0.35	-	-	-	-
Middle L2 Learner x Connection Growth	-0.01	0.01	-0.53	-	-	-	-
Late L2 Learner x Connection Growth	-0.01	0.01	-0.68	-	-	-	-
Middle L2 Learner x Unit Growth	0.01	0.01	1.22	-	-	-	-
Late L2 Learner x Unit Growth	0.01	0.01	1.16	-	-	-	-
Middle L2 Learner x Unit Replacement	-0.01	0.01	-0.73	-	-	-	-
Late L2 Learner x Unit Replacement	0.00	0.01	0.20	-	-	-	-
Middle L2 Learner x L1 French	0.00	0.01	-0.13	-	-	-	-
Late L2 Learner x L1 French	0.00	0.01	-0.43	-	-	-	-
Middle L2 Learner x L1 Russian	-0.01	0.01	-0.69	-	-	-	-
Late L2 Learner x L1 Russian	-0.01	0.01	-0.74	-	-	-	-
Connection Growth x L1 French	0.02	0.01	1.91	-	-	-	-
Unit Growth x L1 French	0.02	0.01	1.87	-	-	-	-
Unit Replacement x L1 French	0.01	0.01	0.53	-	-	-	-
Connection Growth x L1 Russian	0.07	0.01	5.04	-	-	-	-
Unit Growth x L1 Russian	0.04	0.01	3.50	-	-	-	-
Unit Replacement x L1 Russian	0.00	0.01	0.37	-	-	-	-
Middle L2 Learner x Connection Growth x L1 French	0.01	0.02	0.89	-	-	-	-
Late L2 Learner x Connection Growth x L1 French	0.02	0.01	1.50	-	-	-	-
Middle L2 Learner x Unit Growth x L1 French	-0.01	0.02	-0.85	-	-	-	-
Late L2 Learner x Unit Growth x L1 French	-0.01	0.01	-0.57	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 French	0.02	0.02	1.13	-	-	-	-
Late L2 Learner x Unit Replacement x L1 French	0.01	0.01	0.36	-	-	-	-
Middle L2 Learner x Connection Growth x L1 Russian	0.02	0.02	1.12	-	-	-	-
Late L2 Learner x Connection Growth x L1 Russian	0.03	0.02	1.30	-	-	-	-
Middle L2 Learner x Unit Growth x L1 Russian	-0.01	0.02	-0.73	-	-	-	-
Late L2 Learner x Unit Growth x L1 Russian	-0.01	0.02	-0.51	-	-	-	-
Middle L2 Learner x Unit Replacement x L1 Russian	0.02	0.02	0.98	-	-	-	-
Late L2 Learner x Unit Replacement x L1 Russian	0.01	0.02	0.40	-	-	-	-

Note: Mixed effects model formula: F1 ~ (entrenchment * memory * L1) + (1 | model) + (1 | eval)
*$|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007).

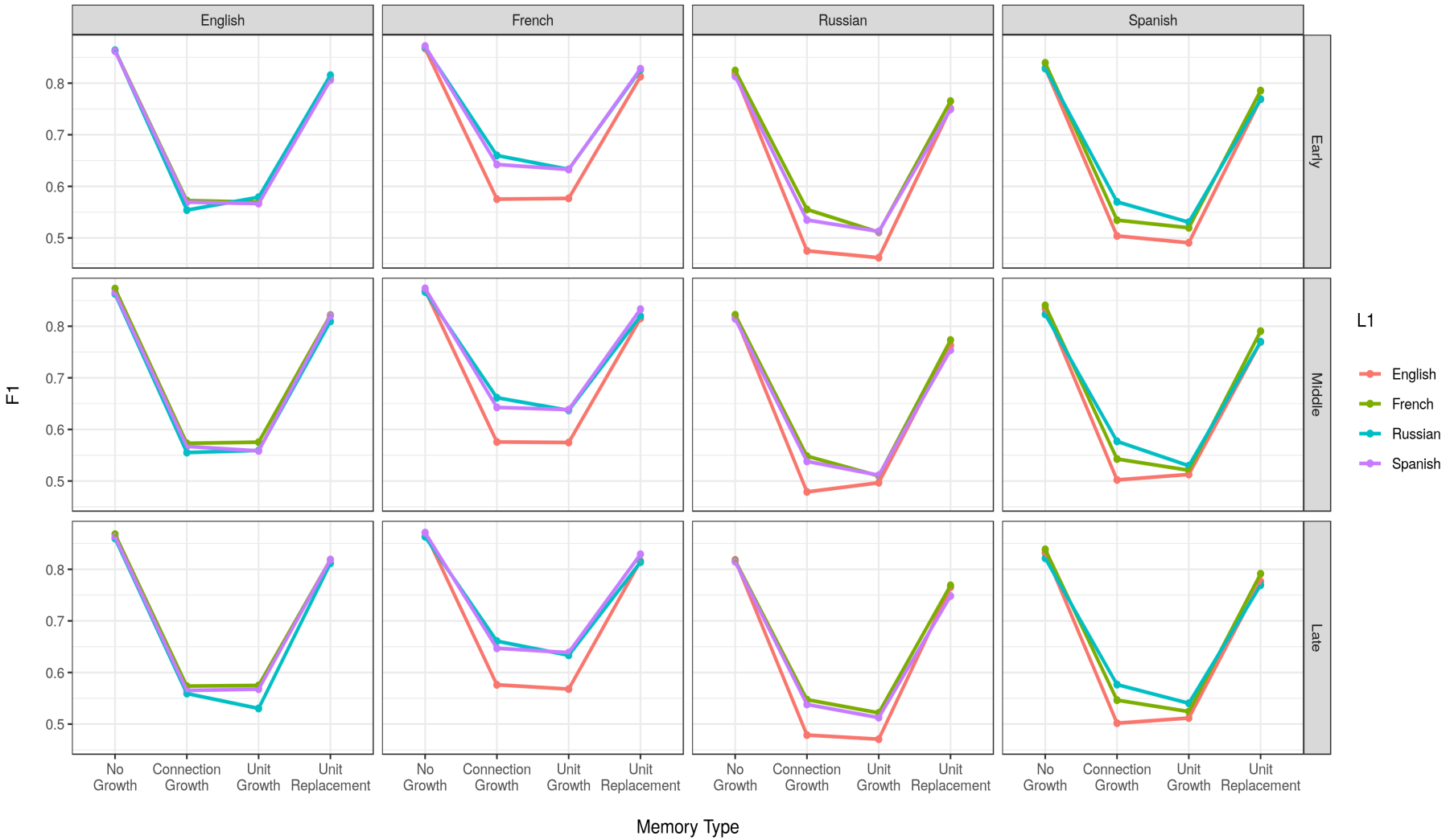

Figure 4.5. Plot visualizing the interaction between entrenchment, memory development, and L1 for each L2 in the word boundary experiment.

4.3 Discussion

The experiment presented above investigated the impact a specific L1 can have on learning to identify word boundaries in a sequence of phonemes when developmental factors like linguistic entrenchment and memory development vary. Previous work by Monner et al. (2013) found support for the entrenchment and less-is-more hypotheses on two different tasks: gender assignment and gender agreement.

This word boundary identification task was chosen for several reasons. One, word boundary identification is a fundamental skill early learners must master before developing more complex linguistic capabilities. Two, it is a skill that develops over time. Third, the identification of word boundaries is often the subject of study in research addressing the impact statistical regularities have on child and adult learners (for a review, see Newport, 2016). Finally, the binary nature of word boundary identification makes it possible to control the number of possible outputs produced by a neural network. A diverse set of languages can be modeled while maintaining the same output space.

An analysis of monolingual performance for each memory development condition indicated baseline performance differences between the languages. English and Spanish performed the best, followed by French and Russian. Connection Growth and Unit Growth led to much worse learning outcomes compared to either the No Growth or Unit Replacement conditions. Unit Replacement also consistently led to worse performance relative to No Growth.

L2 outcomes were not affected by the L1 in the No Growth and Unit Replace-

ment conditions. However, L1 did affect L2 learning outcomes for several language pairs in the Connection Growth and Unit Growth conditions. Again, the Connection Growth and Unit Growth conditions consistently led to poor L2 learning outcomes, regardless of the L2 or L1. Entrenchment did not lead to more negative L2 learning outcomes.

The overall pattern of the results is similar to those obtained in the gender agreement experiment. This experiment did not find consistent entrenchment effects and the influence of the Connection Growth and Unit Growth conditions consistently led to poorer learning outcomes. The results of all three experiments are discussed in the next chapter.

Chapter 5: Conclusion

The dynamic nature of language learning makes it difficult to isolate and control many of the variables believed to influence ultimate L2 learning outcomes (DeKeyser & Larson-Hall, 2005; DeKeyser, 2012). Most research in the field of L2 learning has either been correlational or lacked sufficient statistical power for the claims they make (Hartshorne et al., 2018). Many of the maturational and experiential factors of interest correlate with each other, and to make matters more complicated, they often develop concurrently. These developmental constraints make it impossible to test all possible hypotheses using controlled behavioral experiments. Due to these reasons, three computational experiments were conducted to understand better how different language combinations are affected by manipulations of linguistic entrenchment and memory development.

Previous research by Monner et al. (2013) found support for the hypotheses that linguistic entrenchment negatively affects L2 learning outcomes and that fewer memory resources during the early stages of learning helps alleviate the negative effects of entrenchment. However, their results varied by the learning task and the specific characteristics of the linguistic systems under study. The work presented above exposed the factors of linguistic entrenchment and memory development to a

more complex dataset using a larger variety of languages. Unlike previous research, all models were evaluated on a held-out test set.

The results of the three experiments above are inconsistent with regards to linguistic entrenchment. An entrenchment effect was consistently present in the gender assignment experiment, but the magnitude of the effect was modest. The gender agreement and word boundary experiments only sporadically found a weak entrenchment effect. This result is surprising considering previous research has consistently claimed that their results support the linguistic entrenchment hypothesis. The gender assignment experiment in Monner et al. (2013) found modest entrenchment effect in the L1 Spanish - L2 French language pair and no entrenchment effects in the L1 French - L2 Spanish pair. The results from the three experiments suggest that the effect of entrenchment on second language learning may be overstated in the literature and highly dependent upon the modeling task and architecture. If the entrenchment effect is highly dependent upon learning task and/or network architecture, it is important to know under what hyperparameter settings and which network architectures lead to effects of entrenchment.

Starting with fewer memory resources led to poorer outcomes across learning tasks and language combinations. Support for an advantage in starting small was only found in the gender assignment experiment. In the monolingual model, only L1 French performed better in the Connection Growth condition compared to the No Growth condition. In the L2 models, in the L1 German - L2 French and L1 German - L2 Spanish models, the Connection Growth condition led to better performance on the L2 relative to the No Growth condition. Starting with fewer memory resources in

the gender agreement and word boundary experiments always led to poorer L1 and L2 performance. The results found here are consistent with those found in the L1 French - L2 Spanish language pair of the gender assignment experiment in Monner et al. (2013) and the results obtained by Rohde and Plaut (1999). The former study used nearly identical methods to model memory development, whereas the latter study modeled memory development differently.

There are numerous possible explanations for these findings. The bilingual training phase (2,000,000) was long enough for peak performance to be reached relatively easily regardless of the level of entrenchment. In all three experiments, L2 performance reached its peak only a couple hundred thousand nouns or phrases after the commencement of the bilingual phase. Therefore, there was ample time for the network to learn the new L2 patterns while also maintaining the original L1 patterns. This would indicate that the network extracted all of the statistical regularities of the phonological input well before the bilingual training phase ended. There was evidence of catastrophic interference, especially in the gender assignment experiment. However, this result was likely entirely due to the imbalance between the number of classes in the L1 and L2. If the L2 had fewer gender classes than the L1, performance on the L2 suffered considerably. It would be interesting to know how long this effect would last if the L1 was not maintained after the introduction of the L2.

The specific details of how the networks were trained, the hyperparameters used throughout training, and architectural differences between learning tasks may have led to outcomes different from those obtained in previous studies. The LSTM

architectures and implementations used in the present studies are slightly different from those used in Monner et al. (2013), and very different from those used in other studies (Elman, 1993; Rohde & Plaut, 1999; Li et al., 2004, 2007; Zhao & Li, 2010; Li & Zhao, 2013). In Monner et al. (2013), they used a generalized LSTM algorithm (Monner & Reggia, 2012). Here, the LSTM class and the associated computational graph from PyTorch (Paszke et al., 2017), a machine learning framework for Python, was used. The LSTM implemented here was not the generalized version, which Monner et al. (2013) argue is more biologically plausible and converges quicker than the standard LSTM architecture.

The way in which training was undertaken may also have led to different outcomes. Here, mini-batches of 10 input sequences were used to train the models efficiently. The training routine followed by Monner et al. (2013) appears to have been sequential (one input sequence at a time). The exact details of the hyperparameters and manipulation of the memory development conditions in Monner et al. (2013) are not provided in their published paper, and their code was not available for inspection. Due to the lack of specific detail regarding previous studies, it is difficult to identify which of the many differences led to the divergence in results.

Evaluating model performance on a held-out test set may lead to different learning outcomes. Monner et al. (2013) evaluated their models with the data used to train the model, whereas the experiments here were all evaluated with a held-out test set. There was no a priori reason to believe that the effects of entrenchment would be affected by this change, however.

As stated above, it is possible that the negative influence of linguistic en-

trenchment on L2 learning is overstated in the literature. This is a hypothesis that necessitates further research to identify when entrenchment occurs in these types of models and when it does not. Starting with fewer memory resources consistently led to poorer L1 and L2 performance, regardless of the level of entrenchment. The only exception to this finding was found in the gender assignment experiment. In the gender agreement and word boundary experiments, the negative effect of starting small was very large. The gender agreement was the most difficult task in the study while the word boundary task was the easiest (best overall performance across languages). Overall, the results of these experiments suggest that starting with limited resources and gradually expanding them does not necessarily confer a benefit when learning an L1 or L2.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*(4), 481–509.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bayliss, D. M., Jarrold, C., Baddeley, A. D., Gunn, D. M., & Leigh, E. (2005). Mapping the developmental constraints on working memory span performance. *Developmental Psychology*, *41*(4), 579.

Bley-Vroman, R. (1988). The fundamental character of foreign language learning. In W. Rutherford & M. Sharwood Smith (Eds.), *Grammar and second language teaching: A book of readings* (pp. 19–30). Rowley, MA: Newbury House.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the less-is-more hypothesis. *Language Learning*, *69*, 13–41.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A., 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.

Comrie, B. (1999). Grammatical gender systems: A linguist's assessment. *Journal of Psycholinguistic Research*, *28*(5), 457–466.

Conway, A. R., Jarrold, C. E., Kane, M. J., Miyake, A., & Towse, J. N. (2007). Variation in working memory: An introduction. In A. R. Conway, C. E. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 3–17). New York, NY: Oxford University Press.

Corbett, G. G. (1982). Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics*, 197–232.

Corbett, G. G. (2006). Gender, grammatical. *Encyclopedia of Language & Linguistics*, 749–756.

Corbett, G. G. (2007). Gender and noun classes. In T. Shopen (Ed.), *Language typology and syntactic description: Iii: Grammatical categories and the lexicon* (pp. 241–279). Cambridge, UK: Cambridge University Press.

Cowan, N., & Alloway, T. (1997). The development of working memory. In N. Cowan & C. Hulme (Eds.), *The development of memory in childhood* (pp. 163–200). Psychology Press Ltd.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica: International Journal of Methodology and Experimental Psychology*, *32*(2), 133–143.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433.

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(3), 413–438.

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). Oxford, UK: Oxford Univerity Press.

DeKeyser, R. M. (2012). Age effects in second language learning. In S. Gass & A. Mackey (Eds.), (pp. 442–460). Routledge.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1103.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

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*(03), 379–399.

Gao, W., Lin, W., Chen, Y., Gerig, G., Smith, J., Jewells, V., & Gilmore, J. (2009). Temporal and spatial development of axonal maturation and myelination of white matter in the developing brain. *American Journal of Neuroradiology*, *30*(2), 290–296.

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*(11), 410–419.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. New York, NY: Cambridge.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, *3*(Aug), 115–143.

Gershkoff-Stowe, L., & Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, *34*(1), 37–71.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR*, *abs/1308.0850*. Retrieved from `http://arxiv.org/abs/1308.0850`

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610.

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

Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. (1999). Cambridge University Press.

Harley, B. (1979). French gender rules in the speech of English-dominant, French-dominant and monolingual French-speaking children. *Working Papers on Bilingualism*(19), 129–154.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277.

Hernandez, A. E., & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin*, *133*(4), 638.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holmes, V. M., & de la Bâtie, B. D. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics*, *20*(4), 479–506.

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, *24*(1), 131–161.

Jia, G., Aaronson, D., & Wu, Y. (2002). Long-term language attainment of bilingual immigrants: Predictive variables and language group differences. *Applied Psycholinguistics*, *23*(4), 599–621.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford, CA: Stanford University Press.

Lapkin, S., & Swain, M. (1977). The use of English and French cloze tests in a bilingual education program evaluation: Validity and error analysis. *Language Learning*, *27*(2), 279–310.

Lenneberg, E. (1967). *Biological foundations of language.* New York, NY: John Wiley & Sons.

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.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*(8), 1345–1362.

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in Psychology*, *4*, 828.

Li, P., Zhao, X., & Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*(4), 581–612.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861–883.

Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*(3), 251–285.

MacWhinney, B. (2005). Emergent fossilization. In Z. Han & T. Odlin (Eds.), *Studies of fossilization in second language acquisition* (pp. 134–156). Multilingual Matters.

MacWhinney, B. (2008). A unified model of language acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford University Press.

MacWhinney, B. (2016). Entrenchment in second-language learning. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 343–366). De Gruyter Mouton.

Marchman, V. A. (1993). Constraints on plasticity in a connectionist model of the english past tense. *Journal of Cognitive Neuroscience*, *5*(2), 215–234.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Monner, D., & Reggia, J. A. (2012). A generalized LSTM-like training algorithm for second-order recurrent neural networks. *Neural Networks*, *25*, 70–83.

Monner, D., Vatz, K., Morini, G., Hwang, S.-O., & DeKeyser, R. (2013). A neural network model of the effects of entrenchment and memory development on grammatical gender learning. *Bilingualism: Language and Cognition*, *16*(2), 246–265.

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.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. S. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *COLING* (pp. 3475–3484).

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences*, *10*(1), 147–172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11 - 28.

Newport, E. L. (2016). Statistical language learning: Computational, maturational, and linguistic constraints. *Language and Cognition*, *8*(3), 447–461.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS-W*.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ranjan, C., Ebrahimi, S., & Paynabar, K. (2016). Sequence graph transform (SGT): A feature extraction function for sequence data mining. *arXiv preprint arXiv:1608.03533*.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30–54.

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

Sabourin, L., & Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, *24*(3), 397–430.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Sá-Leite, A. R., Fraga, I., & Comesaña, M. (2019). Grammatical gender processing in bilinguals: An analytic review. *Psychonomic Bulletin & Review*, 1–26.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

Schatz, T., Feldman, N., Goldwater, S., Cao, X., & Dupoux, E. (2019). Early phonetic learning without phonetic categories – insights from machine learning.

Schepens, J., Van der Slik, F., & Van Hout, R. (2013). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In L. Borin & A. Saxena (Eds.), *Comparing approaches to measuring linguistic differences* (pp. 199–230). Mouton de Gruyter.

Schepens, J., Van der Slik, F., & Van Hout, R. (2014). Learning complex features: A morphological account of L2 learnability. *Language Dynamics and Change*, *3*(2), 218–244.

Scherag, A., Demuth, L., Rösler, F., Neville, H. J., & Röder, B. (2004). The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. *Cognition*, *93*(3), B97–B108.

Schmid, H. (2019). *RNNTagger*. Available from `https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/`. Munich, Germany.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523.

Seidenberg, M. S., & Zevin, J. D. (2006). Connectionist models in developmental cognitive neuroscience: Critical periods and the paradox of success. In Y. Munakata & M. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI* (pp. 315–347). Oxford University Press.

Sharoff, S. (2002). Meaning as use: Exploitation of aligned corpora for the contrastive study of lexical semantics. In *LREC.*

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213.

Steinmetz, D. (1986). Two principles and some rules for gender in German: Inanimate nouns. *Word*, *37*(3), 189–217.

Sun, R., Coward, L. A., & Zenzen, M. J. (2005). On levels of cognitive modeling. *Philosophical Psychology*, *18*(5), 613-637.

Surridge, M. E. (1993). Gender assignment in French: The hierarchy of rules and the chronology of acquisition. *International Review of Applied Linguistics in Language Teaching*, *31*(2), 77–96.

Surridge, M. E. (1995). *Le ou la?: The gender of French nouns* (Vol. 1). Multilingual Matters.

Teschner, R. V., & Russell, W. M. (1984). The gender patterns of Spanish nouns: An inverse dictionary-based analysis. *Hispanic Linguistics*, *1*(1), 115–132.

Thomas, M. S., & Johnson, M. H. (2008). New advances in understanding sensitive periods in brain development. *Current Directions in Psychological Science*, *17*(1), 1–5.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42.

Ullman, M. T. (2014). The declarative/procedural model: A neurobiologically motivated theory of first and second language. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 147–172). Routledge.

Uylings, H. B. (2006). Development of the human cortex and the concept of "critical" or "sensitive" periods. *Language Learning*, *56*, 59–90.

van Rossum, G. (1995). *Python tutorial.* Amsterdam, NL: Centrum voor Wiskunde en Informatica.

Vatz, K. L. (2009). *Grammatical gender representation and processing in advanced second language learners of French* (Unpublished doctoral dissertation).

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., ... Rey, A. (2018). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science*.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, *47*(1), 1–29.

Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism*, *13*(5), 505–524.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In *LREC*.