Bilingualism: Language and Cognition 16 (2), 2013, 246–265 © Cambridge University Press 2012 doi:10.1017/S1366728912000454

A neural network model of the effects of entrenchment and memory development on grammatical gender learning*

DEREK MONNER
Department of Computer Science, University of Maryland College Park
KAREN VATZ
Center for Advanced Study of Language, University of Maryland College Park
GIOVANNA MORINI
Department of Hearing and Speech Sciences, University of Maryland College Park
SO-ONE HWANG
Center for Research in Language, University of California San Diego
ROBERT DEKEYSER
Second Language Acquisition Program, University of Maryland College Park

(Received: August 17, 2011; final revision received: June 5, 2012; accepted: June 20, 2012; first published online 30 August 2012)

To investigate potential causes of L2 performance deficits that correlate with age of onset, we use a computational model to explore the individual contributions of L1 entrenchment and aspects of memory development. Since development and L1 entrenchment almost invariably coincide, studying them independently is seldom possible in humans. To avoid this confound, we study neural network models that learn to solve gender assignment and agreement tasks in Spanish and French. We model the learner as a collection of recurrent cell assemblies that subserve working memory and are facilitated by trainable long-term connections. Varying the time-course over which assemblies and connections are added allows us to compare small, growing, child-like networks to fixed-size adult-like ones. Networks undergo variable-length exposure to L1 before L2 onset to control the amount of L1 entrenchment. This model, by allowing us independent control of both variables, lends us a novel glimpse of all sides of their interaction and affords a rare test of the less-is-more hypothesis. Network comparisons suggest that final L2 proficiency declines as L2 onset delays increase relative to L1, implicating an L1 entrenchment effect. However, aspects of memory development during learning play a key role in mitigating these impairments, lending support to less-is-more as a contributor to sensitive periods.

Keywords: age effects, grammatical gender, entrenchment, memory development, recurrent neural networks

1 Introduction

For decades now, language researchers have been attempting to explain the observation that people who learn a second language (L2) later in life tend to have poorer ultimate attainment than those who learn the same language earlier in life; for an illustration of the pattern, see Figure 1a. Cross-linguistically, there is a clear downward trend in many, although not all, measures of language proficiency as age of acquisition increases (DeKeyser, 2012). This phenomenon has been referred to by many names, usually based on the author's thoughts on the phenomenon's likely cause. Since human maturational processes are widely implicated in first language (L1) acquisition, many suspect similar developmental processes to be largely responsible for these observed age effects on L2 acquisition, often referring to a "critical" or "sensitive period" for language learning. Others, who view the issue as a problem inherent in the process of learning, speak of cross-linguistic interference or entrenchment effects. Still others couch the problem in terms of individual differences of the language learners and quality and form of the L2 input. While there is support for all of these accounts of this phenomenon, it is generally difficult to study any of these potential causes in isolation.

In this study we use a neural network model to investigate the individual and compound effects that two of these potential causes of sensitive periods have on ultimate attainment of a learner's first and second languages. The first factor we will consider, entrenchment, can best be understood as previous knowledge that is difficult to change and can perhaps only be altered slowly, thus interfering with the rapid acquisition of newly available information. In this scenario, the longer

* Thanks to two anonymous reviewers for helpful comments. This work was supported in part by NSF IGERT award DGE-0801465.

Address for correspondence:
Derek Monner, A.V. Williams Bldg #3136, University of Maryland, College Park, MD 20742, USA
dmonner@cs.umd.edu

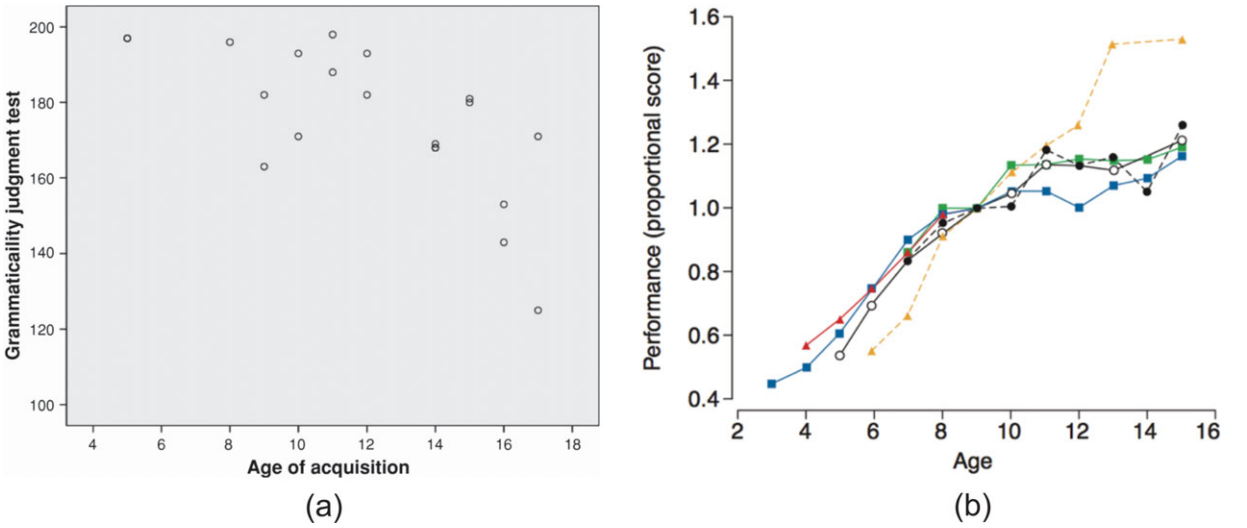

Figure 1. (Colour online) Trends of language-related skills with aging. (a) Scores of participants on an English grammaticality judgment test plotted against age of acquisition (reproduced from DeKeyser, Alfi-Shabtay & Ravid, 2010). (b) Several measures of working memory capacity plotted against the age of the participant (reproduced from Gathercole, 1999).

the learner is exposed to their native language before a second language is introduced, the more their L1 becomes entrenched, making the novel rules and patterns of an L2 more difficult to learn (Hernandez, Li & MacWhinney, 2005). The second factor we consider is the development of aspects of memory, specifically working memory capacity and long-term memory capacity, as implemented by the periodic addition of new units and connections, respectively, to our neural network model. Working memory development is particularly interesting in light of evidence, such as that shown in Figure 1b, that a period of rapid growth of working memory capacity coincides with a period of rapid deterioration of L2 learning ability.

Using only experimentation on human subjects, it is difficult to get a complete picture of the relative contributions of entrenchment and development. While there are exceptions, specifically in the sign language domain, language learning almost invariably starts very early in life, causing L1 acquisition and early L2 acquisition to coincide with many aspects of development. Thus, the contributions of these two factors to the observed differences in ultimate attainment between early and late L2 learners cannot be readily separated from each other. With a computational model, on the other hand, we can examine the interaction of our two chosen factors from all sides, describing the effects of each in isolation as well as their combined impact.

Of course, at present, a computer model cannot learn an entire natural language as human learners can. As such, we chose to model the linguistic sub-tasks of gender assignment and agreement. The factors guiding this choice of tasks included the fact that native and non-native speakers of a language tend to differ significantly, as well as the fact that ultimate attainment tends to vary with age of acquisition. Our model learns to perform gender assignment and gender agreement tasks from naturalistic training data based on word co-occurrence, without having any built-in knowledge of the existence or form of grammatical gender and without being given explicit instruction in the genders of particular words or phrases. Our goal with this model is to provide a better understanding of how the two potential factors we have chosen to study, entrenchment and memory development, contribute individually and in tandem to differences in ultimate language attainment.

Our experiments investigate two related but independent hypotheses. The first of these concerns the effects of language entrenchment: We expect that as the level of L1 entrenchment goes up, L2 learning ability goes down, at least up until some point of maximal entrenchment where the effect levels off. The second line of inquiry concerns the LESS-IS-MORE hypothesis (Newport, 1988, 1990) that states that a learner in the early stages of working memory development will find L2 learning easier than a learner with a fully developed working memory. Our simulations investigate these two hypotheses individually and in tandem, to a greater extent than is normally possible in empirical studies. Additionally, we are able to investigate more specific distinctions within the less-is-more hypothesis, discriminating the effects due to starting small from those due to addition of fresh memory resources.

The remainder of the paper is structured as follows. Section 2 reviews previous research on sensitive period phenomena and relationships to the acquisition of

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

grammatical gender. We also review hypotheses relating sensitive periods to working memory and to L1 entrenchment. Section 3 gives an overview first of neural networks in general and then of the specific neural network model studied herein, including all relevant variations. Section 4 describes separate experiments and results for the gender assignment and gender agreement tasks. Finally, in Section 5, we discuss the implications of our findings.

2 Background

2.1 Sensitive periods

Since Lenneberg (1967) first used the term CRITICAL PERIOD in the context of human language development, a considerable amount of evidence has accumulated that shows a marked decline in the ultimate outcome (not the speed) of language acquisition as age of onset varies from early childhood to late adolescence. This decline has been documented in numerous studies, for both L1 and L2 development, for both spoken and signed languages, and for phonology as well as morphology and syntax (for overviews, see DeKeyser, 2012; Hyltenstam & Abrahamsson, 2003).

Numerous questions remain, however, at least where the L2 is concerned. The most debated one is whether the age effects observed are truly maturational or due to confounds with other variables (e.g., DeKeyser & Larson-Hall, 2005; Long, 2005). Most commonly mentioned in the discussion of potential confounds are the extent of L1 entrenchment (e.g., MacWhinney, 2006), the quantity and quality of input and practice in L2 (e.g., Jia & Aaronson, 2003; Jia, Aaronson & Wu, 2002), the extent to which the learner is motivated to sound like a native speaker (e.g., Bley-Vroman, 1988), and the extent to which formal education took place in the L2 (e.g., Hakuta, Bialystok & Wiley, 2003). An equally important question concerns the nature of inter-individual variation (e.g., whether high levels of some forms of aptitude mitigate the effect of age of onset; Abrahamsson & Hyltenstam, 2008; DeKeyser, Alfi-Shabtay & Ravid, 2010). Finally, there is the question of intra-individual variation depending on the aspects of grammar or pronunciation concerned. In the area of grammar, syntax may be less sensitive to age effects than morphology (Johnson & Newport, 1989), regulars less than irregulars (see Hudson Kam & Newport, 2005, 2009), and salient structures less than non-salient ones (DeKeyser, 2000). Even for a given structure, age effects may be detected with ERP without showing up in the behavioral data (e.g., for subject–verb agreement in Chen, Shu, Liu, Zhao & Li, 2007). In the area of pronunciation, phonetic detail such as precise voice onset time (Abrahamsson & Hyltenstam, 2009) seems particularly problematic for older learners. Some phonetic cues to phonemic status may be easier to pick up than others (vowel duration being easier than closure duration; Baker, 2010); some suprasegmentals such as stress timing may be less sensitive to age than others (Trofimovich & Baker, 2006); and age may even affect different kinds of stress placement differently, the effect being strongest for stress determined by syllable structure (Guion, Harada & Clark, 2004). In sign language, handshape may be more resistant to age effects than location or movement (Morford & Carlson, 2011).

Those researchers who suspect that sensitive periods are maturational in nature have couched their causal explanations in both neurological and psychological terms. Neurological explanations have evolved over time from hemispheric specialization (e.g., Lenneberg, 1967) to myelination (e.g., Long, 1990) to varying rates of neurogenesis, synaptogenesis, or synaptic pruning (e.g., Uylings, 2006). These explanations have focused on the brain as a whole, while others more on specific areas such as the prefrontal cortex (e.g., Petanjek, Judas, Kostović & Uylings, 2008); the amygdala (e.g., Pulvermüller & Schumann, 1994); or the hippocampus, medial temporal lobe, and the basal ganglia (e.g., Ullman, 2004). Psychological explanations, rather surprisingly, came onto the scene later and have included growth of working memory capacity (the less-is-more hypothesis, e.g., Newport, 1990), increased susceptibility to proactive interference (e.g., Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann & Siebert, 2003), and gradual shifts from predominantly procedural/implicit to predominantly declarative/explicit processes (e.g., DeKeyser, 2000; Paradis, 2009; Ullman, 2004). Ultimately, of course, full explanatory adequacy will only be reached if psychological mechanisms can be tied to concurrent neurological developments that together explain the specific learning differences observed.

Empirical research on these issues is usually difficult for many reasons, in large part because the natural confounds of many of the variables involved cannot be experimentally disentangled in research on human learners. Perhaps the only notable exception is in the study of age of acquisition effects in sign language research (Mayberry, Lock & Kazmi, 2002), which is discussed in detail in the supplementary material Section S.1.

2.2 Grammatical gender

The linguistic phenomenon that our models will learn about is grammatical gender, which refers to an arbitrary classification of nouns, often marked by phonological, morphological, and/or semantic properties. Several studies suggest that grammatical gender is subject to sensitive periods. Studies with adult L2 learners of languages like French (Guillelmon & Grosjean, 2001), Spanish (Lew-Williams & Fernald, 2010) and German

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

(Scherag et al., 2004) have shown that non-native adults are slower than L1 speakers at processing nouns, and that their processing does not seem to benefit from patterns in gender agreement that are present in the language. Even childhood learners who begin acquiring French in an immersion program at age six do not achieve native-like gender agreement (Harley, 1979; Lapkin & Swain, 1977), indicating that acquisition of this grammatical component is subject to early age effects. For a more detailed background of the acquisition of grammatical gender, see Section S.2 in the supplementary material.

Gender systems vary in complexity; many Indo-European languages, such as French and German, divide nouns into only two or three gender classes, whereas Bantu languages employ extensive gender systems with up to twenty gender classes (Corbett, 1991). The degree to which grammatical gender is marked throughout a sentence also varies widely. In English, for example, gender is only marked on pronominals with animate reference, whereas gender in the Bantu language Swazi may be marked on adjectives, verbs, adverbs, numerals, and conjunctions.

The languages examined in the current study, French and Spanish, both assign masculine and feminine gender to all nouns; however, subtle differences between the gender classification systems exist. In French, a noun's final phoneme provides cues to gender, though the predictive value of the final phoneme is not always reliable. For example, according to Surridge (1993, 1995), only one "feminine" ([z]), and eight "masculine" ([æ], [ɛ̃], [ã], [ø], [o], [ʒ], [m], [ɛ]) final phonemes indicate gender with more than 90% accuracy; eight "masculine" ([f], [u], [a], [ʁ], [g], [y], [k], [b]) and nine "feminine" ([i], [ɔ̃], [n], [v], [j], [ʃ], [d], [s], [ɲ]) final phonemes indicate gender with 60–89% accuracy; and four final phonemes ([l], [m], [p], [t]) are considered ambiguous and do not provide any indication of the noun's gender. In addition, not everyone agrees with the phonemes' predictability values. For example, Lyster (2006) carried out a final phoneme predictive value analysis based on a corpus different from that of Surridge, and while the results are largely similar, some differences exist. Furthermore, the effect of a noun's phonological ending may be overridden by the noun's morphological ending (Surridge, 1989). Under this hierarchy, a word ending in the typically masculine final phoneme [ʁ] will be feminine when encompassed by the typically feminine morphological suffix *-ure*, as in *coiffure* "hairstyle". Overall, the French gender system is governed by patterns, but it is a complex system with many exceptions.

The Spanish gender system is less complex and more reliable than that of French. According to Teschner and Russell (1984), the majority of Spanish nouns' final phonemes are predictive of gender. Specifically, 90% of nouns ending in the phonemes [a] and [d] are feminine, and 89% of nouns ending in [e], [l], [o], and [ɾ] – which account for the majority of nouns – and also [i], [m], [t], [u], [x], [y], [b], [c], [tʃ] are masculine. Only three final phonemes, [n], [θ], and [s], are considered ambiguous in that they do not predict one gender over another. Morphological gender regularities in Spanish also exist, though they do not override final phonemes, as seen in French. Teschner and Russell identify seven morphological endings that are typically feminine (*-ción*, *-gión*, *-nión*, *-sión*, *-tión*, *-xión*, and *-ez*) and four morphological endings that are typically masculine (*-ón*, *-az*, *-oz*, and *-uz*). Note that these morphological endings encompass two of the ambiguous final phonemes, [n] and [θ], but not phonemes that are predictive of masculine or feminine. Finally, in both languages, animate nouns referring to humans assume semantic gender, so that the words for "man" and "woman" are masculine and feminine, respectively.

Both French and Spanish mark gender on determiners, pronouns, and adjectives. Examples of determiner and adjective markings are shown in sentences 1 and 2.

(1) "The little book is white."
French: **Le** peti**t** livre (MASC) est blan**c**.
Spanish: **El** libro (MASC) pequeñ**o** es blanc**o**.

(2) "The little table is white."
French: **La** peti**te** table (FEM) est blan**che**.
Spanish: **La** mesa (FEM) pequeñ**a** es blanc**a**.

French adjectives may end in almost any phoneme, with the feminine adjective typically marked by an additional and often unpredictable suffix. For example, the adjective *blanc* [blã] "white, MASC" becomes *blanche* [blãʃ] in its feminine form, and *petit* [pəti] "small, MASC" becomes *petite* [pətit]. A number of adjectives have the same phonological form for both masculine and feminine, even when the orthographic form differs. For example, the adjective "difficult" has only one orthographic (*difficile*) and phonological [difisil] form, and the adjective "expensive", while represented by two orthographic forms (*cher*, MASC; *chère*, FEM), are both pronounced [ʃɛʁ].

Spanish adjective formation, on the other hand, is more predictable. The majority of adjectives are marked by an *-o* ending for masculine, and an *-a* ending for feminine, as in *blanco*/*blanca* "white". As in French, not all adjectives have distinct orthographic and phonological masculine and feminine forms. Adjectives ending in *-e*, *-ista*, or a consonant, generally maintain the same form in both masculine and feminine, as in *verde* "green", *idealista* "idealist", and *difícil* "difficult". However, exceptions exist and certain types of adjectives ending in a consonant, such as those referring to nationalities, have a feminine form marked by an *-a* ending, as in *español*/*española* "Spanish" and *alemán*/*alemana* "German". Other exceptions include adjectives ending in

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

-ín, *-ón*, *-or*, such as *juguetón*/*juguetona* "playful" and *hablador*/*habladora* "talkative".

Despite the differences described above, the French and Spanish gender systems are similar in that both classify nouns into masculine and feminine based on phonological regularities, and gender is marked throughout a sentence on determiners, adjectives, and pronouns.

2.3 Memory development and language learning

Our neural network models undergo memory development, in the form of changes in both working memory capacity and long-term memory capacity, in order to examine the effects of maturation on sensitive period effects. At the most simple description, working memory (used interchangeably here with short-term memory) allows pieces of information to be held in the mind for brief periods of time in the absence of the input that caused them. In reality, working memory is most likely composed of a complex interaction of factors, such as attention (Conway, Cowan & Bunting, 2001; Engle, 2002; Kane & Engle, 2003), inhibition or filtering mechanisms (Vogel, McCollough & Machizawa, 2005), rehearsability (Baddeley, 2003; Gathercole & Baddeley, 1993; Wilson & Emmorey, 1997), and "chunking" strategies (Miller, 1956). Thus, although working memory is probably not a unitary construct, the core ability to store and integrate multiple items is critical to many aspects of cognitive functioning, including language processing. Working memory capacity refers to the number of items that can be stored and manipulated for a task. In general, higher capacities are associated with better cognitive function (Baddeley, 2003; Duncan, Seitz, Kolodny, Bor, Herzog & Ahmed, 2000) since lower capacities impose greater informational bottlenecks on processing. In development, working memory capacity grows rapidly from early childhood into adolescence, showing up to a three-fold increase (see Gathercole, 1999). This presents a paradox for language acquisition since higher cognitive function associated with higher memory capacity seems to be inversely correlated with overall language learning ability.

However, this is only a paradox if only the end state of development is considered. In reality, the maturation of working memory as well as language learning occur through time. One possibility is that limited cognitive ability, in particular a small memory capacity, is crucial to early stages of language acquisition, and that memory growth supports full language acquisition. Newport's (1988, 1990) less-is-more hypothesis draws upon data from cases where age of acquisition is NOT confounded with L1 entrenchment: the large proportion of deaf individuals who are not exposed to an accessible form of language early in life. During the language acquisition process and at final language attainment, these late learners have distinct profiles from early learners. As seen among hearing children during early stages of acquisition and word production, young signers (who have been exposed to American Sign Language since birth) morphologically simplify complex signs. This stage is considered to be important for morphological analysis of words and signs. Late learners do not make these types of errors or simplifications, rather processing the forms as "unanalyzed wholes" (Newport, 1990). As adults, these late learners use these complex forms in both ungrammatical and grammatical contexts, suggesting that they have not successfully learned their internal morphology. Early learners, in contrast, progressively develop the complex forms and do not make these types of mistakes as adults.

If the development of working memory is indeed inextricably linked with language acquisition abilities, there are two possible explanations for this relationship. The first is addressed by the less-is-more hypothesis, where the crucial factor is starting with a smaller working memory capacity (Newport, 1990). The rationale is that when a learning system is incapable of processing and holding in memory larger chunks of input, it is forced to analyze the input at lower level of complexity, picking out the highest-level and most prominent patterns while possibly abstracting away much of the detail. Another potential explanation comes from computational modeling, where it has also been demonstrated that controlling the size of the input, perhaps by providing smaller inputs at the beginning of training, contributes to better learning (Elman, 1993). Both models have been experimentally tested in adults, where smaller natural working memory or smaller inputs were associated with better detection of correlations between two binary variables (Kareev, Lieberman & Lev, 1997).

Most studies that directly investigate the relationship between working memory capacity and language learning in children suggest that the development of phonological short-term memory in particular is critical to word learning (Avons, Wragg, Cupplesa & Lovegrove, 1998; Baddeley, Gathercole & Papagno, 1998; Gathercole & Pickering, 2000). Higher spans in phonological short-term memory are linked with larger vocabulary sizes and better performance at learning new words. These working memory capacities are often measured by performance on non-word repetition tasks. However, these correlations leave the causal relationships inconclusive. The ability to temporarily store phonological traces of new utterances may be an important precursor to storing that item in long-term memory. On the other hand, it has been suggested that vocabulary growth leads to a better ability to analyze the representations into phonological segments, which in turn leads to more robust representations of new words (Metsala, 1999). What these two ideas agree on is the importance of the development of decomposed, sublexical representations – such as

phonemes – for language learning. Newport (1990) has made a similar argument about morphology. Drawing upon accompanying behavioral evidence that longer words are learned later in development than shorter words even when frequencies of these words are matched, Brown and Hulme (1996) demonstrate a computational model in which shorter words are maintained in short-term memory for longer given a limited short-term memory, facilitating encoding in long-term memory. A consequence of forming representations for smaller input first may be a better recognition of incremental patterns throughout learning.

2.4 Connectionist modeling

As discussed in Section 1 above, it is often difficult to experimentally separate the various possible causes of age effects when performing empirical research on human subjects. Computational modeling has a key advantage in its ability to independently manipulate a number of variables and to observe their main effects and interactions. Early attempts at computational modeling of linguistic sensitive periods (Goldowsky & Newport, 1993) show support for the less-is-more hypothesis in that a smaller working memory was shown to be better for the learning of some grammatical patterns, and this conclusion was supported by later studies, computational and otherwise (Cochran, McDonald & Parault, 1999; Kareev et al., 1997; Kersten & Earles, 2001). Neurocomputational modeling studies (reviewed in Hernandez & Li, 2007) favor explanations of age-related performance deficits in terms of changes in neural plasticity due the normal accumulation of experience. This idea, that the learning process itself could cause the observed sensitive period effects, is supported by many other modeling studies (reviewed in e.g., Thomas & Johnson, 2008) and has been called the "paradox of success" since learning one task to proficiency can harm the learning of other tasks (Seidenberg & Zevin, 2006). Sensitive period effects can be produced via the learning process itself in a number of ways, including entrenchment, where early experience leaves the learning system in a state not readily compatible with a new learning task; competition for resources between different tasks to be learned; and catastrophic interference, where a new learning task may impact performance on a previously learned task that is not actively maintained.

Previous neural network models that have dealt with aspects of memory development have used varying approaches to limiting working memory. Elman (1993) trained Simple Recurrent Networks (SRNs) on a complex subset of English. This type of network uses recurrent connections to allow the network to access its own previous states, creating an analog of working memory. Elman found that these networks had better eventual performance when this working memory was initially limited to a discrete window of a few steps and gradually increased, consistent with the less-is-more hypothesis. While others have failed to find a difference between developing and mature networks on similar tasks (e.g., Rohde & Plaut, 1999), Elman's study shows one way in which working memory capacity can be modeled in a neural network. As we will explain, our model uses a different approach, directly limiting the capacity of, or physical access to, previous states instead of limiting the network's temporal window of access to these states. Our approach is, in a sense, similar to that of the DevLex models of word and meaning acquisition (Li, Farkas & MacWhinney, 2004; Li, Zhao & MacWhinney, 2007), which utilize growing self-organizing maps to represent semantics and phonology. These maps grow by adding new units to accommodate storage of new lexical and semantic representations; as such, the growth involved more closely resembles long-term memory growth. Our model, in contrast, grows by adding new units that form the substrate for working memory.

There have also been a few notable neural network models that touch on the topic of grammatical gender. MacWhinney, Leinbach, Taraban and McDonald (1989) presented two neural network models of the acquisition of gender, case, and number in German. Both of these models learned to predict the article associated with a given noun, one using hand-coded semantic, phonological, morphological, and case cues, and the other using only observable data in the form of a complete phonological representation of the input noun along with some semantic and case cues. Both models succeeded at learning the nouns they were trained on, and also generalized very well to new nouns. The second model, without the hand-coded cues, outperformed the first. Unfortunately, the static phonological representations in this model only allow it to be applied to words of two syllables or fewer; our model employs temporal phonological representations that allow any word to be encoded. Additionally, Sokolik and Smith (1992) trained a feed-forward neural network to identify a corpus of French nouns as either masculine or feminine. Their study, however, has been widely criticized (Carroll, 1995; Matthews, 1999) for, among other things, using orthographic input, giving explicit gender feedback, and building in language-specific knowledge about gender classes. We believe that our approach adequately addresses these and other concerns, resulting in a model that only utilizes the information available to language learners.

3 Methods

Our intent in the present work is to use neural network models to understand any sensitive periods that arise due to the effects of cross-linguistic interference and aspects

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

of memory development. The first of these two factors is straightforward to implement: Simply teach a network to perform the same task in two languages. By varying the amount of time before the L2 is introduced, we can vary the expected amount of entrenchment of the L1. The second factor is developmental, and involves changes to a neural network's structure and connectivity over the course of the experiment, above and beyond the connection–weight changes that occur during normal training. So that readers who are perhaps only passingly familiar with neural networks can fully grasp the developmental aspects of the model, we include a primer on neural networks in Section S.3 of the supplementary material.

3.1 Our model

In the present work, we use a type of recurrent neural network architecture called the Long Short Term Memory (or LSTM; Gers & Cummins, 2000; Gers & Schmidhuber, 2001; Hochreiter & Schmidhuber, 1997). The LSTM architecture is similar in many ways to the well-known simple recurrent network (SRN) architecture (Elman, 1990), with two notable differences. First, the recurrence in LSTM comes not from a hidden layer and a copy-back context layer as in an SRN, but instead from hidden layer units, called MEMORY CELLS, that maintain their individual states across time-steps. This difference reflects a computational specialization of LSTM towards use as a substrate for working memory, as the maintenance of information across time is less noisy than in SRNs (Munakata, 2004). Combined with the slow weight changes characteristic of most neural network models, this makes the LSTM architecture well suited to its combined use in this study as a long-term categorization memory for learning the gender assignment and agreement tasks and as a working memory for temporarily storing the information relevant to each individual classification. The second difference is that each memory cell in an LSTM hidden layer is supplemented by a set of up to three additional units which serve to multiplicatively gate the inputs into, outputs from, and state retention of each memory cell. The network can learn to use these multiplicative gates to actively select important information to maintain in working memory while simultaneously reducing the kinds of interference that disrupt important working memory representations. A network composed of memory cells can maintain coherent working memory representations of important inputs for longer periods of time than architectures like the SRN. A more detailed primer on LSTM can be found in supplementary material Section S.4.

Our model learns by updating its connection weights based on the principle of gradient descent, utilizing back-propagation of error signals via an algorithm called LSTM-g (Monner & Reggia, 2012). While back-propagation has widely been regarded as neurobiologically implausible, Xie and Seung (2003) revealed gradient descent using back-propagation to be equivalent to a method of Hebbian learning utilized in neurobiologically plausible systems such as Leabra (O'Reilly & Frank, 2006). In light of this, it makes sense to view our use of back-propagation as a computationally expeditious equivalent of more neurobiologically plausible learning methods.

Since the aim of our model is to learn gender properties from speech stimuli, our neural network model is given an input layer able to represent one phoneme of speech at a time. The network is presented with a sequence of such phonemes, one after another, with the sequence as a whole representing a word or noun phrase. This is analogous to listening to spoken sentence fragments. The specific network architectures and desired outputs will differ by experiment and, as such, will be described in detail for each case in Section 4 below.

3.2 Development and network architecture

Since one aim of our model is to investigate the influence of development on learning of gender phenomena, we will next discuss the analogues of maturation in neural networks. Most neural network models have a fixed number of units and connections for the duration of training. Training such a network, starting from randomly assigned connection weights, is tantamount to waiting until a human learner is an adult, or at least fully neurologically developed in the relevant areas, before exposing him or her to any language stimuli. To address cases where language learning happens along with development, we also need to examine situations where the network structure develops during training. In the following paragraphs we examine a few ways of doing this.

In addition to the NO GROWTH condition, where all of the network's units and connections are present at the start of training, we examine a UNIT GROWTH condition in which the network begins with a much smaller number of units and connections (see Figure 2, top row). During the training regimen, new units and their associated connections are gradually added to the network until it reaches maturity, i.e. its maximum number of units and connections, equivalent to the numbers present in the no growth condition. Here, a new unit being added to the network is not necessarily analogous to neurogenesis in humans; instead, we take the view that some of the new connections, created through a process analogous to dendritic outgrowth (Uylings, 2006), happen to project to existing units outside our current view of the network, thus recruiting them for use in processing.

The unit growth condition described above confounds two variables of interest on the cognitive level. Recall

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

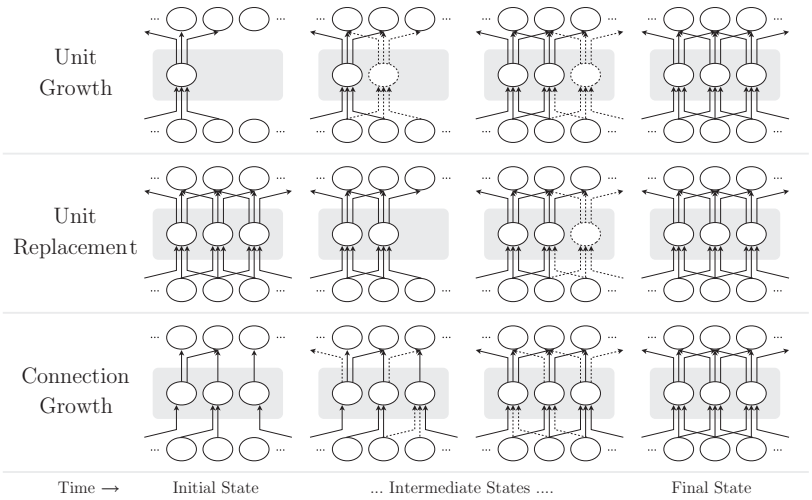

Figure 2. The various network development conditions, from least developed state on the left to most developed state on the right. The top row shows the unit growth condition, initially with few units in the shaded hidden layer; the network recruits new units via new connections (dashed) as time passes. The middle row shows the unit replacement condition, which starts with a full complement of units that are periodically removed and replaced with fresh units and untrained connections. The bottom row depicts the connection growth condition, which begins with a full complement of hidden layer units, though they are sparsely connected to the other units; as time passes, these units develop new connections until they reach a fully connected state.

that the activations of units in a recurrent neural network like ours are the basis of working memory. The network recruits new units during the maturation process, increasing the amount of information it can process at any given instant. We might reasonably expect this to correlate with an increase in cognitive measures of working memory capacity during training. Since these networks start with a small working memory and increase its capacity during training, we can evaluate the less-is-more hypothesis (Newport, 1990) for our model. From our perspective, this hypothesis admits two distinct and independently controllable factors that could lead to better final language performance: (i) starting with a small working memory, and (ii) allocation of new working memory resources during learning. Our unit growth condition possesses both factors, so to investigate them separately, we introduce a third network development condition, termed UNIT REPLACEMENT, that has only the second factor. This condition is not intended to correspond to human maturation; rather, it is included merely as a control to help us separate the effects of starting small from the effects of introduction of untrained resources. In this condition, depicted in Figure 2 (middle row), the network starts in the same state as the no growth condition, with its full complement of units and connections, and thus its full working memory capacity. Periodically, units and their associated connections are removed from the network and replaced with new units and fresh, untrained connections. This happens at a rate commensurate with the rate at which units are added in the unit growth condition. Thus, in both conditions fresh resources are introduced over time, but where the unit growth condition uses these resources to grow the network from its initially small size, the unit replacement condition accepts these fresh resources and discards an equal amount of its existing, trained resources, thereby maintaining a constant size. Since the effective size of the working memory does not change in the unit replacement condition, it allows us to determine if periodic introduction of fresh working memory resources alone, without starting small, can produce any significant benefits.

Working memory is not the only cognitive variable that changes as part of the unit growth condition. The new units that each network recruits must be wired up using new connections. Connections, as the reader will recall, are the basis of long-term memory capacity in a neural network. Thus, a network from our unit growth condition adds both working memory and long-term memory capacity during training. To tease apart these variables, we examine a fourth condition, termed the CONNECTION GROWTH condition, in which all units are present from the beginning but few of the possible connections exist (see Figure 2, bottom row). Since all units are incorporated from the beginning, the network's working memory capacity is fully developed from the start. During training, the network grows new connections at the same rate as in the unit growth condition, giving the network access to new long-term memory storage and allowing us to directly gauge the effects of long-term memory maturation. In addition, this allows us to

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

indirectly assess the contributions of working memory maturation (and compound effects) by subtractive analysis with the unit growth and no growth conditions.

4 Experiments and results

4.1 Gender assignment

In our first set of experiments, we investigate how well neural networks can learn to perform a gender assignment task using realistic sources of information. These networks take single nouns as input and use that information to predict which determiners can appear with that noun. Since nouns commonly occur with determiners in our target languages, French and Spanish, both the input and the output data are readily available to any learner by simply listening to everyday speech. After training, we determine the network's assignment of gender to individual nouns by presenting those nouns as input and observing the network's predictions for determiner pairing. The gender of the most strongly predicted determiner is taken to be the network's gender assignment for the input noun.

Our approach is similar to that taken by the third model from MacWhinney et al. (1989) in that our model uses the complete phonological form of a noun to predict the article to be used with that noun. We diverge from this earlier model in a few important ways. First, we eschew semantic features to investigate what can be learned from phonology alone. Even though phonology is predictive for the majority of words in our target languages, this choice deprives our model of information that learners are known to use (see Section 2.2 above). Second, we present the input noun as a temporal sequence of phonemes instead of a single phonological pattern, the latter of which will always have trouble representing long words or those that do not conform to the prespecified representational form. In addition, our approach corrects the most severe issues with the model of gender assignment by Sokolik and Smith (1992). Where their approach was criticized (Carroll, 1995; Matthews, 1999) for using orthographic input, we use phonemic input instead. Where their network came *a priori* equipped with knowledge of the genders of the training language – and indeed the knowledge that grammatical gender exists at all – our model has no such built-in knowledge. Finally, where their model required explicit feedback about the genders of individual words, our model relies instead upon the co-occurrence of gendered articles with nouns in order to deduce gender assignments. As a result of these differences, our model is more closely aligned with the real-world circumstances of human language learning in most contexts.

An input noun is presented to the network as a temporal sequence of phonemes. Each such phoneme is represented as a set of binary auditory features, with the activations of the network's input layer adjusted to reflect the feature set of each phoneme in turn. We use this representation because such features are universal in the sense that various configurations of these features can represent virtually any phoneme. As such, units representing these features could potentially be a built-in component of the brain of a language learner, or could be learned. That said, we only included enough features here to distinguish all phonemes in our target languages. The full set of phonemes and features are detailed in Table 1. After processing an entire sequence of phonemes representing the input noun, the network activates units in its output layer that correspond to determiners that it predicts to be compatible with the input noun. The network learns to perform this behavior by observing determiner–noun pairings and adjusting its connection weights accordingly.

The left half of Figure 3 shows the general architecture of the networks we train to perform this gender assignment task. The networks have an input layer of units corresponding to the on and off states of the features that make up the input phonemes. Units in the input layer project to units in a single hidden layer of memory cells. The intrinsic self-recurrence of the memory cells forms the substrate for working memory in the network. Finally, the hidden layer projects forward to the output layer which consists of nine units representing the definite and indefinite singular determiners of our target languages: *le*, *la*, *l'*, *un*, and *une* in French, and *el*, *la*, *un*, and *una* in Spanish. We do not posit that units representing these words could be built into the brains of language learners, nor that the words are represented in single units. However, since these determiners form a small closed class of words, we feel it is not too large a leap to presume that the learner represents these frequent determiners as distinct entities before much gender learning takes place. Our single-unit representation for each determiner is the simplest possible in this context, though other representations would likely work as well.

For this set of experiments, we used the 600 French words from the Sokolik and Smith (1992) paper as the input data for our model, and a set of 600 equivalent words from Spanish. For each trial during training, we first select a language and then select a noun at random from our corpus. We pair the noun with either a gender-matched definite or indefinite determiner from the appropriate language to form a simple noun phrase. The noun is given as input to the network, which then predicts applicable determiners and adjusts its weights in such a way that, in the future, it will be more likely to predict the determiner that actually co-occurred with the input noun. A network is considered to have assigned the correct gender for an input noun if an article of the appropriate gender is most active after presentation.

To determine a baseline level of performance on the gender assignment task, we trained networks on either French or Spanish only and recorded their performance.

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

Table 1. *Binary feature representations of phonemes.*

	consonantal	sonorant	continuant	strident	nasal	lateral	trill	voice	labial	round	coronal	anterior	distributed	dorsal	high	low	back	radial	adv tongue root
h	–	+	+	–	–	–	–	–	–	–	–	–	–	+	+	–	+	–	–
ɟ	–	+	+	–	–	–	–	+	–	–	–	–	–	+	+	–	–	–	–
ɛ	–	+	+	–	–	–	–	+	–	–	–	–	–	+	–	–	–	+	–
e	–	+	+	–	–	–	–	+	–	–	–	–	–	+	–	–	–	+	+
ə	–	+	+	–	–	–	–	+	–	–	–	–	–	+	–	–	+	–	–
a	–	+	+	–	–	–	–	+	–	–	–	–	–	+	–	+	+	+	–
i	–	+	+	–	–	–	–	+	–	–	–	–	–	+	+	–	–	+	+
ɔ	–	+	+	–	–	–	–	+	+	+	–	–	–	+	–	–	+	+	–
o	–	+	+	–	–	–	–	+	+	+	–	–	–	+	–	–	+	+	+
y	–	+	+	–	–	–	–	+	+	+	–	–	–	+	+	–	–	+	+
u	–	+	+	–	–	–	–	+	+	+	–	–	–	+	+	–	+	+	+
œ	–	+	+	–	–	–	–	+	+	+	–	–	–	+	–	–	–	+	–
ø	–	+	+	–	–	–	–	+	+	+	–	–	–	+	–	–	–	+	+
ɛ̃	–	+	+	–	+	–	–	+	–	–	–	–	–	+	–	–	–	+	–
ã	–	+	+	–	+	–	–	+	–	–	–	–	–	+	–	+	+	+	–
ɔ̃	–	+	+	–	+	–	–	+	+	+	–	–	–	+	–	–	+	+	–
æ̃	–	+	+	–	+	–	–	+	+	+	–	–	–	+	–	–	–	+	–
k	+	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	+	–	–
t	+	–	–	–	–	–	–	–	–	–	+	+	–	–	–	–	–	–	–
p	+	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–
g	+	–	–	–	–	–	–	+	–	–	–	–	–	+	+	–	+	–	–
d	+	–	–	–	–	–	–	+	–	–	+	+	–	–	–	–	–	–	–
b	+	–	–	–	–	–	–	+	+	–	–	–	–	–	–	–	–	–	–
ɣ	+	–	+	–	–	–	–	+	–	–	–	–	–	+	+	–	+	–	–
ð	+	–	+	–	–	–	–	+	–	–	+	+	–	–	–	–	–	–	–
θ	+	–	+	–	–	–	–	–	–	–	+	+	–	–	–	–	–	–	–
β	+	–	+	–	–	–	–	+	+	–	–	–	–	–	–	–	–	–	–
ʃ	+	–	+	+	–	–	–	–	–	–	+	–	+	–	–	–	–	–	–
s	+	–	+	+	–	–	–	–	–	–	+	+	–	–	–	–	–	–	–
f	+	–	+	+	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–
ʁ	+	–	+	+	–	–	–	+	–	–	–	–	–	+	–	–	+	–	–
ʒ	+	–	+	+	–	–	–	+	–	–	+	–	+	–	–	–	–	–	–
z	+	–	+	+	–	–	–	+	–	–	+	+	–	–	–	–	–	–	–
v	+	–	+	+	–	–	–	+	+	–	–	–	–	–	–	–	–	–	–
ʎ	+	+	–	–	–	+	–	+	–	–	–	–	–	+	+	–	–	–	–
l	+	+	–	–	–	+	–	+	–	–	+	+	–	–	–	–	–	–	–
ɲ	+	+	–	–	+	–	–	+	–	–	–	–	–	+	+	–	–	–	–
ŋ	+	+	–	–	+	–	–	+	–	–	–	–	–	+	+	–	+	–	–
n	+	+	–	–	+	–	–	+	–	–	+	+	–	–	–	–	–	–	–
m	+	+	–	–	+	–	–	+	+	–	–	–	–	–	–	–	–	–	–
j	+	+	+	–	–	–	–	+	–	–	–	–	–	+	+	–	–	–	–
ɾ	+	+	+	–	–	–	–	+	–	–	+	+	–	–	–	–	–	–	–
w	+	+	+	–	–	–	–	+	+	–	–	–	–	+	+	–	+	–	–
r	+	+	+	–	–	–	+	+	–	–	+	+	–	–	–	–	–	–	–

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

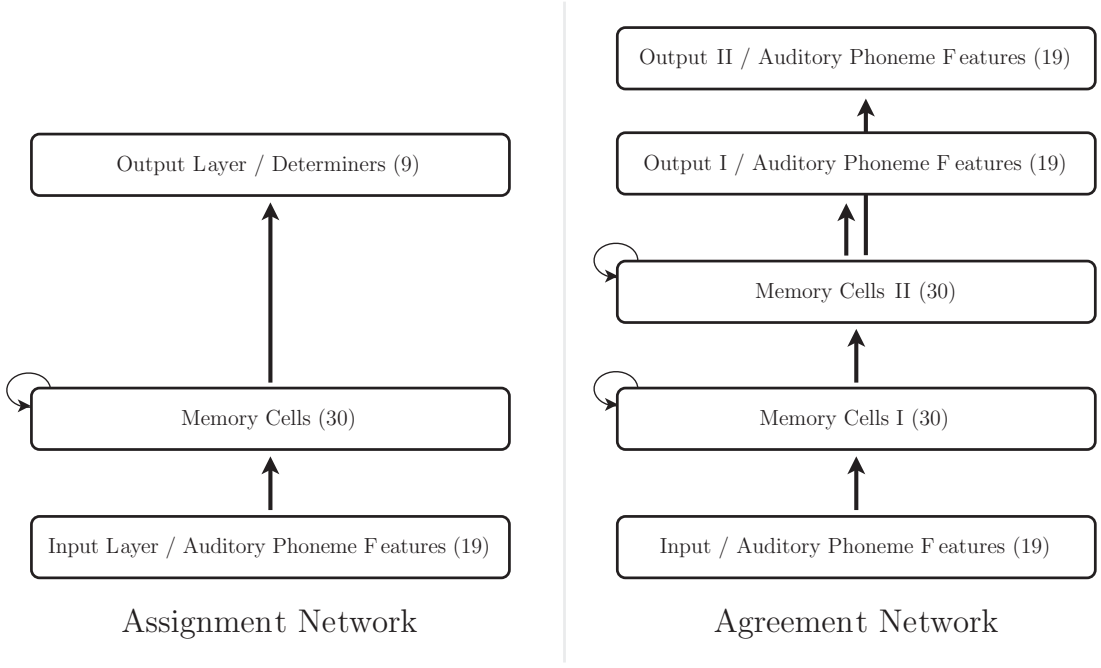

Figure 3. On the left is shown the architecture of networks used in the gender assignment task. The assignment network takes features of auditory phonemes as input, passes them through a hidden layer of self-recurrent memory cells, and maps a sequence of such inputs onto an output layer of units representing determiners in two languages. On the right is the architecture of networks used in the gender agreement task. The agreement network also takes auditory phoneme features as input, but passes them through a series of two hidden layers of memory cells. After processing each input phoneme, the network uses its two output layers to predict the next two phonemes that it will be given as input. Though no recurrent connections are depicted at this level for either network, each individual memory cell is self-recurrent, remembering its activation from the previous step.

The results are shown in the left half of Figure 4. As one would expect, networks trained on French alone scored well in excess of 90% after training, while scoring at chance on Spanish; similarly, Spanish-trained networks performed well on their native language and at chance on French. It is worth noting that Spanish performance was consistently a few percentage points better than French performance, likely due to the phonemic cues to gender assignment in Spanish being simpler and more reliable than those in French. Performance was consistent across the four development conditions, suggesting that, alone, network development has little impact on outcomes for the gender assignment task, at least in the first language.

With a baseline level of performance established for networks that are "native" to either French or Spanish, we next investigated the performance of bilingual networks under a number of different learning conditions designed to assess the role of L1 entrenchment. Each condition varies the length of time t which the network spends learning the task on L1 alone before L2 is introduced (Zhao & Li, 2010). We describe the conditions in terms of two periods, the first of which consists of training only in L1 for t trials, where t varies widely across conditions. This is immediately followed by the second period, in which L1 and L2 trials are mixed with equal probability. The duration of the second period is always two million trials in an effort to ensure that the networks have time to reach peak performance on both languages. While this second mixed training period will undoubtedly create competition and interference between the two languages, the amount of interference should be the same in each condition because the size and mix proportion of the second training period are the same across conditions. In contrast, the amount of L1-only training prior to the introduction of L2 is varied across conditions, meaning that networks that start with different values of t will end up in different states – reflecting differing levels of entrenchment – when L2 training begins.

A network whose training regimen has $t = 0$ is a native bilingual in the sense that L1 and L2 are presented at precisely the same time, and in the same proportions. Thus, such a network should exhibit no L1 entrenchment. Networks trained with higher values of t, having had a longer time with exposure only to L1, should exhibit more entrenchment. Given this, the prevailing ideas about L1 entrenchment offer a number of predictions about the final, peak L1 and L2 performance of the networks:

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

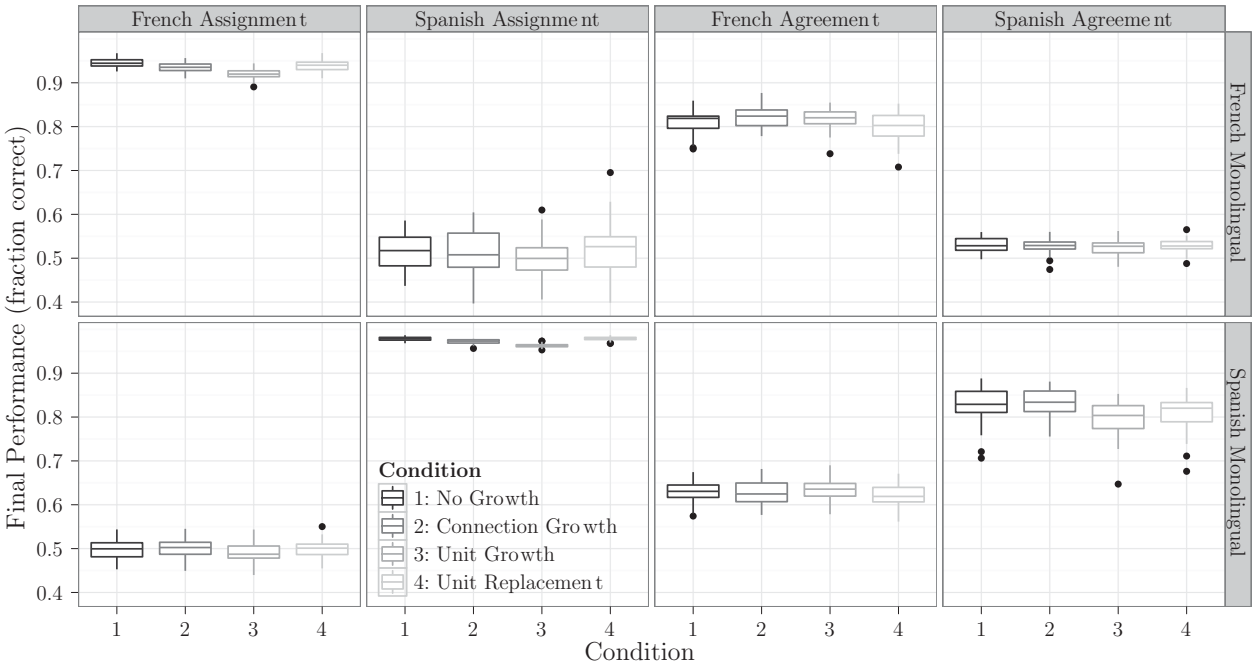

Figure 4. Results for monolingual networks, of all four developmental varieties, on the gender assignment and agreement tasks. We trained 30 separate networks in each developmental condition for each language. Each network was trained for two million trials in one language, and then evaluated on both languages.

1. Networks' final L1 performance should not decrease as t increases.
2. Networks trained with $t = 0$, as native bilinguals, should not exhibit impairment in either language with respect to the other.
3. Networks should show increasing degradation of final L2 performance as t increases, at least until the networks have mastered L1 to a point at which the effect of entrenchment saturates.

These predictions can be investigated by plotting the final L1 and L2 performance of fully trained networks on the gender assignment task versus the value of t with which they were trained. We trained 30 separate networks for each of 15 values of t as well as for each of the four maturation conditions and each of two languages; thus a total of 3,600 networks were trained to produce the following figures. For conditions in which the network matures during training, each of these networks begins training in its most immature state and develops over the course of the first 400,000 trials, at which point it reaches maturity – i.e. architectural parity with the networks in the no growth condition. Thus, some networks in the connection growth and unit growth conditions (i.e. those with $t = 0$) are first exposed to L2 in their most immature state, while others (i.e. $t = 400{,}000$ and above) are not exposed to L2 until after reaching maturity.

After both training periods were complete, we recorded the fraction of inputs to which each network assigned the correct gender, for both languages, and plotted them in the left half of Figure 5. These graphs depict the final performance of the networks on the y-axis versus the value of t – i.e. the duration of the L1-only training period and thus the delay before L2 onset relative to L1 – on the x-axis. Thus, the expected L1 entrenchment increases from negligible to maximal as we move from left to right in each figure; another way of saying this is that the networks towards the left of the x-axis are closer to true bilinguals whereas the networks closer to the right edge are late L2 learners. The y-axis values always depict final performance after the conclusion of training. These graphs show fitted curves for each of the different network maturation conditions, and for each such curve, the shaded area behind it represents the 95% confidence interval.

To examine the first prediction above, we first look at the performance of the various networks in their native language. Table 2 shows the results of a statistical analysis on the performance results, comparing the means for each condition at the first and last t-values using a two-proportion z-test. In addition to statistical significance, the table also provides an indication of the magnitude of the performance change. This answers the question, for each condition, of whether performance is statistically flat, increasing, or decreasing as t values increase. The table also shows codes for the magnitude of the significant changes in performance. The prediction of non-decreasing performance with increasing t appears to be largely borne out. When native language and task match, performance

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

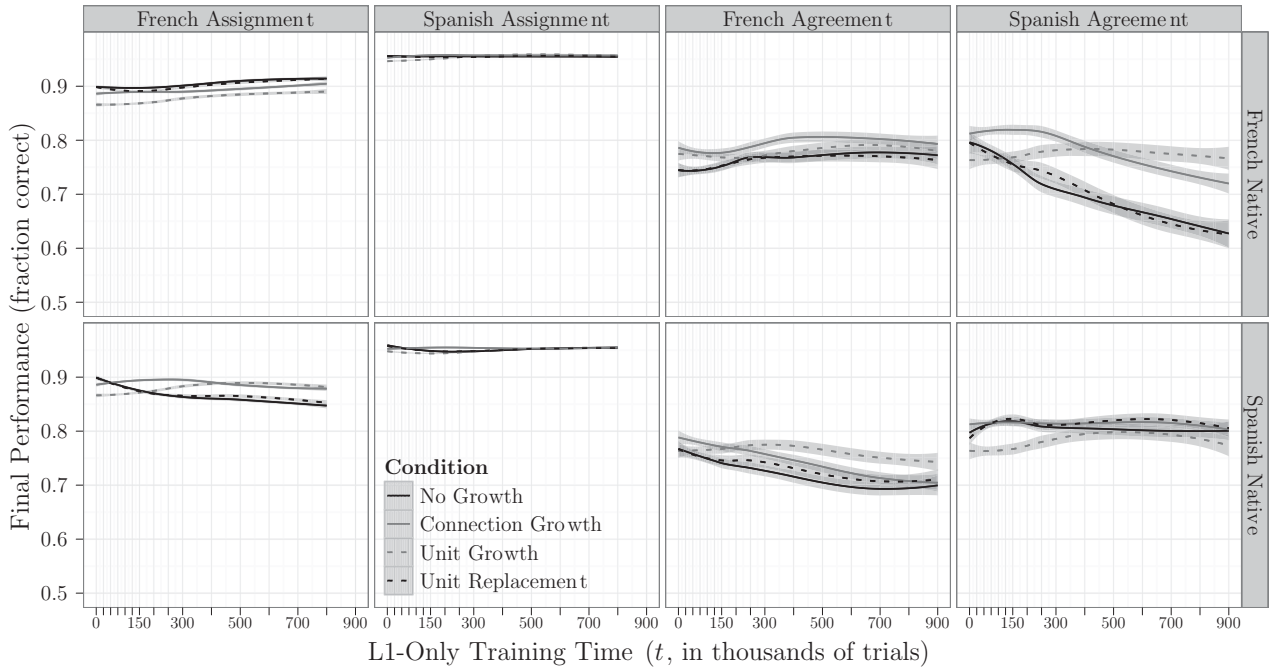

Figure 5. Results for bilingual networks on the gender assignment and agreement tasks. Each network was assigned an L1 and trained initially on only that language before the other language was introduced. The x-axis varies the time t each network spent with L1 in isolation before L2 was introduced. Note that $t = 0$ corresponds to a native bilingual network for which neither language is ever prioritized, while larger values of t correspond to increased time spent with L1 alone, and thus presumably increased levels of L1 entrenchment. The y-axis shows the final performance on the language task after training was complete. We trained 30 separate networks in each combination of developmental condition, native language, and t-value (shown as ticks on x-axis). The networks were trained for t trials of L1 alone before a period of 2 million trials of L1 and L2 in equal proportion. The lines depicted for each combination of development condition and task are smoothed-average curves shown over the sampled values of t, which are indicated by the vertical lines in each pane.

is flat or occasionally weakly increasing as t values rise. Figure 5 shows us that, as with the monolingual networks, bilingual networks have a slightly harder time learning French than Spanish as an L1. Differences in Spanish performance between the different maturational variants of Spanish-native networks were minimal, while the French-native networks that grew their working memory capacity during training showed a slight disadvantage. However, the expected general pattern of flat or improving performance with increasing t held for all conditions.

We can investigate the second prediction above by examining each network's performance on its second language. We do this by comparing second-language performance of networks with $t = 0$ on the x-axis to the native-language performance. We see that across all maturational conditions, the true bilingual networks (those with $t = 0$) perform well when compared to the native networks in both languages, lending support to the second prediction above.

Moving on to the third prediction, we can clearly see a t-related performance deficit in the no growth condition for L2 French; increasing Spanish exposure before French is introduced causes the final French performance of the network to decrease at a rate that is at first rapid but eventually slows for larger delays. The maturational properties in play for the connection growth and unit growth conditions, however, appear to have helped these networks compensate for the expected declines in French performance due to Spanish entrenchment. Networks in the unit replacement condition tended to perform at levels comparable to the no growth networks, suggesting that introduction of new working-memory resources without starting small may not be sufficient to gain a significant reprieve from the deleterious effects of increasing L1 entrenchment. In the case where French was the L1 and Spanish the L2, no appreciable t-related performance decreases were observed. We expect that this is due again to the relative ease of the task for Spanish as compared to French.

At least in the case of French as an L2, the data shown in Figure 5 and Table 2 seems to support both the predictions of performance declining due to increased entrenchment and of maturation during learning helping to overcome these difficulties. Next we trained networks on the more difficult task of gender agreement, the results of which are reported in the next section.

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

Table 2. *Significance and magnitude of performance change as* t *increases.*

Task	Language	Development	Slope	Significance	Magnitude
French assignment	French native	No growth	Increasing	***	*
		Connection growth	Increasing	***	*
		Unit growth	Increasing	***	*
		Unit replacement	Increasing	***	*
	Spanish native	No growth	Decreasing	***	**
		Connection growth	Flat		
		Unit growth	Increasing	***	*
		Unit replacement	Decreasing	***	**
Spanish assignment	French native	No growth	Flat		
		Connection growth	Flat		
		Unit growth	Increasing	***	*
		Unit replacement	Flat		
	Spanish native	No growth	Flat		
		Connection growth	Flat		
		Unit growth	Increasing	***	*
		Unit replacement	Flat		
French agreement	French native	No growth	Increasing	***	*
		Connection growth	Increasing	**	*
		Unit growth	Increasing	***	*
		Unit replacement	Flat		
	Spanish native	No growth	Decreasing	***	**
		Connection growth	Decreasing	***	***
		Unit growth	Decreasing	***	*
		Unit Replacement	Decreasing	***	**
Spanish agreement	French native	No growth	Decreasing	***	*****
		Connection growth	Decreasing	***	****
		Unit growth	Increasing	***	*
		Unit replacement	Decreasing	***	*****
	Spanish native	No growth	Increasing	***	*
		Connection growth	Decreasing	***	*
		Unit growth	Increasing	***	*
		Unit replacement	Increasing	***	**

Significance: *** for $p < .001$, ** for $p < .01$, * for $p < .05$
Magnitude: ***** for m > 12%, **** for m > 9%, *** for m > 6%, ** for m > 3%, * for m > 1%

4.2 Gender agreement

Our second set of experiments explores how neural networks perform on a gender agreement task. During a trial, networks in these experiments receive a noun phrase (e.g., *el mecanismo interno* "the internal mechanism" in Spanish) presented as an unsegmented sequence of phonemes (e.g., [elmekanismointerno]) as input. The network's job at every point in this phoneme sequence is to predict the next few phonemes that it will hear. As such, the network uses a phonemic representation of everyday speech as both the input and the training signal. After training, the network's gender agreement performance is evaluated using noun phrases of the form

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

Table 3. *Determiners used by the gender agreement model.*

French	Spanish
a, le, l', un, une, ce, cette, cet, aucun, aucune, chaque, tel, telle, sa, son, ma, mon, ta, ton, notre, votre, leur	el, la, un, una, este, esta, ese, esa, aquel, aquella, ningún, ninguno, ninguna, cualquier, cualquiera, cada, su, tu, mi, nuestra, nuestro, vuestra, vuestro

determiner–noun–adjective – common constructions in our target languages of Spanish and French. To determine gender agreement, we give the network the determiner and noun as input, followed by the portion of the adjective that is gender-neutral, and ask the network to predict the correct ending for the adjective. If the network predicts the gender-appropriate ending more strongly than the gender-inappropriate ending, we consider the network's answer to be correct.

The noun phrases we used as training data for the gender agreement task were extracted from the French and Spanish versions of Wikipedia (2011). We downloaded archives containing the complete text of each version of Wikipedia and applied part-of-speech tags to each word using TreeTagger (Schmid, 1994). We then extracted all noun phrases of the forms determiner–noun, determiner–noun–adjective, and the less frequent determiner–adjective–noun, where the determiner is one from Table 3. From this list of noun phrases we removed any phrases containing words that were not in our language dictionaries – Lexique 3 for French (New, 2006) and CUMBRE for Spanish (CUMBRE, n.d.). Finally, we extracted the most frequent 100,000 noun phrases for each language. These phrases comprise the training data. On each training trial, we chose a phrase probabilistically, based on the phrases' corpus frequencies; we used this phrase as the input – and training signal – for the network on that trial.

The right side of Figure 3 presents the architecture of the networks trained on the gender agreement task. The input layer is the same as it was for the gender assignment experiments, with each input unit corresponding to a binary auditory feature of a phoneme. These networks, however, have two hidden layers of memory cells instead of one. This is because the gender agreement task involves two separate levels of segmentation of the input. To perform the task effectively, we expect that any learner needs to divide the phoneme sequence first into morphemes and words and, at a higher level, into noun phrases in which gender agreement must be maintained. Previous experiments with these types of networks on language tasks (Monner & Reggia, 2012) have shown a network with two hidden layers to be more effective in this case than networks with a single hidden layer.

The network's output layers are each identical to the input layer because the network is predicting upcoming phonemes. There are two such output layers because the network must predict not only the next phoneme that will occur in the input, but the phoneme after that as well. We require the network to make predictions of two future phonemes because some of the gendered adjective endings that we would like to predict consist of two phonemes. For example, the French adjective for "particular" is *particulier* [paʁtikylje] in the masculine and *particulière* [paʁtikyljɛʁ] in the feminine; we can see in the phonetic spellings that the gendered endings of these adjectives differ across two phonemes, with [-e] ending the masculine form and [-ɛʁ] ending the feminine form. Since we can only show the network the gender-neutral portion of the phoneme sequence (i.e. [paʁtikylj-]) without giving away the gendered form intended by the speaker, we must have the network predict two subsequent phonemes (either of which may be null if subsequent phonemes do not exist) in order to capture gendered endings with two phonemes such as [-ɛʁ].

When evaluating performance on the gender agreement task after training, we use only phrases of the determiner–noun–adjective form because it is the only form that is adjective-final. Our testing paradigm requires an adjective-final form because the network must predict the gender-appropriate ending of the last word, and only adjectives generally have two distinct gendered endings. Gender-neutral adjectives, and adjectives where the two gendered forms are orthographically distinct but phonetically identical (e.g., in French, the masculine *architectur**al*** and the feminine *architectur**ale*** are both pronounced [aʁʃitɛktyʁal]), are present during gender agreement training but ignored during the performance evaluation.

To determine a baseline level of performance on the gender agreement task, we trained sets of networks on either French or Spanish only and recorded their performance. The results are shown in the right half of Figure 4. As was the case with the gender assignment task from the previous section, we find here that networks trained on French do well on French and perform at chance on Spanish. Networks trained on Spanish perform as expected on that language and do significantly worse on French.

We use the same experimental setup as in the gender assignment task to investigate the effects of L1 entrenchment alone (i.e. the no growth condition) and together with network maturation (i.e. the unit growth, unit replacement, and connection growth conditions) in the gender agreement task. As before, training consists of two periods, the first consisting of t trials in which inputs come exclusively from the designated L1, and the second consisting of two million trials where inputs may

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

be drawn from either language. We trained 30 networks in each maturation condition and for each value of *t*, the duration of the initial L1-only training period. The results are shown in the right half of Figure 5 above.

We can examine the networks' performance on their first languages, broken out by language and maturation condition as before, by looking at the bottom half of Table 2. As expected, we do not see decreasing performance with increasing *t* in any of the conditions where the task and native language match.

Next we examine the final performance scores on L2 for networks in each condition of the gender agreement task as a function of *t* on the x-axis. The results for both languages here are similar to what we observed in the gender assignment task for the case of L2 French. The mature networks in the no growth condition show a marked susceptibility to L1 entrenchment, with L2 performance decreasing by as much as 17% as *t* is increased, delaying the onset of L2 relative to L1. However, the networks in the unit growth condition were largely able to mitigate this performance decrease by introducing new units and connections during learning. Performance of networks in the connection growth condition fall between these two. The addition of new connections to the networks appears to successfully stave off entrenchment effects when the level of entrenchment is small, but for values of $t > 200{,}000$ the entrenchment effects again start to become apparent. This tells us that addition of new units and new connections both help to counteract deficits due to entrenchment. Viewed from the cognitive perspective, growth in long-term memory capacity – in the connection growth condition – during training helped to mitigate the effects of L1 entrenchment, as did growth in working memory capacity, as evidenced by the superior performance of the unit growth condition over the connection growth condition for higher values of *t*. However, as shown by the unit replacement networks again tending to track the performance of the no growth networks, the addition of fresh neural resources is not all that is required to reap a performance benefit. Instead, it seems that starting small, either in terms of working memory capacity or long-term memory capacity, or both, is an essential factor that, combined with growth of neural resources, leads to the performance increase.

5 Discussion

The data presented in Section 4 (with one exception, discussed in detail below) appears to support the predictions of established ideas of L1 entrenchment: Increasing levels of entrenchment of the L1 caused increasing difficulty in acquiring an L2. The most dramatic of these can be seen clearly in the no growth conditions, where we witness an initially steep decline in learnability of the L2 task as time spent on the L1 task increases. The simulation results also largely agree with conclusions of empirical studies of gender learning in both early and late bilinguals (discussed in Section S.2 of the supplementary material) in that early L2 learners perform much like native speakers, whereas later L2 introduction leads to poorer performance.

The simulations also bore out the predictions of the less-is-more hypothesis, with the networks that undergo working memory development outperforming those that started with full-sized working memory capacities. Our experimental efforts to separate the effects of starting with a small working memory from those of simply adding fresh memory resources showed a distinct advantage to growth combined with starting small. This not only provides a small but important clarification to the mechanism behind the less-is-more hypothesis, but is a result for which an empirical investigation would be difficult if not impossible. We treat the results pertaining to each hypothesis in separate sections below.

5.1 Entrenchment

As mentioned earlier, our simulations provided one exception to our hypothesis about entrenchment, in the form of French-native learners of the Spanish gender assignment task attaining near-native-like performance levels on their L2 task. This may be explained, in whole or in part, by the inherent similarity of French and Spanish; see our discussion of the empirical study by Sabourin, Stowe and de Haan (2006) in Section S.2 of the supplementary material. When two languages are very similar, one might expect L2 learning to be easier where it agrees with L1 and harder where it disagrees. For example, the fact that a noun ending in [o] is a very reliable predictor of masculine gender in both French and Spanish may underlie the unexpected ease with which our French-native networks learned the Spanish gender assignment task: Since masculine nouns ending in [o] are so prevalent in Spanish, the transfer of this concordant rule from L1 French would immediately improve accuracy by leaps and bounds. The reverse – transferring the rule from L1 Spanish to L2 French – would not be as beneficial since masculine nouns ending in [o] are far less prevalent in French than in Spanish, thus leading to less of an impact on the learner's overall accuracy. On the other hand, it may be more difficult for native French speakers to learn Spanish's association between [a] and feminine gender given that [a] is associated with the masculine in French. In our simulations, this rule may have had less of an impact because [a] is a less reliable cue in French; or perhaps it is the case that discordant rules from L1 can be easily overcome. The simulations reported here certainly do not fully explore interactions between language similarity, rule transfer, and ease of L2 learning. To better grasp the significance of interactions between concordant and

https://doi.org/10.1017/S1366728912000454 Published online by Cambridge University Press

discordant rules like the examples above, we hope in the future to study an expanded model that includes more languages of varying levels of similarity.

5.2 Memory development

While our modeling approach does not directly implement cognitive constructs such as working memory capacity, we argued in Section 3.3 that the connection growth condition could be reasonably conceived as representing growth from an initially small long-term memory capacity, and the unit growth condition as growth of both long-term and working memory capacities from small beginning states. Allowing the networks to mature in either of these conditions helped to mitigate the negative impacts of L1 entrenchment, especially for longer delays in L2 onset. The fact that the connection growth condition generally improves upon the no growth condition suggests to us that growth of long-term memory capacity may be a key maturational factor during language learning. For the longest delays, the unit growth condition appears to have had the greatest positive impact, which suggests to us that growth of working memory capacity also has a positive influence in combating entrenchment effects.

The unit replacement condition, on the other hand, demonstrated the effects of adding fresh long-term and working memory resources to the network without starting small, and without changing the network's overall size. Since the networks in this condition did not do substantially better than those in the no growth condition, we have to conclude that the only thing lacking in the unit replacement condition – beginning from resources of modest capacity, or starting small – is an essential factor underlying the performance gains made by the unit growth and connection growth networks. This lends support to the less-is-more hypothesis, and further constrains it in the sense that it is now clearer that initial size is crucial; the effect is not caused by resource acquisition alone.

The less-is-more hypothesis is usually presented at the cognitive level, suggesting that a system with limited cognitive resources will latch on to the low-hanging organizational fruit, learning representations efficient enough to accommodate its small memory capacity. This can serve as a boon later on, when new memory capacity is added and can tackle more complex stimuli. This proposal also makes intuitive sense at the level of neural information processing for a variety of reasons. A network that has its full complement of resources when learning begins naturally learns to use all the resources at its disposal to widely distribute its learned interpretations of its L1 experiences. If an L2 is introduced later, the distributed L1 experience cannot be easily or quickly consolidated to use only a subset of the neural resources so as to free up some of these for the L2 alone. Instead, the L2 and L1 experiences intermix and interact, exacerbating L1 entrenchment effects and prolonging performance deficits in L2 due to resource competition from L1 (Thomas, 2009). On the other hand, a network that begins training with more modest resources will be forced to attempt to encode the L1 using only the limited resources available. Though these may initially be insufficient for a full understanding of L1, the limitations will force the network to adopt more efficient and less widely distributed encodings of the L1. This may entail segmenting the input into smaller generative chunks, like phonemes and morphemes. This consolidation of L1 knowledge in the resources that were added early leaves the later-added neural resources free to adapt to novel data such as that presented by an L2. If this story is correct, starting with fewer resources and building them up during language learning are key strategies to developing more modular representations for each language, which helps to avoid the deleterious effects of L1 entrenchment and resource competition with L2.

Our simulation results also showed that our networks' final performance when learning only one language was generally the same or worse in the developing conditions compared to the pre-developed no growth condition. This stands apart from previous results showing that late L1 acquisition of sign languages is impaired proportionally to age of acquisition (e.g., Mayberry, 1993; Mayberry & Lock, 2003), the explanation for which is thought by many to be developmental in nature. We see two potential explanations for this discrepancy. The first and most obvious is that our model does not account for the mechanism, developmental or otherwise, that underlies these impairments in late L1 acquisition. A second possibility that affords our model some explanatory power rests on the idea that the observed performance deficits in a late-learned L1 are due to entrenchment and/or interference from home sign systems developed by the learners prior to exposure to a conventional sign language (Seidenberg & Zevin, 2006). Under this view, a late-learned L1 functions more like an L2, creating a situation that is more directly comparable to our bilingual networks than the monolingual ones, which had no prior exposure to any type of communication system which could interfere or become entrenched. While this interpretation minimizes the discrepancy between our model and empirical findings, it remains a controversial hypothesis regarding the origin of late L1 learning deficits.

We do not mean this work to in any way suggest that entrenchment and memory development explain all the age effects we see in second language learning. As many researchers have pointed out, cognitive maturation is typically confounded with a variety of other changes that take place in the same time frame, such as social development, changing patterns of input and interaction, and schooling in the L2. While we acknowledge that the factors we have studied here do not explain all the age

effects observed in humans, we do believe they are part of a larger picture involving many of the variables outlined above. Our simulations confirm that entrenchment – a natural consequence of learning different tasks in stages – can indeed cause large deficits in second language performance. Our comparison of developmental conditions bears out the predictions of the less-is-more hypothesis, showing that memory development – that is, starting from a small memory and growing it during learning – can help to prevent disruptions due to entrenchment. While much more work is necessary to determine how cognitive maturation contributes to age effects, this study contributes to a better understanding of how memory development in particular could be an important part of that picture.

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.

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59 (2),* 249–306.

Avons, S. E., Wragg, C. A., Cupplesa, W. L., & Lovegrove, W. J. (1998). Measures of phonological short-term memory and their relationship to vocabulary development. *Applied Psycholinguistics, 19,* 583–601.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36,* 189–208.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158–173.

Baker, W. (2010). Effects of age and experience on the production of English word-final stops by Korean speakers. *Bilingualism: Language and Cognition, 13 (3),* 263–278.

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. New York: Newbury House.

Brown, G. D. A., & Hulme, C. (1996). Non-word repetition, STM, and age-of-acquisition: A computational model. In S. E. Gathercole (ed.), *Models of short-term memory*, pp. 129–148, Hove: Psychology Press.

Carroll, S. E. (1995). The hidden dangers of computer modeling: Remarks on Sokolik and Smith's connectionist learning model of French gender. *Second Language Research, 11 (3),* 193–205.

Chen, L., Shu, H., Liu, Y., Zhao, J., & Li, P. (2007). ERP signatures of subject–verb agreement in L2 learning. *Bilingualism: Language and Cognition, 10 (2),* 161–174.

Cochran, B. P., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language, 41,* 30–58.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of WM capacity. *Psychonomic Bulletin & Review, 8,* 331–335.

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

CUMBRE (n.d.). Corpus del Español Contemporáneo de España e Hispanoamérica. Madrid: SGEL.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22 (4),* 499–533.

DeKeyser, R. M. (2012). Age effects in second language learning. In S. Gass & A. Mackey (eds.), *Handbook of second language acquisition*, pp. 442–460. London: Routledge.

DeKeyser, R. M., 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. M., & 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. 89–108. Oxford: Oxford University Press.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., & Ahmed, A. (2000). A neural basis for general intelligence. *Science, 289 (5478),* 457.

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

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11 (1),* 19–23.

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

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove: Lawrence Erlbaum.

Gathercole, S. E., & Pickering, S. J. (2000). Assessment of working memory in six- and seven-year old children. *Journal of Educational Psychology, 92,* 377–390.

Gers, F. A., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation, 12 (10),* 2451–2471.

Gers, F. A., & Schmidhuber, J. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks, 12 (6),* 1333–1340.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. In E. Clark (ed.), *Proceedings of the 24th Annual Child Language Research Forum*, pp. 124–138. Stanford, CA: Center for the Study of Language and Information (CSLI).

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

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish–English bilinguals' acquisition of English word

stress patterns. *Bilingualism: Language and Cognition, 7,* 207–226.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science, 14 (1),* 31–38.

Harley, B. (1979). French gender 'rules' in the speech of English-dominant, French-dominant and monolingual French-speaking children. *Working Papers in Bilingualism, 19,* 129–156.

Hernandez, A., & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin, 133 (4),* 638–650.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences, 9 (5),* 220–225.

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

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1 (2),* 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59,* 30–66.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in second language acquisition. In C. J. Doughty & M. H. Long (eds.), *Handbook of second language acquisition*, pp. 539–588. Oxford: Blackwell.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* B47–B57.

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.

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

Kane, M. J., & Engle, R. W. (2003). Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132,* 47–70.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126 (3),* 278–287.

Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language, 44,* 250–273.

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,* 279–314.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

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,* 447–464.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks, 17 (8–9),* 1345–1362.

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

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

Long, M. (2005). Problems with supposed counter-evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics in Language Teaching* (*IRAL*)*, 43 (4),* 287–316.

Lyster, R. (2006). Predictability in French gender attribution: A corpus analysis. *Journal of French Language Studies, 16,* 69–92.

MacWhinney, B. (2006). Emergent fossilization. In Z. Han & T. Odlin (eds.), *Studies of fossilization in second language acquisition*, pp. 134–156. Clevedon: Multilingual Matters.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language, 28,* 255–277.

Matthews, C. A. (1999). Connectionism and French gender attribution: Sokolik and Smith re-visited. *Second Language Research, 15,* 412–427.

Mayberry, R. I. (1993). First-language acquisition after childhood differs from second-language acquisition: The case of American Sign Language. *Journal of Speech and Hearing Research, 36,* 51–68.

Mayberry, R. I., & Lock, E. (2003). Age constraints on first versus second language acquisition: Evidence for linguistic plasticity and epigenesis. *Brain and Language, 87,* 369–383.

Mayberry, R. I., Lock, E., & Kazmi, H. (2002). Linguistic ability and early language exposure. *Nature, 417,* 38.

Metsala, J. L. (1999). Young children's phonological awareness and non-word repetition as a function of vocabulary development. *Journal of Educational Psychology, 91,* 3–19

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Monner, D., & Reggia, J. A. (2012). A generalized LSTM-like training algorithm for second-order recurrent neural networks. *Neural Networks, 25,* 77–83.

Morford, J. P., & Carlson, M. L. (2011). Sign perception and recognition in non-native signers of ASL. *Language Learning and Development, 7,* 149–168.

Munakata, Y. (2004). Computational cognitive neuroscience of early memory development. *Developmental Review, 24,* 133–153.

New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. *Actes de la Conférence Traitement Automatique*

des Langues Naturelles (TALN 2006), Louvain, Belgique. http://www.lexique.org (retrieved March 18, 2009).

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

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

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and nasal ganglia. *Neural Computation, 18 (2),* 283–328.

Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Amsterdam: John Benjamins.

Petanjek, Z., Judaš, M., Kostović, I., & Uylings, H. B. (2008). Lifespan alterations of basal dendritic trees of pyramidal neurons in the human prefrontal cortex: A layer-specific pattern. *Cerebral Cortex, 18,* 915–929.

Pulvermüller, F., & Schumann, J. H. (1994). Neurobiological mechanisms of language acquisition. *Language Learning, 44 (4),* 681–734.

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

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

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 morphosyntactic language aspects. *Cognition, 93,* B97–B108.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing, 12,* 44–49.

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.), *Attention & Performance XXI: Processes of Change in Brain and Cognitive Development*, pp. 585–612. Oxford: Oxford University Press.

Sokolik, M. E., & Smith, M. E. (1992). Assignment of gender to French nouns in primary and secondary language: A connectionist model. *Second Language Research, 8,* 39–58.

Surridge, M. E. (1989). Le facteur sémantique dans l'attribution du genre aux inanimés en français. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique, 34,* 19–44.

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* (*IRAL*)*, 31,* 77–95.

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

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

Thomas, M. S. C. (2009). Competition as a mechanism for producing sensitive periods in connectionist models of development. In J. Mayor, N. Ruh & K. Plunkett (eds.), *Progress in Neural Processing 18: Proceedings of the Eleventh Neural Computation and Psychology Workshop*, pp. 349–360. Singapore: World Scientific.

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

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28 (1),* 1–30.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92,* 231–270.

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

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438,* 500–503.

Wikipedia: The free encyclopedia. (2011). FL: Wikimedia Foundation, Inc. http://www.wikipedia.org (retrieved January 15, 2011).

Wilson, M., & Emmorey, K. (1997). A visuospatial "phonological loop" in working memory: Evidence from American Sign Language. *Memory and Cognition, 25,* 313–320.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation, 15 (2),* 441–454.

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