Evolving Connectionist Models to Capture Population Variability across Language Development: Modeling Children's Past Tense Formation

Maitrei Kohli*
University of London
Birkbeck College
Department of Computer Science & Information Systems
maitrei@dcs.bbk.ac.uk

George D. Magoulas
University of London
Birkbeck College
Department of Computer Science & Information Systems
Knowledge Lab
gmagoulas@dcs.bbk.ac.uk

Michael S. C. Thomas
University of London
Birkbeck College
Department of Psychological Sciences
m.thomas@psychology.bbk.ac.uk

Abstract Children's acquisition of the English past tense has been widely studied as a testing ground for theories of language development, mostly because it comprises a set of quasi-regular mappings. English verbs are of two types: regular verbs, which form their past tense based on a productive rule, and irregular verbs, which form their past tenses through exceptions to that rule. Although many connectionist models exist for capturing language development, few consider individual differences. In this article, we explore the use of populations of artificial neural networks (ANNs) that evolve according to behavioral genetics principles in order to create computational models capable of capturing the population variability exhibited by children in acquiring English past tense verbs. Literature in the field of behavioral genetics views variability in children's learning in terms of genetic and environmental influences. In our model, the effects of genetic influences are simulated through variations in parameters controlling computational properties of ANNs, and the effects of environmental influences are simulated via a filter applied to the training set. This filter alters the quality of information available to the artificial learning system and creates a unique subsample of the training set for each simulated individual. Our approach uses a population of twins to disentangle genetic and environmental influences on past tense performance and to capture the wide range of variability exhibited by children as they learn English past tenses. We use a novel technique to create the population of ANN twins based on the biological processes of meiosis and fertilization. This approach allows modeling of both individual differences and development (within the lifespan of an individual) in a single framework. Finally, our approach permits the application of selection on developmental performance on the quasi-regular task across generations. Setting individual differences within an evolutionary framework is an important and novel contribution of our work. We present an experimental evaluation of this model, focusing on individual differences in performance. The experiments led to several novel findings, including: divergence of population attributes during selection to favor regular verbs, irregular verbs, or both; evidence of canalization, analogous to Waddington's developmental epigenetic landscape, once selection starts targeting a particular aspect of the task domain; and the limiting effect on the power of selection in the face of stochastic selection (roulette wheel),

Keywords
Neural network based modeling, behavioral genetics, class imbalance, quasi-regular mappings, English past tense, evolution, individual differences, development, genetic computing

* Corresponding author.

sexual reproduction, and a variable learning environment for each individual. Most notably, the heritability of traits showed an inverse relationship to optimization. Selected traits show lower heritability as the genetic variation of the population reduces. The simulations demonstrate the viability of linking concepts such as heritability of individual differences, cognitive development, and selection over generations within a single computational framework.

1 Introduction

In artificial life systems, interactions between evolution and learning have attracted considerable attention in the literature, and several computational models have been proposed to investigate the way evolution affects learning. In this work, we focus on language learning, an area where computational models have made several contributions towards a better understanding of language development and evolution [15, 28, 43].

Language learning is considered one of the most complex tasks. Nevertheless, most children acquire language naturally, effortlessly, and quickly compared to other areas of cognitive development. Language is like the majority of complex systems that exist in nature and that empirically exhibit hierarchical structure [42].

Two opposing theories of language acquisition dominate the linguistic and psycholinguistic communities (refer to [55] for a review). The *nativist* approach, proposed by Chomsky [5, 6] and promoted by Pinker, claims that the linguistic capability at least with respect to grammar is innate; therefore, certain linguistic universals are inherited by language learners, encoded in the genome by some prior process of evolutionary selection; only the established parameters need a little tweaking in order for language to be fully acquired [30].

The second view is the *emergentist* approach. It asserts that language emerges as a result of various challenging constraints, which are all consistent with other general cognitive abilities. No dedicated provisions for universal grammar are required. According to this view, the complexity of language emerges from the exposure of relatively simple developmental processes to a massive and complex environment [24, 25].

Computational models provide an insight into language acquisition processes and the nativist-versus-emergentist debate. Artificial neural networks or connectionist networks offer an intuitive framework in which empirical phenomena in language acquisition can be explained by virtue of interactions between a language-learning system that incorporates general properties of computations in the brain and statistical properties of the linguistic environment to which it has been exposed [17]. Computational models have been extensively applied to investigate the mechanisms of language development, including simulating early phonological development, lexical segmentation, vocabulary development, the acquisition of pronouns, the development of inflectional morphology, syntax comprehension, syntax production, metaphor comprehension, and reading [50] (for reviews, see [4, 26]).

One particular focus of research has been the field of inflectional morphology, which considers the alteration of the phonological forms of words to change their meaning (such as tense for verbs and number for nouns). Within this field, the acquisition of the English past tense has drawn a great deal of attention, under the assumption that it taps the main cognitive processes involved in the acquisition and use of morphological knowledge [17]. Children's acquisition of the English past tense has been the focus of a great deal of empirical research, mostly due to its *quasi-regular* mappings [50].

Quasi-regular domains are interesting because of the presence of systematic input-output mappings with a minority of exceptions [50]. The majority of English verbs, viz. regular ones, form their past tense by following a rule for stem suffixation, also referred to as the *+ed* rule. This rule allows for three possible phonological suffixes [18]—/*d*/ (e.g., raise–raised); /*t*/ (e.g., clap–clapped); /*ed*/ (e.g., visit–visited). The rule is applied productively to novel forms (e.g., wug–wugged). However, there are around 200 irregular verbs that form their past tenses by exceptions to the aforementioned

rule, (e.g., go–went; eat–ate; ring–rang, hit–hit). Although irregular verbs do not follow the productive rule, there are some irregular verbs that share characteristics of the regular verbs. For instance, many irregular verbs have regular endings, /*d*/ or /*t*/, but with either a reduction of the vowel (e.g., say–said; do–did) or the deletion of a stem consonant (e.g., has–had; make–made) [23]. This overlap between regular and irregular verbs adds to the complexity of the task domain. (See the mapping between written and spoken forms of English for another example of a quasi-regular domain within language [33]).

Due to this dual and fuzzy nature, there is an ongoing debate in the field of language development about the processing structures necessary to acquire the domain, and whether it has a duality to reflect the structure of the problem space (refer to [51] for a review). Is it necessary for the system to contain a prior processing assumption that the domain includes a productive rule, requiring symbolic computational structures? If so, how are the exception cases accommodated? Or can productivity emerge from undifferentiated associative mechanisms exposed to quasi-regular domains?

There are two main theories. The first is a dual-route account, proposed by Pinker [29], according to which two separate mechanisms are involved in learning the mappings: a rule-based system for learning regular mappings, and a rote-memory system, which supports the irregular mappings. Rumelhart and McClelland [41] challenged this dual-mode model by proposing a model based on the principles of parallel distributed processing. Their alternative model demonstrated that a two-layered feedforward neural network could learn mappings between phonological representations of verbs and their corresponding past tense forms, both regular and irregular, as well as demonstrating productivity of the rule to novel verbs. This model, though extremely influential, had several drawbacks (refer to [18] for details).

The model was based on the backpropagation algorithm and inspired many subsequent connectionist models of acquisition of inflections ([8, 10, 38, 39], to name a few). Subsequent connectionist models addressed many of the drawbacks of the initial model. For example, Plunkett and Marchman [39] took the main idea from Rumelhart and McClelland's model and modified it into three-layered feedforward architecture with more realistic phonological representations.

The line of research inspired by Rumelhart and McClelland employed artificial neural networks to simulate a wide range of phenomena related to past tense acquisition. However, the majority of this work was concerned with capturing the developmental profile of the average child. Recently artificial neural network models have been extended to explore causal factors of atypical development, for example, in the cases of developmental language disorder and the Williams syndrome [19]. To our knowledge, very little work has been concerned with capturing the wide range of variability that typically developing children exhibit in acquiring this aspect of language. Thomas, Forrester, and Ronald [50] modeled the effects of socioeconomic status (SES) on language development, combining development and individual differences in a single framework. The key innovation of this model was that it addressed individual differences arising from variations in SES of the families in which children are raised. Such variation was simulated as a modulation of the language environment with which children interact, but importantly, captured against a background of variation in the computational power of each individual's language-learning systems.

Recently, two innovations in this line of research have raised interesting questions of relevance to research in artificial life and evolutionary computation. The first innovation is the application of past tense modeling to individual differences between children with respect to their origin in *genetic and environmental factors*. For example, to some extent language delay runs in families, implying a heritable component, while differences in SES—a proxy measure of the quality of the environment—also explain some of the variation in language development [50]. The second innovation is the use of *multi-scale modeling* to reconcile data from multiple levels of description, including genetic, neural structure, cognitive processes, behavior, and the environment. The operation of genetic factors on behavior is captured as the outcome of an extended development process involving interaction with a structured learning environment. This framework, using past tense as an illustrative cognitive domain, has for example explored the relationship of statistical gene-behavior associations (as reported in genome-wide association studies) to developmental mechanisms. The specification of a

genetic level in the model allows simulation of identical and fraternal twins, thereby simulating the kinds of twin study designs used to assess the heritability of high-level behavior [48].

Genetic algorithms were initially proposed to simulate biological evolution, and in that context they have found several applications in artificial life. However, they quickly gained prominence as global search methods and have been extensively applied for optimization, where selection across generations aims to improve the performance of learning systems on a target task. By contrast, the existing multi-scale models take the presence of genetic variation as a starting point. This raises the following questions: Where does the existing genetic variation in populations come from? How does this variation respond to the operation of selection? How do measures of heritability alter across generation's through the operation of selection? What are the implications of using a quasi-regular domain as the target problem for optimization? What parts of the problem domain are optimized across generations, and what factors determine this? The relationship between learning and genetic inheritance is a crucial concern in the study of artificial life [27, 28].

To address these questions, in this work, we build on our previous investigation that combined concepts of behavioral genetics with the idea of parametrically diverse populations of learning systems, where genes (representing intrinsic factors) and environment (expressed via training data sets) interact throughout development to shape differences in individual classifier behaviors [22]. We extend the framework to an evolutionary context by introducing selection in the populations' optimization process across generations, focusing on learning a particular task: the English past tense. The use of selection on performance in a quasi-regular task and the resulting findings make our English past tense acquisition model novel and different from others proposed in the literature. In this context, we present our synergistic approach to capture population variability stemming from genetic and environmental influences and to analyze effects of selection on behavioral outcomes.

This approach not only captures the heterogeneity observed in acquiring a new ability, but also helps in understanding how the quality of an environment interacts with intrinsic constraints, leading to an individual's overt behavior. It shows, for example, the different behaviors emerging due to interaction of quality of the training set with good (or poor) learning rate (i.e., ability to learn, similar to neuroplasticity) and good (or poor) numbers of hidden units (i.e., capacity to learn, somewhat similar to neurogenesis). It also highlights how applying selection results in changes in overt behavior across generations.

In behavioral genetics, factors affecting language development are attributed to genetic and environmental influences [20]. To model genetic influences, we encode variation in neuro-computational parameters of ANNs that modulates overall learning efficiency. These parameters relate to how a network (individual) is built (the number of hidden units), its processing dynamics (slope of the logistic function within processing units), and how it adapts (learning rate). The effects of environmental influences are simulated via a filter applied to the training set. This filter alters the quality of information available to the learning system. One factor identified to correlate with variations in language and cognitive development is SES, in terms of parent income and education levels. Although this measure is a proxy for the potentially multiple causal pathways by which environmental variation influences development, one line of evidence supports the view that SES modulates levels of cognitive stimulation: Children in lower-SES families experience substantially less language input and also a narrower variety of words and sentence structures [50]. When implemented as a filter, the result is the creation of a unique subsample of the training set for each simulated individual, based on their SES.

Although intrinsic and extrinsic parameters vary independently in this formulation, gene-environment interactions can occur. According to principles of behavioral genetics, both genes (intrinsic factors) and environment (training data sets) interact throughout development to shape differences in individual behaviors (performance) [37]. Here, connectionist networks contain a range of parameters that can increase or decrease the ability and/or capacity of the network to acquire a new ability, but the structure, or the quality, of the environment affects the way these intrinsic parameters behave. For example, within a modeling context, a certain number of ANN hidden units might be highly beneficial for a specific condition of the environment (say, the number of training

examples available), but if these conditions were to change drastically (say, a large expansion of the training set), the same number of hidden units might not be able to accommodate the change. Thus, the system's performance would alter.

Apart from having genetic and environmental variation, our model also incorporates *selection* and its effects. As is shown later in this article, applying selection on performance on the English past tense problem leads to two novel findings: (i) Selection can stochastically target different aspects of a quasi-regular task, depending on the initial conditions, potentially producing divergent populations. This in turn results in the emergence of different and varied behavioral (performance) patterns, while still optimizing on the target task. (ii) The amount of performance variation explained by genetic similarity, the so-called heritability metric [20], plays an important role in identifying which aspect of this quasi-regular task is being targeted by selection.

The rest of this article is organized as follows. First, we give an overview of the proposed hybrid computational model, combining neural networks and evolution, and its inspiration in behavioral genetics. We then explain the methodology for the implementation of the model and the past tense data set used in the simulations. Finally, we present the results and discuss their implications.

2 Behavioral-Genetics-Inspired Hybrid Computational Model

Behavioral genetics (BG) is a field of study that examines the role of genetics in individual differences in human behavior. Behavior is the most complex phenotype in that it reflects the functioning of the whole organism; it is dynamic and changes in response to the environment [35]. This field is concerned with the study of individual differences, that is, knowing what factors make individuals within a group differ from one another. It also estimates the importance of genetic and environmental factors that cause individual differences. Thus, the behavior (phenotype) is the result of genetic factors together with environmental factors.

Twin studies are the workhorse for behavioral genetic research. The twin design provides a quasi-experimental scenario triggered to measure respective contributions of nature and nurture to individual differences [20]. Twins are matched for age, family, and other social influences. They are either genetically identical (genetic relatedness of 1.0 for *monozygotic* (MZ), or identical, twins) or as similar as siblings (genetic relatedness of 0.5 for *dizygotic* (DZ), or fraternal, twins) and, to an approximation, share the same environment (applicable for both MZ and DZ twins, based on the *equal-environments assumption*) [37]. The difference between MZ and DZ twin pairs in their similarity in performance, along with assumptions about their similarity of environment, allows inferences to be drawn about the influence of genetic relatedness on behavior [36]. If MZ twins are more similar in a trait than DZ twins—and the environment plays an equivalent role in making each pair similar—then the greater similarity of MZ twins must stem from their greater genetic similarity.

The extent to which greater genetic similarity predicts greater trait similarity can be used to derive a measure of heritability [12]. Formally, the heritability statistic is defined as the proportion of observed or phenotypic variance that can be explained by genetic variance [20]. There has been increasing acceptance that in humans, many high-level behaviors show marked heritability [20], a finding that would have been surprising to many researchers in the latter part of the 20th century. For our purposes, the measure of heritability derived from the twin method is useful because it is a single metric derived from performance that scales however large the number of parameters by which the heritable influence is delivered in the underlying mechanism. Thousands of gene variants may influence thousands of parameters that combine in complex ways to shape behavior, but by measuring trait similarity in twins, a single measure of genetic influence is available.

In twin designs, environmental influences are defined as being of two types, shared (or between-family) and non-shared (or unique and within-family). Shared environmental influences are those that serve to make members of a family similar to each other and different from members of other families. Shared environmental influences include family structure, socioeconomic status, and parental

education, to name a few [34]. By contrast, non-shared environmental influences are factors that serve to make individuals different from one another. These environmental influences do not operate on a family-by-family basis, but rather on an individual-by-individual basis. Examples include peer groups, perinatal traumas, and parental treatment [34, 37]. Measurement error in twin designs also contributes to the non-shared environment.

To measure heritability and the proportions of variance explained by shared and non-shared environments, we use a technique based on Falconer's equations [12] as described in [22]. Linear algebra is used to derive estimates of heritability. Broadly, since DZ twins are half as genetically similar (on average) as MZ twins, the difference in the correlation (Pearson's formula) between MZ and DZ twins shows about half the genetic influence on behavior; doubling the difference in correlations between MZ and DZ twins gives an estimate of heritability.

Our base model, prior to implementing sources of variation and the use of twins, was inspired by that proposed by Plunkett and Marchman [39]. They suggested that both the regular and the exception verbs could be acquired by an otherwise undifferentiated three-layer backpropagation network, trained to associate representations of the phonological form of each verb stem to a similar representation of its past tense. This became our base model; we introduced the sources of variations, relying on the Rprop algorithm [40] for training; and then introduced selection across generations of twin pairs. Table 1 provides a high-level description of the BG-inspired past tense model, and each step in the table is discussed in subsections below [21].

2.1 Simulating Variations Due to Genetic Influences

Artificial neural networks depend on a range of parameters that increase or decrease their ability to acquire a new task. In the current instantiation, our approach employed three free parameters to

Table 1. High-level description of BG-inspired English past tense model.

1. Simulate variations in genetic influences.
    - Encode neurocomputational parameters into genome.
    - Calibrate range of variation of each of these parameters.
2. Simulate variations in environmental influences.
    - Apply SES-based filter to data set to generate unique training subset for each twin pair.
3. Generate initial population of ANN twins, $G(0)$, such that each individual is an ANN characterized by its own genetic and environmental influences. Set $i = 0$.
4. REPEAT
    (a) *Train* each individual (ANN twin) using some local search mechanism.
    (b) Evaluate *fitness* of each individual ANN according to training performance result for regular verbs, irregular verbs, and combined performance. Also calculate heritability by comparing similarity of identical and fraternal twin pairs.
    (c) *Select* parents from $G(i)$ based on their fitness on combined (overall) performance.
    (d) Apply search operators to parents to produce offspring, which form $G(i + 1)$.
5. UNTIL termination criterion is met.

constrain the learning abilities of ANNs, which were assumed to be under genetic influence. The first two parameters—the number of hidden units and the learning rate (or the initial learning rate of Rprop)—have been used in almost all applications involving ANNs. These are formational parameters, since the former corresponds to how the network is built and thus relates to a network's capacity to learn, whereas the latter governs how networks adapt and hence provides a network with the ability to learn. These parameters would thus be influential in distinguishing between fast and slow learners.

We also used another parameter, the slope, or steepness, of the logistic threshold function within the artificial neurons. This corresponds to the activation dynamics acting within each network. Modulation of this parameter leads to steeper or shallower slopes in the threshold function. A shallow slope negates the opportunity of a processing unit to make large output changes in response to small changes in input; a steep slope ultimately leads to very sensitive but binary response characteristics subject to entrenchment effects. Therefore, too shallow or too steep values of this parameter will hinder the learning process [31, 49].

In order to constrain learning, these properties were encoded into a genome. The genome was the measure of the base composition of an individual. In other words, it served as a set of instructions about how to form an organism of a particular species or group. Encoding parameters in the genome allowed the individuals in a population to have a different genotype, that is, different values of each of the free parameters but from within the same fixed range. It thus led to variability in a population by giving each network a different ability or capacity to learn new tasks. As described in Section 3.2, the artificial genome was composed of binary genes, where the values of several genes were used to determine the value of each parameter, a scheme called polygenic coding [48].

2.2 Simulating Variations Due to Environmental Influences

Variations in shared environmental influences were simulated through variations in the environmental factor (EF), which implemented the kind of variation in the richness of language environment associated with differences in the SES of the families in which children are raised. SES effects can be implemented in three main ways: by manipulating the quality and quantity of the information available, by altering the motivation of the learner to utilize the available information through differences in reward and punishment schedules, or by manipulating the computational properties of learning systems (as, for instance, differences in stress levels or diet might influence brain processes in children) [50].

For this work, we focus on EF as a manipulation of the quality and quantity of information available to the learners. We assumed that, in principle, there is a perfect environment, or full training set, available to any learner. This comprised all of the verbs available in the language and their accepted past tense forms. We then modeled an individual's EF by a number selected at random from the range 0.6–1.0. This gives a probability that any given verb in the full training set would be included in that individual's training set. The range 0.6–1.0 defines the range of variation of EF in the population, and ensures that all individuals are exposed to more than half of the past tense domain. Twin pairs raised in the same family were exposed to the same training set, such that EF would lead to effects of shared environment. The variance in performance that cannot be inferred from shared environment is representative of effects of unique or non-shared environmental influences. In the absence of measurement error (because none was added), unique environment effects arose from stochastic factors such as the initial weights of ANNs or the random order of presentation of training items.

The learning speed and fast convergence of many feedforward neural networks depend to some extent on their initial values of weights and biases [45, 56]. For this reason, in our approach, initial values of weights were one way to capture unique environments. The initialization method used in this work is similar to that proposed by [3] and uses the interval $\left[-\frac{a}{\sqrt{d_{\text{in}}}}, +\frac{a}{\sqrt{d_{\text{in}}}}\right]$, wherein a is chosen in such a way that the weight variance corresponds to the points of maximum curvature of the

activation function (its value is 2.38 for the standard sigmoid function [45]), and d_{in} is the fan-in (the total number of inputs) of a neuron in the network.

3 Model Implementation Methodology to Capture Individual Differences

Using the concepts explained in the previous sections, we built a model to learn English past tenses and also capture individual differences in performance. The starting point of this work was to estimate the proportions of the variance contributed by variances in structural parameters (genes), training set (shared environment), and initial weights (non-shared environment). The methodology adopted can be summarized as follows.

3.1 Design ANNs

The first step was to design ANNs incorporating neurocomputational parameters that constrained their ability to learn. We selected three free parameters, each of which corresponded to how the network is built: the number of hidden units; the slope of the logistic activation function; and how it adapts, the initial learning rate of Rprop.

3.2 Calibrate Range of Variation

In the second step, the range of variation of each of these parameters was calibrated to avoid the presence of genes in the population that produced networks with no learning ability. This established the range of variation in the population prior to the operation of selection. To this end, we began with random values for all parameters and trained 100 neural networks for 1000 epochs while varying the values, in steps of 5 for hidden units and 0.01 otherwise, for each of these parameters individually. The calibration process was carried out for all parameters, until values were identified beyond which the learning failed (less than 20% accuracy—lower bound), as well as the values that resulted in successful learning (80% accuracy or more—upper bound for range). This method provided a range of parameter values from poor up to very good performance. These values were then encoded in the artificial genome. Encoding the parameters within a fixed range allowed variation in the genome between members of population, which then produced variations in computational properties. The range of variation of the parameter values served as the upper and the lower bound used for converting the genotype (encoded values) into its corresponding phenotype (real values). For the encoding, we used a binary representation, whereby each gene had two variants (alleles), with 10 bits per parameter, split into two chromosomes. Paired chromosomes allowed sexual reproduction to be simulated. In turn, this permitted networks to be generated with different degrees of relatedness, either MZ or DZ twins, and enabled utilization of the twin design to estimate heritability. The parameters and their range of variation are given in Table 2.

The genotype-phenotype mapping was implemented as follows: This step involved decoding the binary representation of the population into vectors of real values. The genotypes are the concatenated binary strings of given length and are decoded into real-valued phenotypes over a specified interval using standard binary coding [54]. There are a number of ways in which binary-to-real conversions can be done; in this work we make use of the Matlab genetic algorithm Toolbox (http://codem.group.shef.ac.uk/index.php/ga-toolbox) library function called *bs2rv*, which has a decoding matrix. This matrix has the following parameters to accomplish the binary-to-real conversion: length of each binary string (*len*); lower and upper bounds for each encoded gene (neurocomputational parameter) (*lb* and *ub*); type of encoding (binary or Gray) (*code*); type of scaling to be used for each string, (arithmetic or logarithmic) (*scale*); and finally whether or not to include the lower and/or upper bound in the representation range (*lbin* and *ubin*).

Table 2. Genome representing ANN parameters and their ranges.

Parameter	Range of variation
No. of hidden units	10–500
Initial learning rate	0.07–0.1
Slope of logistic	0.0625–4.0

3.3 Breed the Population

The next step concerned creating the population of ANN twins using the genome. We simulated the biological processes of meiosis and fertilization to create 50 pairs of MZ and 50 pairs of DZ twins (refer to [7] for details about biological meiosis and fertilization). Table 3 presents the algorithm used for creating the population of ANN twins.

We began this process by creating a population of n members with random binary genomes. These n members were then split into two groups of size n/2, representing fathers and mothers. Next, the genome of each individual was split into two equal halves, resulting in two chromosomes per individual. Each chromosome contained half the information to code for each parameter. Crossover was applied m times on these chromosomes. Each crossover resulted in two sperms or eggs. Sperms and eggs were then combined to create offspring employing positional recombination, such that for each parameter, half the encoded information came from the sperm and other half from the egg. Thus, every crossover and fertilization led to two offspring, resulting in a total of $2m$ possible offspring. Mutation was not performed, for simplicity. The mutation rate is extremely low in humans; mutation tends to reduce the average genetic similarity between siblings below 50%, violating the assumptions of the twin design. The offspring genotypes were converted to phenotypes using the parameter values given in Table 2.

Although in biology meiosis creates two sperms or two eggs from the crossover operation, the likelihood of both of the pair ending up in organisms is very small. If this happened, the mean

Table 3. Meiosis- and fertilization-based method for creating a population of ANN twins.

1. Generate initial population $G(0)$ of n members at random.
2. Split the population members into two groups of size $n/2$ representing fathers and mothers.
3. REPEAT
   (a) For each parent, split genome into two equal halves, resulting in two chromosomes per individual, such that each chromosome carries half the information for each encoded parameter.
   (b) Apply crossover m times on each chromosome pair, every crossover resulting in either two sperms or two eggs.
   (c) Combine the sperms and eggs using positional recombination such that half of the encoded genetic information comes from sperm and other half from egg, resulting in $2m$ possible offspring.
   (d) Verify the genetic similarity between twin pairs, and accordingly choose MZ and DZ twin pairs, picking only one offspring per crossover.
4. UNTIL population of desired size n is obtained.

genetic similarity of the population would start to be affected. We therefore only selected one of the pair of sperms or eggs generated by the crossover to generate offspring, while the other was discarded.

To verify the genetic similarity between twin pairs, we used the Hamming distance metric to assess the similarity amongst offspring. Let us assume that $2m = 6$, and crossover is applied three times, so that *xover1* results in offspring *(o1, o2), xover2* results in *(o3, o4)*, and *xover3* results in offspring *(o5, o6)*. First, we randomly pick any one offspring out of the possible six; let us assume that is *o1*. For the reasons explained above, we discard *o1*'s corresponding offspring, *o2*. Next, the similarity of *o1* with the remaining four offspring is checked using the Hamming distance formula. The offspring that is at most 50% similar is chosen as *o1*'s corresponding DZ twin; assume *o4*. This implies that *(o1, o4)* form a pair of DZ twins. Subsequent to *o4*'s selection, its corresponding twin from crossover, *o3,* is discarded. Now, out of the remaining two twins, one is chosen randomly and replicated, and the resulting two constitute the MZ twin pair.

This process was repeated until we achieved the desired population size. When simulating multiple generations, the internal similarity of the gene pool should not be increased by inbreeding. If related individuals were to breed with each other, the average similarity between individuals would increase over the generations. For this reason, we separated twin pairs into breeding and non-breeding populations, and only bred from the breeding twin of each pair, while the non-breeding twin was available to compute heritability. Breeding therefore always took place between unrelated individuals, preserving the mean genetic similarity within populations across generations.

3.4 Apply Variation in the Environment

An individual's EF was modeled by a randomly chosen number between 0.6 and 1.0. This gave a probability that any given pattern in the full training set was included in that individual's training set. This filter was applied at each generation to create unique training subsets for all members of the population in that generation. The range 0.6–1.0 defined the range of variation of environmental quality, and ensured that all individuals were exposed to more than half of the training data set. In accord with the *equal-environments assumption* [36], twin pairs raised in the same family were assigned the same training subset.

3.5 Train ANN & Assess their Performance

The population of twin ANNs was trained on the past tense data set using the Rprop algorithm [1, 40]. Rprop is a training algorithm for supervised learning in feedforward neural networks. It takes into account only the sign of the partial derivative over all patterns and maintains a separate learning rate for each weight that is adapted during training. It is a superior algorithm in terms of convergence speed, accuracy, and robustness with respect to the training parameters and therefore is a very popular choice for the training of multilayer feedforward neural networks in various applications and is included in several packages (e.g., R and Matlab). Rprop works as follows: Every time t at which the partial derivative of the error E with respect to a weight $w_{ij}^{(t)}$ changes its sign, which indicates that the last weight update was too big and the algorithm has jumped over the local minimum, the update value $\Delta_{ij}(t)$ is decreased by multiplying with a user defined factor $\eta-$. If the derivative retains its sign, the update value is slightly increased by multiplying with a factor η^{+}in order to accelerate convergence in shallow regions. Additionally, in case of a change in sign, there should be no adaptation in the succeeding learning step. In practice this can be achieved by setting the derivative $\frac{\partial E}{\partial w_{ij}}(t-1) = 0$ in the adaptation rule, which leads to no update. Lastly, the update values are calculated and all the weights are updated after the gradient information of the whole pattern set is computed. For a detailed explanation of the Rprop algorithm and of its variants, the reader is referred to section 2 of Anastasiadis et al. [1] and to [40].

In this work, the performance was assessed on the full training set, as well as, on another novel dataset that was created to test the generalization ability of the networks (see below). The continuous outputs produced by networks were converted to binary by applying a threshold. Then the

performance was assessed using recognition accuracy based on Hamming distance as explained below in Table 4.

3.6 Apply Selection

Based on the performance of the networks on the full training set, members were selected from the breeding population to produce offspring to populate the next generation. To this end, a stochastic selection metric, the standard roulette wheel, was applied at the end of training (1000 epochs) (for details about roulette wheel selection refer to http://www.edc.ncl.ac.uk/highlight/rhjanuary2007g02.php). An important and novel aspect of our approach to the past tense acquisition problem was the combination of the roulette wheel method with the sexual reproduction method. The selected members entered the breeding pool and then bred with a randomly chosen member from that pool. After selection, only the offspring form the next generation of the population—parents (or members of the previous or breeding population)—were discarded. Despite the use of sexual reproduction, we did not include gender effects in the method or its outcomes.

As a result of sexual reproduction, the best properties of parents did not always get transferred to offspring. This is for two reasons: (i) An individual (parent) can only pass one copy of each gene to its offspring. Therefore, there is an equal chance that either a maternally inherited gene or a paternally inherited gene will get transmitted to the offspring [20]. Since, after selection for the breeding pool, the members breed randomly, the best properties do not always get transferred effectively, since the advantageous gene may not be inherited. (ii) Inherited traits must combine with environmental conditions during the offspring's development to produce its behavioral phenotype. Inheritance of strong learning abilities may not be associated with good performance if the offspring is exposed to an impoverished learning environment. It is the combination of inherited liabilities and environmental conditions that makes each individual unique [16].

3.7 Repeat

The entire process was iterated until ANN parameters did not markedly change across generations or performance started to converge, that is, the learning error reached a small value.

4 English Past Tense Data Set

The data set was based on the "phone" vocabulary from the Plunkett and Marchman [39] past tense model. The past tense domain was modeled by an artificial language created to capture many of the important aspects of the English language, while retaining greater experimental control over the similarity structure of the domain [39]. Artificial verbs were monosyllabic phoneme strings that followed one of three templates—CCV, VCC, and CVC, wherein C is a consonant and V is a vowel. There were 508 verbs in the data set. Each verb had three phonemes—initial, middle, and final. The phonemes were represented over 19 articulation binary features encoding English phonology (e.g., voicing, tongue position, closed or open lips) [14]. A network thus had $3 \times 19 = 57$ input units and $3 \times 19 + 5 = 62$ units at the output. The extra five units in the output layer were used for representing the affix for regular verbs in binary format. As an example, consider the word *bag* and its past tense *bagged*. It would be represented as: first phoneme /b/ `0100111000000000000`; second phoneme /æ/ `1011101000001001000`; and third phoneme /g/ `0100100010000000000`. So, the word *bag* becomes

```
0100111000000000000  1011101000001001000  0100100010000000000.
```

Its past tense, *bagged* is made by adding an extra 5-bit suffix to the original verb. The suffix, /ed/ in our example, is represented as 00101. A detailed schematic of the phonological coding scheme can be found in [46, Figure 3], available here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6075940/figure/f0015/.

Table 4. Recognition-accuracy-based performance calculation algorithm.

Input: Actual output of network, Y_n

Desired output, Y_d

Output: Performance accuracy, A

Variables: $I \rightarrow$ total number of patterns Y_n

$J \rightarrow$ total number of patterns in Y_d

$P_i \rightarrow$ a pattern in Y_n, where $i < I$

$P_j \rightarrow$ a pattern in Y_d, where $j < J$

$h_{\text{dist}} \rightarrow$ Hamming distance between phonemes of P_i and P_j

$h_{\text{allm}} \rightarrow$ Hamming distance between allomorphs (or the last 5 bits) of P_i and P_j

Match; corr; err

```
1.  initialize Match = false; corr = 0; err = 0
2.  for (i = 1: I; i < I; i + +)   Repeat
3.    Split Pi into three phonemes and allomorph
4.    for (j = 1: J; j < J; j + +) do
5.      Split Pj into three phonemes and allomorph
6.      Calculate h_dist between corresponding phonemes of Pi and Pj
7.      If h_dist < preset threshold (for all three phonemes) do
8.        Calculate h_allm between respective allmorphs
9.        If h_allm == 0, do
10.         corr = corr + 1;
11.         Match = true;
12.         Break;
13.       Else
14.         err = err + 1;
15.       end
16.     end
17.   end
18.   If (j == J AND Match == false) do
19.     err = err + 1;
20.   end
21. end
22. A = (Corr/I) * 100;
23.   Return A
```

In the training data set, there were 410 regular and 98 irregular verbs. Regular verbs followed the "add -ed" rule, with three different phonetic versions of the -ed inflection, */ed/*, */d/*, and */t/*. There were three types of irregular verbs: vowel change, no change, and arbitrary. In the data set, out of 410 regulars, there were 271 */ed/* verbs, 90 */d/* verbs, 49 */t/* verbs. As this was an unbalanced data set, generating a classifier was challenging, as the classifiers tend to map (label) every pattern with the majority class.

A second data set was also created to assess the generalization performance of the model. It measured the degree to which an ANN could reproduce in the output layer properly inflected novel items presented in the input, according to the regular rule. The generalization set comprised 410 novel verbs, each of which shared two phonemes with one of the regular verbs in the training set, for example *wug–wugged* [18, 47]. Three different degrees of similarity were used to create a generalization data set. In the first case, the first phoneme of the training set verb stem was changed. In the second case, the first two phonemes of verb stems were changed. Both of these changes were however consistent with phonotactics, that is, a C was replaced by another C and a V by another V. In the third case, however, the first two phonemes were changed in such a way that the conformity to phonotactics was violated. This use of novel verbs is standard practice for generalization testing in the context of tense formation [18].

5 Experimental Design

In order to explore the behavior of the model in different lineages (i.e., sequences of development, selection, and breeding), three replications of the model were tested, each having a twenty-generation duration. The experiments were conducted on Condor, a platform that supports running high-throughput computing on large collections of distributive owned computing resources [44]. It follows a master-slave configuration, which has proved suitable for training neural network architectures [32].

Each lineage was characterized by its own initial population (produced with random binary genomes) and unique initial weights. The evolutionary methodology was then applied to each of these model instantiations, so that they all shared the same range of variation for genetic and shared environmental influences. At the same time, however, they were unique, for each of them began with a different initial population created from random binary genomes. The three replications (r_1, r_2, and r_3) of the model permitted evaluation of the robustness of the method. Once the genomes and parameters were generated for the initial set of 50 MZ and 50 DZ twins in a lineage, these were instantiated as three-layered feedforward networks and were trained using the batch version of the Rprop algorithm [1, 40]. The stopping condition was an error goal (mean squared error) of 10^{-5} within 1000 epochs.

Since the focus of these experiments was to explore individual differences in performance and heritability over generations, instead of seeking an "optimal" solution as typically happens when genetic algorithms are used, three replications were considered adequate, as they allowed comparing a population of 12000 ANN models across replications.

Table 5 summarizes the settings used in the experimental design. Empirical data from young children performing the past tense task [18, 49], were used to benchmark the performance of the proposed model, which has been the subject of considerable research in the literature.

6 Results and Discussion

The overall accuracy of the model on regular verbs was higher than that on irregular verbs. The mean performance on the full training set ranged between 74% and 80% for regular verbs, and between 34% and 40% for irregular verbs. The model was able to efficiently generalize the past tense rule to novel items with a mean accuracy rate of around 60%.

Table 5. Experimental settings.

No. of generations	20
Size of population	Breeding = 100; Non-breeding = 100 Total $r_1+r_2+r_3$ across generations = 12,000 ANNs
Size of data sets	Training = 508 Generalization = 508
Training mode	Batch
Max. training epochs	1000
Goal MSE	10^{-5}
Increment to weight change, delta_inc	1.2
Decrement to weight change, delta_min	0.5
Maximum weight change, deltamax	50.0
Minimum weight change, deltamin	10^{-6}
Initial weight update (Rprop learning rate), delta0	Values from genome
Hidden units; Steepness of logistic	Values from genome
Selection operator	Roulette wheel, applied at the end of training (1000 epochs)
Crossover	6 crossovers/chromosome; different operators used
Environmental factor	Probability value between 60% and 100%

The performance of the model compares well with empirical data for children reported in the literature [2, 52]. The behavioral data in [2] comprise performance results of 442 6-year-old children on a past tense test. They were tested on 11 regular verbs and 8 irregular verbs. The average accuracy achieved by the children on regular verbs was 90%, whereas for irregular verbs it was 38%. The performance is also consistent with the performance reported in the developmental study of [52] for 5–7-year-old children: For regular verbs, accuracy rates were 60% (5-year-olds), 75% (6-year-olds), and 80% (7-year-olds); for irregular verbs, accuracy rates were 25% (5-year-olds), 58% (6-year-olds), and 50% (7-year-olds).

We compared the model's performance with that of two other past tense models from [18, 49]. In the former model, 1000 networks were trained for 1000 epochs in various degrees of environmental and genetic variation scenarios. The experimental setting, which closely matched our experiments (referred to as G-wide and E-narrow), resulted in average accuracy of 80% for regular verbs and 38% for irregular verbs. In the latter case, the model comprised networks trained for 400 epochs, with results averaged over 10 replications with different random seeds. The results corresponding to 6-year-olds fall in the range 60%–80% for regular verbs and 20%–40% for irregular verbs, achieved in the window of epochs 51–70. Their model also achieved over 80% generalization accuracy.

Changes in performance levels, heritability estimates, and parameter values over generations were initially analyzed using independent linear regressions. Individually reliable trend lines at the $p = .05$

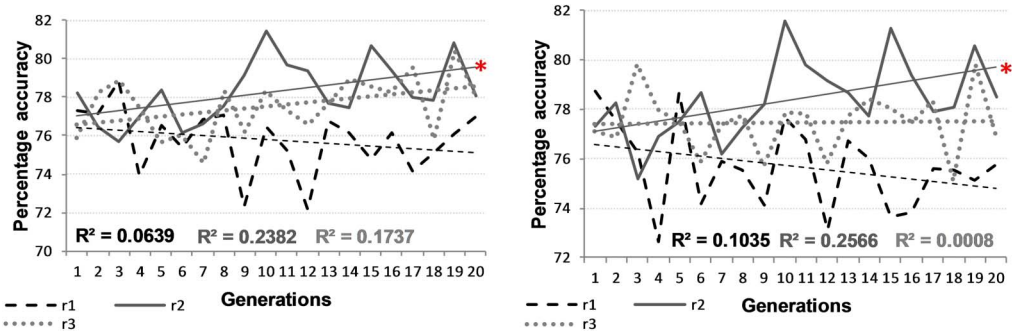

Figure 1. Mean performance per generation for breeding (left) and non-breeding (right) twin populations on regular verbs. The three lines show three different lineages. Statistically reliable ($p < .05$) trend lines are marked with an asterisk (*).

level are shown in the figures. Given the overall design, which combined repeated measures (e.g., regular verb performance, irregular verb performance, generalization) and between-group measures (replication population; breeding versus non-breeding populations), trajectory analysis was used to assess overall patterns in the component linear regressions [47].

Figure 1 depicts the mean accuracy with which breeding and non-breeding twin populations formed past tenses for regular verbs across a sequence of generations, for three replications with differential initial genomes. These graphs summarize the results from 12,000 networks. Figure 2 shows equivalent data for irregular verbs, and Figure 3 presents the generalization results. In each case, a zigzagged line indicates the mean accuracy level of the 100 networks for each population at each generation, while a straight line represents the general trend observed in that replication scenario. The trend line was derived from a linear regression line based on the least-squares method, predicting mean performance level from generation number. R^2 values were small, reflecting the nonmonotonic changes in performance over generations. This is in line with changes in mean trait levels in animal populations following selective breeding, such as the open field behavior of mice [11, 37]. An asterisk in these figures signifies replications where the change over generations was statistically significant at the $p < .05$ level using linear methods.

We initially considered performance in application of the past tense rule, comparing the measures of regular verb performance against generalization, for the three replications and breeding versus non-breeding populations (12 trajectories). A fully factorial ANCOVA revealed no overall change in performance across the generations ($F(1, 108) = 2.23$, $p = .138$, $\eta_p^2 = .020$). However, this masked a differential pattern between replications, with some showing rising performance and others no change ($F(2, 108) = 8.65$, $p < .001$, $\eta_p^2 = .138$). This pattern was common across

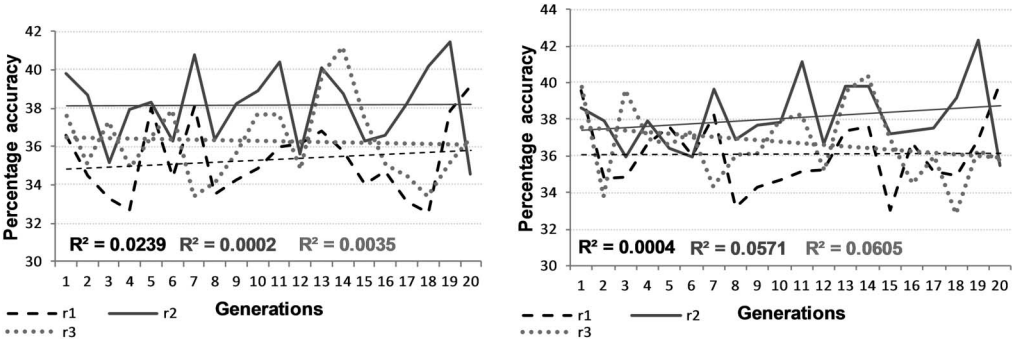

Figure 2. Mean performance per generation for breeding (left) and non-breeding (right) twin populations on irregular verbs. The three lines show three different lineages. Trend lines are not statistically reliable ($p < .05$).

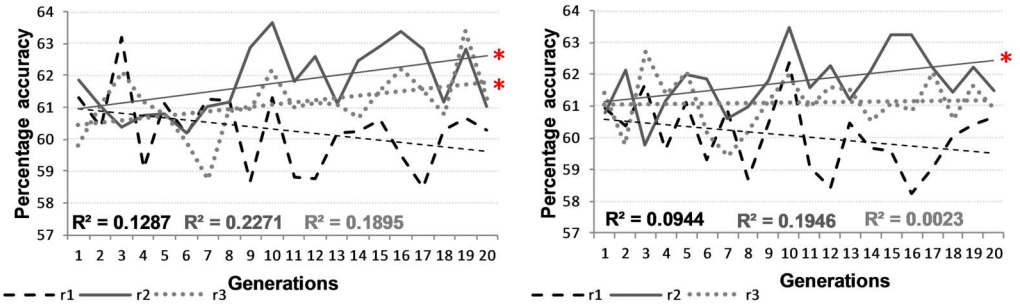

Figure 3. Mean generalization accuracy per generation for breeding (left) and non-breeding (right) twin populations. The three lines show three different lineages. Statistically reliable ($p < .05$) trend lines are marked with an asterisk (*).

measures and breeding versus non-breeding populations. Regular verb performance was reliably higher than generalization ($F(1, 108) = 6288.30$, $p < .001$, $\eta_p^2 = .983$).

Irregular verb performance, by contrast, showed no individual population with rising performance across generations, though the replication populations showed consistently different levels of accuracy ($F(2, 108) = 3.27$, $p = .042$, $\eta_p^2 = .057$). Comparison with regular verb performance indicated that the relationship between performance and generation was reliably modulated by measure ($F(2, 108) = 4.53$, $p = .013$, $\eta_p^2 = .077$). Regular verb performance was also reliably higher than irregular verb performance ($F(1, 108) = 9958.42$, $p < .001$, $\eta_p^2 = .989$).

Most notable in Figures 1 to 3 is the presence of some downward trends in performance over generations, despite the operation of selection. Selection should serve to improve performance over generations, since genes conveying an advantage in learning are more likely to be transmitted to the next generation. The probabilistic nature of this transmission—the mode of sexual reproduction does not guarantee that the advantageous genes of an individual selected to breed will appear in the offspring, and the selection mechanism is itself probabilistically related to the final performance level—accounts for the slow change in population mean performance over generations. It does not account for why performance should become *worse* over generations.

The explanation is suggested by the fact that opposite trends were observed for regular verbs and irregular ones (with generalization patterning with regular verbs). When performance across generations was worsening for regular verbs, it was improving for irregular verbs, and vice versa. This is most evident in a comparison of replication 1, where irregulars showed a trend to improve over generations but regulars to worsen, and replication 3, where regulars showed a trend to improve but irregulars to worsen (three-way interaction of verb type by generation by replication: $F(1, 36) = 4.64$, $p = .038$, $\eta_p^2 = .114$).

Because the learning domain of the English past tense is quasi-regular, good performance across all mappings could in principle be achieved by scoring strongly on regular verbs, strongly on irregular verbs, or strongly on both (with regular verbs the more powerful driver, being in the majority). If optimizing the same computational parameters enhanced both types of mapping, then selecting for either strong regular or strong irregular performance should enhance the performance of the population on the other mapping type as well. However, it is known that the two types of mappings are differentially sensitive to different parameters in ANNs. For example, regular mappings benefit from steeper sigmoid functions, while irregular mappings require more hidden units [48]. The combination of (a) selection by mean performance that could be driven by either stronger regular or irregular verb performance, and (b) parameters that favor learning of either regular or irregular mappings, together sets the stage for possible divergence of gene pools over generations. Even in the face of selection, some lineages may become specialized for regular verbs at the expense of irregular verbs, while other lineages may become specialized for irregular verbs at the expense of regular verbs. Yet others may show increased performance in both verb types across generations. Which path a given starting population follows will depend on the distribution of parameters created by the

initial genomes, the set of individual environments, and stochastic factors involved in selection and breeding. Once genes are lost from the population that are beneficial for the non-preferred verb type, these cannot be regained. Selection can then improve performance only by continuing to select for genes for the preferred verb type.

This phenomenon is analogous to Waddington's epigenetic landscape, an idea proposed by Conrad Waddington to account for restriction of fate in development [13, p. R459]. In his model, Waddington compared the process of cellular differentiation to a marble, representing a pluripotent cell, on top of a hill. The hill contains many paths or valleys that the marble can roll down, and each path will eventually lead to a distinct final differentiated state, such as a blood cell or a skin cell. He described each of the valleys as an individual developmental pathway or "chreode." As the marble moves down the hill, the paths and final destinations available become more limited, representing the increased differentiation of the cell [53]. This is what makes an initial pluripotent cell become a specialized cell, and reversing this process is impossible under normal circumstances. In this case, the restriction of fate occurs over generations in selection of genes for parameter values more appropriate to one or the other verb type.

Changes in the frequency of different gene variants (here, binary values 0 or 1) in the gene pool should alter the range of genetic variation across generations, since any effective operation of selection would reduce genetic diversity. Given that the range of environmental variation (EF of 0.6 to 1.0) remained consistent across generations, genetic diversity could explain less of the phenotypic variation, and a corresponding reduction in heritability would be expected. To test this idea, we examined correlations in performance between MZ and DZ network twin pairs, using Falconer's equations to derive estimates of heritability [37]. Heritability was estimated as twice the difference between MZ and DZ correlations; unique environmental effects, as the extent to which MZ correlations were less than 1; and shared environment effects, as the remaining variance (i.e., 1–{heritability}–{unique environment}). Strictly speaking, these equations assume an additive model, which only holds for MZ correlations that are no more than twice DZ correlations. In our results the correlations sometimes violated this condition. However, for consistency, we plot heritability estimates according to the same formulas across conditions, though the values sometimes appear outside of the range 0 to 1 as the assumptions of the additive model become violated. The plotted data should therefore be seen as proportional to the heritability and environmentability observed in populations, rather than direct estimates under an additive model.

Figure 4 shows the estimates of heritability (variance due to genetic factors) for regular (a) and irregular verbs (b). These six trajectories were compared in a fully factorial ANCOVA. Heritability reliably reduced over generations ($F(1, 54) = 5.54$, $p = .022$, $\eta_p^2 = .093$), and this pattern was not modulated by the measure or replication population. Though replication 2 showed the steepest reduction in heritability, the difference in the pattern across replications was not reliable ($p = .107$).

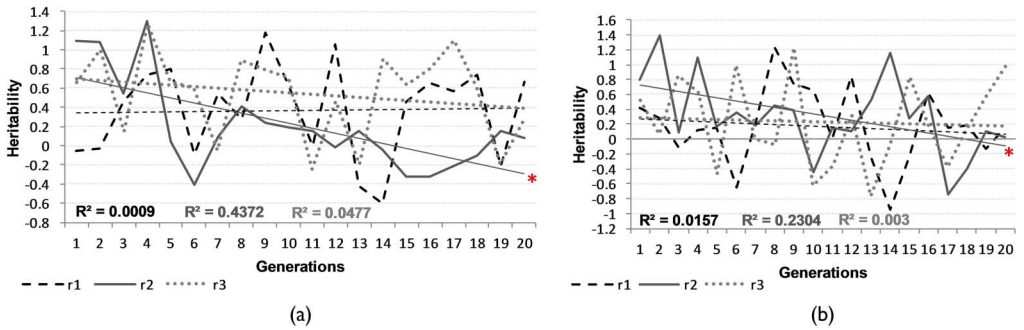

Figure 4. Heritability or proportion of variance due to genetic (or structural) factors for (a) regular and (b) irregular verbs. The three lines show three different lineages. Statistically reliable ($p < .05$) trend lines are marked with an asterisk (*).

If a lineage becomes increasingly optimized on a task (or a specific aspect of the task domain), the range of the intrinsic parameters relevant to that task (so-called *domain-relevant* parameters) should decrease across generations, as only the genes producing the preferred parameter values are retained. For example, if populations are improving on irregular verbs, which require more capacity to hold nonsystematic mappings, then across generations, networks with larger numbers of hidden units would have a greater chance to get selected in the breeding pool. Across generations, the variability in the range of the number of hidden units would reduce. By contrast, the range of variation in other parameters less relevant to irregular-verb performance would be less affected. Optimization and heritability should therefore have an inverse relationship in this case.

Figure 5 plots the performance levels and heritability estimates for regular verbs, which showed the strongest optimization, for all replications across all generations. It shows a reliably negative relationship between performance and heritability as predicted ($R^2 = .079$, $F(1, 118) = 10.12$, $p = .002$). The more the population-mean performance increased, the lower the heritability estimate. However, the lineages behaved differently in detail.

In replication (lineage) 1, regular-verb performance and rule generalization dropped across generations while irregular-verb performance improved. Heritability for regular verbs was initially higher than that for irregular verbs, centered on 0.4, and it then increased across generations, implying lack of selection for parameter sets specialized for regularity. By contrast, heritability of irregular verbs was lower, centered on 0.2, and decreased across generations, implying selection for, and narrowing of the range of, parameter sets specialized for irregularity. Note that this process of specialization caused the *overall* accuracy to drop, because irregular verbs form a minority of the data set (there were only 98 irregular verbs and 410 regular verbs).

In replication (lineage) 2, regular-verb performance, irregular-verb performance, and generalization all increased across generations. Heritability of regular verbs dropped from around 0.8 to around zero. A similar pattern was observed for irregular verbs, with heritability dropping from high values to almost zero. In this lineage, optimization caused a narrowing of the range of genetic variation relevant to learning of both regular and irregular verbs.

In replication (lineage) 3, regular-verb performance and generalization improved while irregular-verb performance dropped. The heritability of regular verbs decreased from 0.6 to 0.2 while the heritability of irregulars remained stable, at around 0.2. These two observations suggest that the range of intrinsic parameters being targeted by selection initially accommodated both regulars and irregulars, but as generations progressed, there was a narrowing in this range for parameters more suited to regular verbs.

When heritability of a particular aspect of the task reduces, it implies that the variance in performance is less due to genetic factors and more due to shared and non-shared environmental factors.

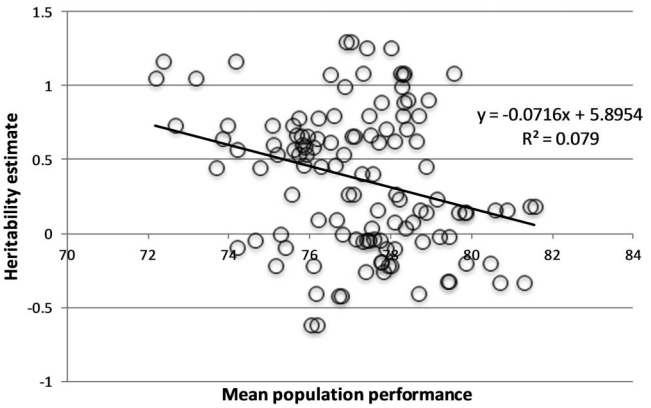

Figure 5. Negative association between heritability and population mean performance on regular verbs (data for all generations and lineages).

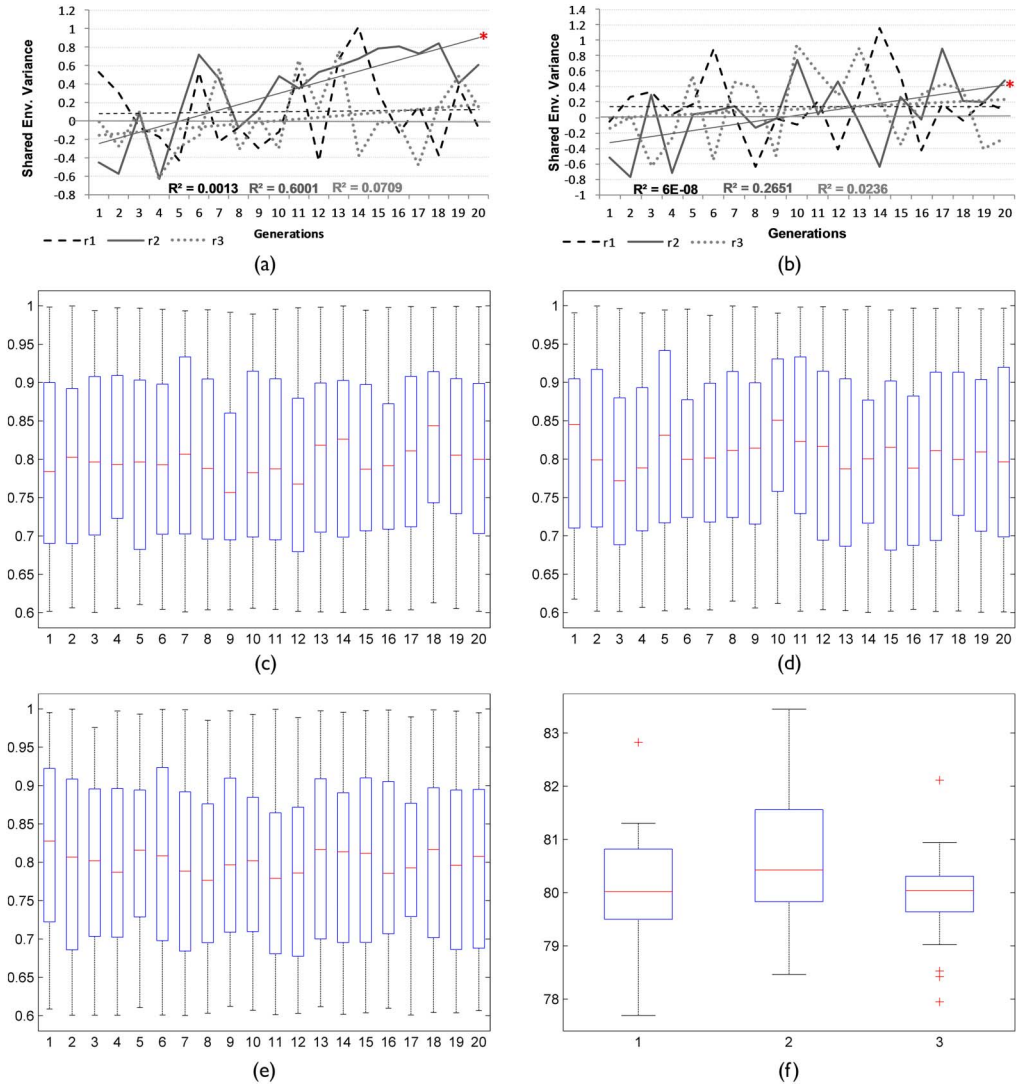

Figure 6. (a) Proportion of variance due to shared environmental factors—regular verbs. (b) Proportion of variance due to shared environmental factors—irregular verbs. (c) Percentage of EF per generation: replication 1. (d) Percentage of EF per generation: replication 2. (e) Percentage of EF per generation: replication 3. (f) Percentage of EF per replication.

Figures 6(a) and 6(b) display the variance due to shared environmental factors, in this case the filtered training data sets. The effect of shared environment reliably changed over generations ($F(1, 54) = 8.42, p = .005, \eta_p^2 = .135$), though this was driven primarily by replication 2, illustrated by an interaction of population × generation ($F(2, 54) = 3.65, p = .033, \eta_p^2 = .119$). The pattern was common across regular and irregular verbs. Figures 6(c)–(6f) confirm that, though stochastically sampled, the range and mean level of EF were constant across generations for all lineages.

Figures 7(a) and 7(b) present the variance in performance due to non-shared environmental factors or initial weights in our implementation. Analyses revealed no reliable effects, with non-shared environmental effects consistent across generations and modulated neither by measure type nor by replication population. The graphs show that the differences in initial weights led to large variability in behavioral outcomes. In cases when intrinsic factors were not very suitable for the task domain,

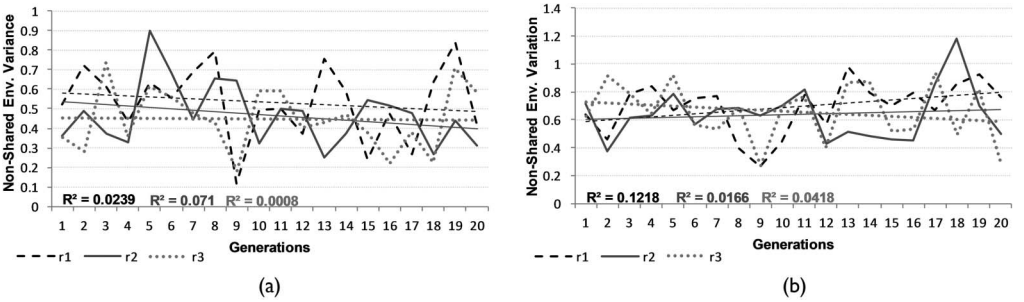

Figure 7. Proportion of variance due to non-shared environmental factors: (a) regular verbs; (b) irregular verbs.

having good initial weights might give networks a fighting chance, exaggerating the effect of intrinsic stochastic factors.

Heritability is a useful statistic because it is scalable across potentially very large numbers of computational parameters (and their interactions) that contribute to the variation in learned high-level behaviors, or in this case, the outcome of learning for a set of ANNs. However, in the current simulations, relatively few parameters were encoded in the genome and permitted to vary across populations and between generations. Our final step of analysis then, was to examine the change in mean parameter values for a given lineage across generations. This should reveal the domain-relevant parameters that were selected in those cases where performance on one verb type was enhanced at the expense of the other, and should therefore in turn reveal the drivers behind changes in heritability. Figure 8 depicts changes in mean parameter values for number of hidden units, initial learning rate, and slope of the logistic activation function. For hidden units, there was a reliable reduction in number across generations ($F(1, 54) = 190.55$, $p < .001$, $\eta_p^2 = .779$), with the reduction occurring at different rates across the three replication populations ($F(2, 54) = 33.79$, $p < .001$, $\eta_p^2 = .556$).

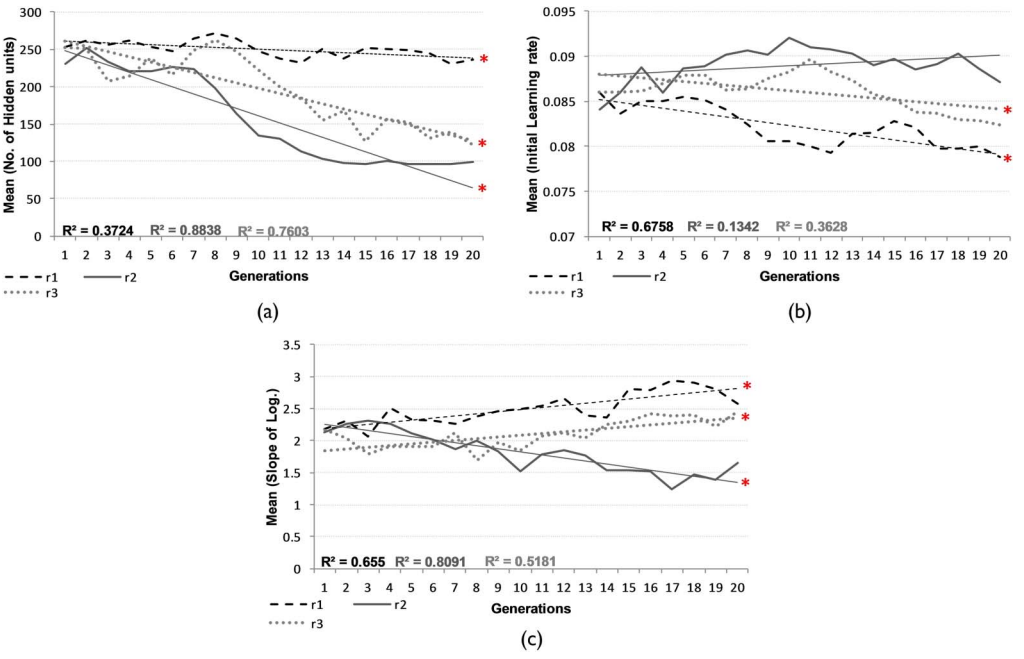

Figure 8. (a) Change in the mean value of the number of hidden units per generation. (b) Change in the mean value of the initial learning rate per generation. (c) Change in the mean value of the slope of logistic activation per generation.

For the learning rate, the same pattern was observed: reduction across generations ($F(1, 54) = 22.69$, $p < .001$, $\eta_p^2 = .296$) modulated by replication, with the reduction appearing in only two of the three replications ($F(2, 54) = 12.22$, $p < .001$, $\eta_p^2 = .312$). Lastly, for the slope of logistic activation, there was an increase in two of the populations across generations and a reduction in the other (main effect of generation: $F(1, 54) = 12.99$, $p < .001$, $\eta_p^2 = .325$; interaction of generation × replication: $F(2, 54) = 61.06$, $p < .001$, $\eta_p^2 = .693$). Overall, replications 1 and 3 showed a common pattern of reduction in hidden units, reduction in learning rate, and increase in temperature. For replication 1, the reduction in hidden units was milder, the learning rate fell lower, and the temperature rose higher. Replication 2 showed a different pattern: a greater fall in hidden units, no change in learning rate, and a drop in temperature (i.e., slope of logistic activation).

The three chosen parameters provided networks with capacity to learn (more hidden units can accommodate more input-output mappings) and/or ability to learn (optimum values of initial learning rate and steepness of logistic activation allow discovery of connection weights for those mappings). Irregular verbs belong to the category of nonsystematic mappings, which are more demanding of computational capacity. Figure 9 depicts the variation in the ranges of the three parameters across generations. In Figure 9, on every box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and the outliers are marked separately by +. The height of the box represents the interquartile range (IQR) of the data set, which is the difference between the 75th percentile and 25th percentile. The lines at the end of the whiskers mark the highest and lowest values of the data set that are within 1.5 times the inter quartile range of the box edge.

The figure reveals the parameters being targeted by selection in each lineage. While the range of variation of computational parameters appears uniformly spread throughout the lineages, the range skews towards values that make the parameters act in a domain-relevant way, and produces the observed changes in heritability.

Lineage (replication) 1 improved irregular-verb performance at the expense of regular-verb, and this was reflected by maintenance of high levels of hidden units. Learning rates declined, while genes for steeper logistic slopes were selected.

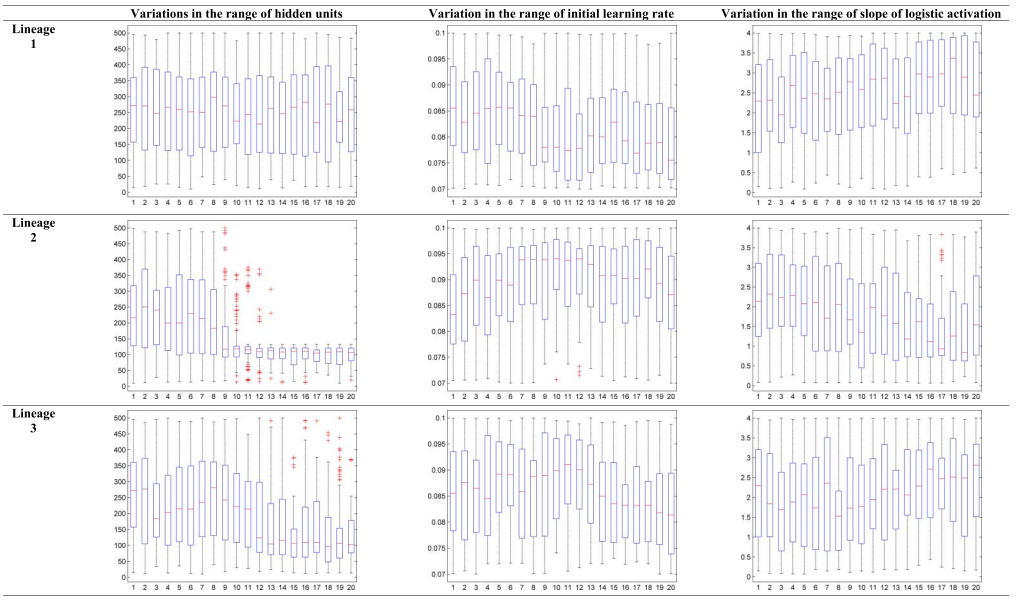

Figure 9. Range of variation of intrinsic parameters across generations. Boxes represent the interquartile range (IQR), whiskers the range excluding outliers, and + mark outliers (outside 1.5 × IQR).

Regular verbs have systematic input-output mappings, which are less demanding of computational capacity. Lineage (replication) 2 improved regular-verb performance at the expense of irregular-verb, and this was reflected by an increase in learning rate. Both hidden-unit numbers and logistic slope declined.

In lineage (replication) 3, the main improvement over generations was on regular verbs. As with lineage 2, there was a decline in hidden-unit number, but unlike lineage 2, there was also a decline in learning rate. Instead, the logistic slope showed an increase, which lineage 1 suggested was better adapted to accommodating irregular mappings.

7 Findings

Research in artificial life has been interested in the relationship between evolution and learning as complementary optimization techniques for ANN [9, 28]. Evidence suggests that evolution can select network parameters and starting states that improve learning outcomes. Here we have considered this relationship in a model where heritable network properties were passed between generations via a simulated process of sexual reproduction, where networks were exposed to learning environments of variable quality, and where the problem domain was quasi-regular, providing greater challenges for optimization. The main findings were as follows:

i. Applying selection on the individual's performance level in a quasi-regular task, such as past tense acquisition, results in the emergence of divergent behaviors depending on initial conditions—both genetic and environmental.

ii. Once selection starts targeting a particular aspect of task domain, there is restriction of fate, or canalization, in a fashion analogous to Waddington's epigenetic landscape, but now in an evolutionary sense rather than a developmental one. From an initial pluripotent state, the pathway of a lineage can become optimized to regular or irregular properties of the domain in a way that cannot be reversed.

iii. Assumptions on the framework, inspired by behavioral genetics, restricted optimization in three ways. First, because the quality of the learning environment varied independently of the quality of the genotype, it was harder to assign credit or blame to the genotype on the basis of the outcome of learning. Second, if preferential properties of the genome were to be identified, their transmission to the next generation would be compromised by probabilistic (roulette) selection to reproduce; by sexual reproduction, which only probabilistically passes on advantageous traits; and by death (removal of the advantaged individual from the population).

iv. Heritability is a metric that identifies the net contribution of genotypic variation to phenotypic variation and that is invariant to the number of parameters contributing to that variation. In this case, it acted as an identifier of the aspect of the quasi-regular task being targeted by selection. Highly heritable behavior indicates that the trait is not being selected for, whereas behavior with reduced heritability implies selection and optimization. Therefore *an inverse relationship exists between heritability and optimization.*

v. When selection operated, a higher proportion of variance due to shared environmental factors (filtered training sets) was associated with stronger learning.

vi. Non-shared environmental factors (initial weights) led to larger values of behavioral variance. This effect was larger when intrinsic properties were not particularly suitable and chance superior initial weight configurations were necessary to have good learning outcomes.

Finally, as in all simulations that employ stochastic factors, the results would benefit from further replications—here, of lineage on the one hand, and extended selection across generations on the

other hand—to ensure that convergence has been achieved. In the training of a single ANN, regular verbs are learnt first; with continued training, learning of irregular verbs follows, as these exception mappings are accommodated around the connectivity that supports the regular past tense rule [39, 41]. However, evolution is not like development. Once the genes are lost for the computational properties that enable learning of irregulars alongside regulars, this potential cannot be regained by further selection over more generations. Additional trials in the context of transfer learning reported in [22] reveal that this behavior is also encountered in other data sets.

8 Conclusion

In this article, we have introduced a novel computational approach, inspired by principles of behavioral genetics, to model the performance of 6-year-old children on English past tense acquisition. We analyzed the proportion of variance accounted for by ANN computational parameters (encoded in an artificial genome) and filtered training sets and initial weights (equivalent to the effect of environment), highlighting the effects of selection and sexual reproduction. Our model was able to identify the causal factors leading to behavioral and performance variability within the population. The model demonstrated that divergent behavioral outcomes can emerge when selection is applied to a quasi-regular task, where selection of parameters becomes more specialized in one aspect of the task across generations. Importantly, the model showed that heritability and optimization have an inverse relationship, with heritability identifying which aspects of the task domain are being targeted by selection.

Several avenues require further investigation, including replication across lineages and more extended selection to verify convergence. Theoretically, more complex genome representations may allow encoding more computational parameters and increased genetic variability. Alternative selection schemes may alter the divergence of population genotypes over generations, as might nonrandom assignment of environments to genotypes (gene-environment correlations), which is observed to occur in human populations [20].

References

1. Anastasiadis, A. D., Magoulas, G. D., & Vrahatis, M. N. (2005). New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, *64*, 253–270.
2. Bishop, D. V. M. (2005). DeFries–Fulker analysis of twin data with skewed distributions: Cautions and recommendations from a study of children's use of verb inflections. *Behavior Genetics*, *35*(4), 479–490.
3. Bottou, L. Y. (1988). Reconnaissance de la parole par reseaux multi-couches. In *Proceedings of the International Workshop Neural Networks Application, Neuro-Nimes, 88* (pp. 197–217). EC2 and Chambre de Commerce et d'Industrie de Nimes.
4. Chater, N., & Christiansen, M. H. (2008). Computational models of psycholinguistics. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 477–504). Cambridge, UK: Cambridge University Press.
5. Chomsky, N. (2014). *Aspects of the theory of syntax* (p. 11). Cambridge, MA: MIT Press.
6. Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1–61.
7. Cooper, G. M., & Hausman, R. E. (2000). *The cell: A molecular approach*, 10. Washington: ASM Press.
8. Cottrell, G. W., & Plunkett, K. (1991). Learning the past tense in a recurrent network: Acquiring the mapping from meaning to sounds. In K. J. Hammond & D. Gentner (Eds.), *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 328–333). Hillsdale, NJ: Lawrence Erlbaum Associates.
9. Crispo, E. (2007). The Baldwin effect and genetic assimilation: Revisiting two mechanisms of evolutionary change mediated by phenotypic plasticity. *Evolution: International Journal of Organic Evolution*, *61*(11), 2469–2479.
10. Daugherty, K., & Seidenberg, M. S. (1992). Rules or connections? The past tense revisited. In K. J. Hammond & D. Gentner (Eds.), *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 259–264). Hillsdale, NJ: Lawrence Erlbaum Associates.

11. DeFries, J. C., Gervais, M. C., & Thomas, E. A. (1978). Response to 30 generations of selection for open-field activity in laboratory mice. *Behavior Genetics*, *8*(1), 3–13.

12. Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Harlow, Essex, UK: Longmans Green, 3.

13. Ferrell Jr., J. E. (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. *Current Biology*, *22*(11), R458–R466.

14. Fromkin, V., Rodman, R., & Hyams, V. (2011). *An introduction to language* (11th ed.). Boston: Wadsworth, Cengage Learning.

15. Gong, T., & Shuai, L. (2013). Computer simulation as a scientific approach in evolutionary linguistics. *Language Sciences*, *40*, 12–23.

16. Griffiths, P. E. (2009). The distinction between innate and acquired characteristics. In *Stanford Encyclopedia of Philosophy*. Stanford, CA: Metaphysics Research Laboratory, Stanford University. Accessible at https://plato.stanford.edu/entries/innate-acquired/.

17. Karaminis, T., & Thomas, M. S. C. (2015). *The multiple inflection generator: A generalized connectionist model for cross-linguistic morphological development*. Available at http://www.bbk.ac.uk/psychology/dnl/old_site/personalpages/KT_TheMultipleInflectionGenerator.pdf.

18. Karaminis, T., & Thomas, M. (2010). A cross-linguistic model of the acquisition of inflectional morphology in English and Modern Greek. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32), 730–735.

19. Karmiloff-Smith, A., & Thomas, M. S. C. (2003). What can developmental disorders tell us about the neurocomputational constraints that shape development? The case of Williams syndrome. *Development and Psychopathology*, *15*(4), 969–990.

20. Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2016). *Behavioral genetics*. New York: Macmillan Higher Education.

21. Kohli, M., Magoulas, G. D., & Thomas, M. (2012). Hybrid computational model for producing English past tense verbs. In C. Jayne, S. Yue, & L. Iliadis (Eds.), *International Conference on Engineering Applications of Neural Networks* (pp. 315–324). Berlin, Heidelberg: Springer.

22. Kohli, M., Magoulas, G. D., & Thomas, M. S. (2013). Transfer learning across heterogeneous tasks using behavioural genetic principles. In Y. Jin & S. A. Thomas (Eds.), *2013 13th UK Workshop on Computational Intelligence (UKCI)* (pp. 151–158). New York: IEEE Press.

23. Lupyan, G., & McClelland, J. L. (2003). Did, made, had, said: Capturing quasi-regularity in exceptions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *25*, 740–745.

24. MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*(1), 199–227.

25. MacWhinney B. (2005). A unified model of language acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). New York: Oxford University Press.

26. Mareschal, D., & Thomas, M. S. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, *11*(2), 137–150.

27. Mitchell, M., & Forrest, S. (1994). Genetic algorithms and artificial life. *Artificial Life*, *1*(3), 267–289.

28. Munroe, S., & Cangelosi, A. (2002). Learning and the evolution of language: The role of cultural variation and learning costs in the Baldwin effect. *Artificial Life*, *8*(4), 311–339.

29. Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

30. Pinker, S. (1994). *The language instinct: The new science of language and mind* (p. 7529). London: Penguin UK.

31. Plagianakos, V. P., Magoulas, G. D., & Vrahatis, M. N. (2006). Evolutionary training of hardware realizable multilayer perceptrons. *Neural Computing & Applications*, *15*(1), 33–40.

32. Plagianakos, V. P., Magoulas, G. D., & Vrahatis, M. N. (2006). Distributed computing methodology for training neural networks in an image-guided diagnostic application. *Computer Methods and Programs in Biomedicine*, *81*(3), 228–235.

33. Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56.

34. Plomin, R., & DeFries, J. C. (1980). Genetics and intelligence: Recent data. *Intelligence*, *4*(1), 15–24.

35. Plomin, R. (1990). The role of inheritance in behavior. *Science*, *248*(4952), 183–188.

36. Plomin, R., & Spinath, F. M. (2004). Intelligence: Genetics, genes, and genomics. *Journal of Personality and Social Psychology*, *86*(1), 112.

37. Plomin, R., DeFries, J. C., & McClearn, G. E. (2008). *Behavioral genetics*. New York: Macmillan.

38. Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*(1), 21–69.

39. Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43–102.

40. Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proceedings of the IEEE International Conference on Neural Networks* (pp. 586–591). San Francisco: IEEE Press.

41. Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*, *II*, 216–271. Cambridge, MA: MIT Press.

42. Simon, H. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467–482.

43. Smith, A. D. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 281–293.

44. Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The Condor experience. *Concurrency and Computation: Practice and Experience*, *17*(2–4), 323–356.

45. Thimm, G., & Fiesler, E. (1995). Neural network initialization. In J. Mira & F. Sandoval (Eds.), *International Workshop on Artificial Neural Networks* (pp. 535–542). Berlin, Heidelberg: Springer.

46. Thomas, M. S. (2018). A neurocomputational model of developmental trajectories of gifted children under a polygenic model: When are gifted children held back by poor environments? *Intelligence*, *69*, 200–212.

47. Thomas, M. S. C., Annaz, D., Ansari, D., Serif, G., Jarrold, C., & Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research*, *52*, 336–358.

48. Thomas, M. S., Forrester, N. A., & Ronald, A. (2016). Multiscale modeling of gene–behavior associations in an artificial neural network model of cognitive development. *Cognitive Science*, *40*(1), 51–99.

49. Thomas, M. S. C., Ronald, A., & Forrester, N. A. (2009): *Modelling the mechanisms underlying population variability across development: Simulating genetic and environmental effects on cognition* (DNL technical report). London: Birkbeck, University of London.

50. Thomas, M. S., Forrester, N. A., & Ronald, A. (2013). Modeling socioeconomic status effects on language development. *Developmental Psychology*, *49*(12), 2325.

51. Thomas, M. S., & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 23–58). Cambridge, UK: Cambridge University Press.

52. Van der Lely, H. K., & Ullman, M. T. (2001). Past tense morphology in specifically language impaired and normally developing children. *Language and Cognitive Processes*, *16*(2–3), 177–217.

53. Waddington C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. London: Allen & Unwin.

54. Whitley, D., Starkweather, T., & Bogart, C. (1990). Genetic algorithms and neural networks: Optimizing connections and connectivity. *Parallel Computing*, *14*(3), 347–361.

55. Wintner, S. (2010). Computational models of language acquisition. In A. Gelbukh (Ed.), *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 86–99). Berlin, Heidelberg: Springer.

56. Yam, J. Y., & Chow, T. W. (2000). A weight initialization method for improving training speed in feedforward neural network. *Neurocomputing*, *30*(1–4), 219–232.