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

Maitrei Kohli[*+]
Birkbeck, University of London

George D. Magoulas[*]
Birkbeck, University of London

Michael S.C. Thomas[**]
Birkbeck, University of London

Abstract

Children's acquisition of English past tense has been widely studied as 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, very few consider individual differences. In this paper, we explore the use of populations of artificial neural networks (ANNs) that evolve according to Behavioural 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 Behavioural 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 the neuro-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 artificial neural network twins based on the biological processes of meiosis and fertilization. This approach allows modelling 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. This is an important aspect that distinguishes our work from others reported in literature, setting individual differences within an evolutionary framework. We present an experimental evaluation of this model focusing on individual differences in performance. The experiments led to some interesting findings such as: applying selection on the individual's performance level in a quasi-regular task such as past tense acquisition results in the emergence of divergent behaviours depending on initial conditions – both genetic and environmental; once selection starts targeting a particular aspect of task domain, it starts behaving similar to Waddington's epigenetic landscape; and selection based on a stochastic method such as roulette-wheel, when combined with sexual reproduction method for population generation, has a limiting effect on final behavioural (or performance) levels achieved. The findings validate the effectiveness of the method within an evolutionary setting and provide the basis for future work to capture population-level differences within a developmental setting.

Keywords
Neural networks-based modelling, behavioural genetics, class imbalance, quasi-regular mappings, English past tense, feed forward neural networks, genetic computing

---

+ Contact Author
* Department of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, Bloomsbury, London WC1E 7HX, UK
Email: {maitrei, gmagoulas}@dcs.bbk.ac.uk
** Department of Psychological Sciences, Birkbeck, University of London, Malet Street, Bloomsbury, London WC1E 7HX, UK
Email: m.thomas@psychology.bbk.ac.uk

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 [13, 23, 39].

Language learning is considered one of the most complex tasks. Nevertheless, most children acquire it naturally, effortlessly, and quickly compared to other areas of cognitive development. Language is like the majority of complex systems which exist in nature and which empirically exhibit hierarchical structure [38].

Two opposing theories of language acquisition dominate the linguistic and psycholinguistic communities (refer to [49] for a review). The nativist approach, proposed by Chomsky [4, 5], and promoted by Pinker, claims that the linguistic capability at least with respect to grammar is innate; therefore, certain linguistic universals are given to the language learners for free; only the established parameters need little tweaking in order for language to be fully acquired [25].

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 [20, 21].

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 [15]. 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 [45]; (for reviews, see [3]; [22]).

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 plurals for nouns). Within this field, the acquisition of 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 [15]. Children's acquisition of English past tense has been the focus of great deal of empirical research, mostly due to its *quasi-regular* mappings [45]. Quasi-regular domains are interesting because of the presence of systematic input-output mappings but the presence of a minority of exceptions [45].

The majority of English verbs, viz. regular, form their past tense by following a rule for stem suffixation, also referred to as *+ed* rule. This rule allows for three possible phonological suffixes [14] - */d/* e.g. raise – raised; */t/* e.g. clap – clapped; */ed/* e.g. visit – visited. 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 [19]. This overlap between regular and irregular verbs adds to the complexity of task domain. (See the mapping between written and spoken forms of English for another example of a quasi-regular domain within language, [28]).

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. (Refer to [42] 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? Or can productivity emerge from associative mechanisms exposure to quasi-regular domains?

There are two main theories. The first is a dual route account, proposed by Pinker [24], 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 [37] 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 feed-forward neural network can 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 [14], for details).

This Backpropagation algorithm-based model inspired many subsequent connectionist models of acquisition of inflections like [7, 9, 34, 35] to name a few. Subsequent connectionist models addressed many of the drawbacks of the initial model. For example, Plunkett and Marchman [35] took the main idea from Rumelhart model and modified it into a three-layered feed-forward 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 past tense acquisition related phenomena. 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 Specific Language Impairment and Williams syndrome [16]. 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 [45] modelled the effects of socio-economic 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, simulated as modulation of the structured learning environment, against a background of variation in the computational power of individual's 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 modelling 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 variance in language development [45]. The second innovation is the use of *multi-scale modelling* to reconcile data from multiple levels of description, including genetic, neural structure, cognitive processes, behaviour, and the environment, where behaviour itself 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-behaviour 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 behaviour [46].

In artificial life research, Genetic algorithms are usually employed for optimisation, where selection across generations aims to improve the performance of learning systems on a target task. By contrast, the existing multi-scale models took 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 through the operation of selection? What are the implications of using a quasi-regular domain as the target problem for optimisation? What parts of the problem domain are optimised across generations and what factors determine this?

To address these questions, in this work, we build on our previous investigation that combined concepts of Behavioural Genetics with the idea of parametrically diverse populations of learning systems, where genes (representing intrinsic factors) and environment (expressed via training datasets) interact throughout development to shape differences in individual classifier behaviours [18]. We extend the framework to an evolutionary context by introducing selection in the populations' optimisation process across generations, focusing on learning a particular task: 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 literature. In this context, we present our synergistic approach to capture population variability stemming from genetic and environmental influences and to analyse effects of selection on behavioural outcomes.

This approach not only captures the heterogeneity observed in acquiring a new ability but also helps in understanding how the quality of environment interacts with intrinsic constraints, leading to an individual's overt behaviour. It shows, for example, the different behaviours emerging due to interaction of quality of 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 behaviour across generations.

In Behavioural Genetics, factors affecting language development are attributed to genetic and environmental influences [33]. To model genetic influences, we encode the variation in neuro-computational parameters of ANNs, thereby modulating their learning efficiency. These parameters relate to how a network (individual) is built (the number of hidden units), its processing dynamics (slope of 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 structure [45]. 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 Behavioural Genetics, both genes (intrinsic factors) and environment (training datasets) interact throughout development to shape differences in individual behaviours (performance) [32]. 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 modelling context, a certain number of ANN hidden units may be highly beneficial for a specific condition of the environment (say 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 may not be able to accommodate the change. Thus, the system's performance will alter.

Apart from having genetic and environmental variation, our model also incorporates "selection" and its effects. As is shown later in the paper, applying selection on performance on the English past tense problem leads to two novel findings: (i) selection targets different aspects of a quasi-regular task depending on different initial conditions, potentially producing divergent populations. This in turn results in emergence of different and varied behavioural (performance) patterns, while still optimising on the target task; (ii) the amount of performance variation explained by genetic similarity, the so-called heritability metric [33] plays an important role in identifying which aspect of this quasi-regular task is being targeted by selection.

The rest of the paper is organised as follows. First, we give an overview of the proposed hybrid computational model, combining neural networks and evolution, and its inspiration in Behavioural Genetics. We then explain the methodology for the implementation of the model and the past tense dataset used in the simulations. Finally, we present the results and discuss their implication.

2. Behavioural Genetics Inspired Hybrid Computational Model

Behavioural Genetics is a field of study that examines the role of genetics in individual differences in human behaviour. Behaviour is the most complex phenotype as it reflects the functioning of the complete organism; it is dynamic and changes in response to the environment [30]. This field is concerned with the study of individual differences, i.e. 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 behaviour or phenotype is the result of genetic factors together with environmental factors.

In Behavioural Genetics, twin studies are widely employed to untangle genetic and environment effects on behaviour. 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*) [32]. The difference in the similarity in performance between MZ or DZ twin pairs, along with assumptions about their similarity of environment, allows inferences to be drawn about the influence of genetic relatedness on behaviour [31].

Environmental influences are defined as being of two types, shared (or between-family) and non-shared (or unique and within-family). Shared, or between-family, environmental influences are those which are shared amongst family members and serve to make members of a family (in this case, twins) similar to each other and different from members of other families. Shared environmental influences often tend to include family structure, socioeconomic status, and parental education to name a few [29]. By contrast, non-shared, or within-family, environmental influences are factors that are not common amongst family members, serving to make individuals different from one another. These environmental influences often 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 [29, 32].

Our work exploits the notion of a “population of twins”, i.e. ANNs with some degree of similarity in their neuro-computational parameters, to simulate and thereby investigate the process of disentangling genetic and environmental influences on performance.

Within the field of Behavioural Genetics “heritability” is a central concept, and it is also an important aspect of the current work. The Heritability statistic is defined as the proportion of observed or phenotypic variance that can be explained by genetic variance. In simpler terms, heritability is the amount of population variability explained by genetic similarity [33]. There has been increasing acceptance that in humans, many high-level behaviours show marked heritability [33], a finding that would have been surprising to many researchers in the latter part of the 20th Century.

In computational terms, heritability can be interpreted as the amount of performance variation accounted for by structural similarity. Additionally, twin studies provide a valuable tool for exploring environmental influences, especially family or shared environment, against a background of heritability.

To measure heritability and the proportions of variance explained by shared and non-shared environments, we use a technique based on Falconer’s equations [10] as described in [18]. 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 between MZ and DZ twins shows about half the genetic influence on behaviour; 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 [34]. 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 [36] for training; and then introduced selection across generations of twin pairs.

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, or genes, to constrain the learning abilities of ANNs. The first two parameters, 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 [26, 43].

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/capacity to learn new tasks.

Simulating variations due to environmental influences: Variations in shared environmental influences were simulated through variations in the Environmental Factor (EF), in this case 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 utilise 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) [45].

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 modelled 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, 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. Here, it included any measurement error, as well as stochastic factors such as the initial weights of ANNs.

The learning speed and fast convergence of many feed forward neural networks depend to some extent on their initial values of weights and biases [41, 50]. For this reason, in our approach, initial values of weights were one way to capture unique environments. The initialisation method used in this work is similar to that proposed by [2] and uses the interval: $[-\frac{a}{\sqrt{d_{in}}}, +\frac{a}{\sqrt{d_{in}}}]$; wherein a is chosen in a way that weight variance corresponds to the points of maximum curvature of activation function. This value is 2.38 for standard sigmoid function [41]; and d_{in} is fan-in of neuron or 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 captured the individual differences in performance. The starting point of this work was to estimate the proportion of variance attributed by variances in structural parameters (genes), training set (shared environment), and initial weights (non-shared environment). The methodology adopted can be summarised as follows:

Design ANNs*:* the first step was to design ANNs incorporating neuro-computational parameters that constrained their ability to learn. We selected three free parameters, each of which corresponded to how the network is built: number of hidden units; its activation dynamics, i.e. slope of logistic function; and how it adapts, i.e. learning rate, or the initial learning rate of Rprop.

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. 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, as well as the values which resulted in successful learning. 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 binary representation, whereby each gene had two variants or alleles, with 10 bits per parameter, split into two chromosomes. The parameters and their range of variation are given in Table 1.

Parameters	Range of variation
No of hidden units	10 - 500
Learning rate	0.07 - 0.1
Slope of logistic	0.0625 - 4.0

Table 1: Genome representing ANN parameters and their ranges

Breed the population: the next step concerned breeding the population of ANN twins using the genome. We simulated the biological processes of meiosis and fertilisation to create 50 pairs of MZ and 50 pairs of DZ twins. This method was chosen because it is the closest simulation of actual biological processes (refer to [6] for details about biological meiosis and fertilisation).

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 of these individuals was split into two equal halves, resulting in two chromosomes per individual. Each chromosome contained half the information to code for each parameter. Then crossover was applied multiple times, say *m* times on these chromosomes. Each crossover results in two sperms or eggs. Now these sperms and eggs need to be combined to get the desired offspring. To do this positional recombination was used to combine the sperms and eggs, such that for each parameter, half the encoded information came from sperm and other half from egg. Thus, every crossover and fertilisation led to 2 offspring and resulting in total *2m* possible offspring. The genotypes of these resulting offspring were converted to phenotype using the parameters values given in Table1.

Although in biology, meiosis creates two sperm/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 genetic similarity of the population would start to be affected. We therefore only selected one of the pair of sperm/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, leading to: *xover1* results in offspring *(o1, o2)*; *xover2* results in *(o3, o4)* and *xover3* leads to offspring *(o5, o6)*. First, we randomly pick any one offspring out of the possible six; let us assume that is *o1*. Because of the reasons explained above, we discard *o1*'s corresponding offspring, *o2*. Next, the similarity of *o1* is checked with the remaining four offspring using the Hamming distance formula. The offspring that is at most fifty percent 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, any one is chosen randomly and replicated, and they comprise 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. In other words, 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.

Apply variation in the environment: an individual's EF was modelled by a randomly chosen number between 0.6-1.0. 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 ensures that all individuals were exposed to more than half of the training dataset. In accord with the *Equal*

Environments Assumption [32], twin pairs raised in the same family were assigned the same training subset.

ANN training and performance assessment: the population of twin ANNs was trained on the past tense dataset using the Rprop algorithm. The performance was assessed on the full training set, as well as on another novel dataset that was created to test the generalisation ability of the networks (see below). The continuous outputs produced by networks were converted to binary by applying threshold. Then the performance was assessed using recognition accuracy based on Hamming distance as explained in [17].

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). An important and novel aspect of our approach for 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 randomly a chosen member from that pool. After selection, only the offspring form the next generation of populations – parents (or members of previous/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 mainly because (i) an individual (parent) can only pass one copy of each gene (or intrinsic parameter) 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 [33]. Since, after getting selected in breeding pool, the members breed randomly, the best properties do not always get transferred effectively, since the advantageous gene may not be inherited. (ii) Although some traits are inherited from parents during reproduction, these inherited traits are tendencies and offspring inherit the predisposition to exhibit that behaviour. Most traits, however, are the product of a developmental process involving interaction with the environment – usually skills and behaviours that are acquired by experience in the organism's lifetime and make it compatible with its environmental and survival needs. Environmental traits are not transferred genetically from one generation to another. It is the combination of inherited and environmental traits that make each individual unique [12].

Repeat: the entire process was iterated until ANN parameters did not markedly change across generations or performance started to converge, i.e. the learning error reached a small value.

4. English Past Tense Dataset

The dataset was based on the "phone" vocabulary from Plunkett and Marchman [34] past tense model. The past tense domain was modelled 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 [34].

Artificial verbs were monosyllabic and were constructed from English phonemes. There were 508 verbs in the dataset. Each verb had three phonemes – initial, middle, and final. The phonemes were represented over 19 binary features. A network thus had 3×19 = 57 input units and 3×19 + 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.

In the training dataset, there were 410 regular and 98 irregular verbs. These were divided into four types: regular verbs that formed their past tense by adding */ed/* - e.g. visit – visited; regular verbs which formed their past tense by adding */d/* - e.g. tame – tamed, regular verbs which suffixed */t/* - e.g. clap – clapped, and finally the irregular verbs, e.g. hide – hid or go – went. In the dataset, out of 410 regulars, there were 271 */ed/* verbs, 90 */d/* verbs, 49 */t/* verbs. As this is an imbalanced dataset, generating a classifier is challenging as the classifier tends to map/label every pattern with the majority class.

A second dataset was also created to assess the generalisation performance of the model. The main intent was to measure 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 generalisation set comprised 508 novel verbs, each of which shared two phonemes with one of the regular verbs in the training set, for example *wug – wugged* [14, 43], i.e. generalisation set consists of novel regular verbs. This use of novel verbs is standard practice for generalisation testing in context of tense formation.

5. Experimental Design

In order to explore the behaviour of the model in different lineages, i.e. combinations of genetic and environmental influences, three replications of the model were tested, each having a twenty-generation duration. The experiments were conducted on Condor, which is a platform that supports running high throughput computing on large collections of distributive owned computing resources [40]. It follows a master-slave type configuration, which has proved suitable for training neural network architectures [27].

Each scenario was characterised by its own initial population (produced with random binary genomes) and unique values for the other heuristics involved, such as initial weights. The evolutionary methodology was then applied to each of these model instantiations, such 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. Thus, having three replications ($r1_,$ $r_{2,}$ *and* r_3) of the model aided in evaluating the robustness of the method.

For each generation, there were 50 pairs of DZ and 50 pairs of MZ twins with their computational parameters encoded into a genome. These were split in breeding and non-breeding, where the former is the population containing the 1st twin out of each of the twin pairs (100 networks) and the latter is the population containing the remaining 2nd twin of a twin pair (100 networks). These were instantiated as three-layered feed-forward networks and were trained using the batch version of the Rprop algorithm. The stopping condition was an error goal (mean squared error) of 10^{-5} within 1000 epochs. The networks were trained on the filtered training sets, but performance was always assessed on the full training set and then tested on the previously unseen generalisation set. The filter applied was based on EF values of each twin pair. These values represent the probability of including a particular data point (or training pattern) of the full training set into an individual's filtered training set. This varied between 60% and 100% so that each individual would come across at least half of the training set. Twin pairs had the same filtered training set. In order to breed twins, different crossover operators were employed like single point, multi point and more.

Table 2 summarises the settings used in the experimental design. Moreover, empirical data from young children performing the past tense task [14,43], were also used to benchmark the performance of the proposed model with respect to this age group, which has been the subject of considerable research in the literature.

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 Datasets	Training= 508 Generalisation= 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 (SES)	Probability value between 60% and 100%

Table 2: Experimental settings

6. Results & 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 generalise the past tense rule in novel items with the mean accuracy rate of around 60%.

The performance of our model compares well with empirical data for children reported in the literature [1, 47]. The behavioural data in [1] comprise of performance results of 442 6-year old children on past tense test. They were tested on 11 regular verbs and 8 irregular verbs. The average accuracy achieved by children on regular verbs is centred around 90%, whereas the average accuracy for irregular verbs is centred on 38%. It also agrees to a large extent with the performance reported in the developmental study of [47] for 5-7 year old children are: for regular verbs, accuracy rates are 60% (5 year olds), 75% (6 year olds) and 80% (7 year olds); for irregular verbs, accuracy rates are 25% (5 year olds), 58% (6 year olds) and 50% (7 year olds).

We compared our model's performance with two other past tense models from [43] and [14]. In the former model, 1000 networks were trained for 1000 epochs in various degrees of environmental and genetic variation scenarios. The experimental setting that 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 latter case, the model comprised of network 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 of 60%-80% in case of regular verbs and between 20%-40% for irregular verbs, achieved in the window of 51-70 epochs. Their model also achieved over 80% generalisation accuracy.

We also analysed the results, initially using independent linear regressions to assess performance / heritability / parameter changes for each population over the generations. Individually reliable trend lines at the p=.05 level are shown on the following figures. Given the overall design, which combined repeated measures (e.g., regular verb performance, irregular verb performance, generalisation) and between group measures (replication population; breeding vs. non-breeding populations), trajectory analysis was used to assess overall patterns in the component linear regressions [44].

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 summarise the results from 12,000 networks. Figure 2 shows equivalent data for irregular verbs, while Figure 3 represents the generalisation 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 relatively small, reflecting the non-monotonic 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 behaviour of mice [8, 32]. A red star in these figures indicates replications wherein patterns emerging were most statistically significant, as discussed below.

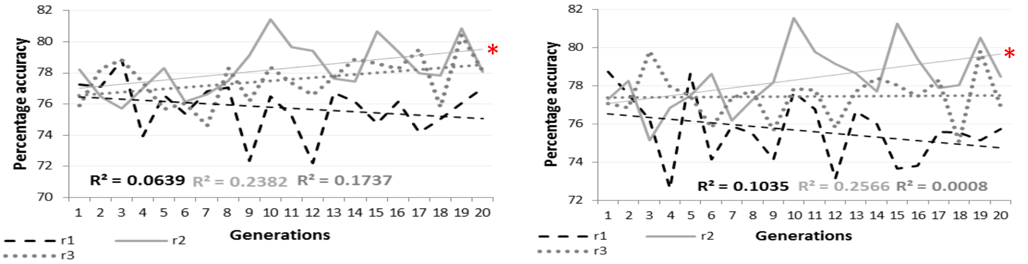

Figure 1: Mean performance per generation for breeding (left) and non-breeding (right) twin populations on regular verbs

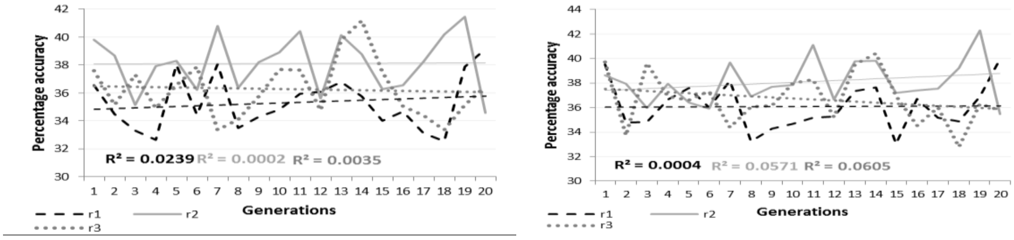

Figure 2: Mean performance per generation for breeding (left) and non-breeding (right) twin populations on irregular verbs

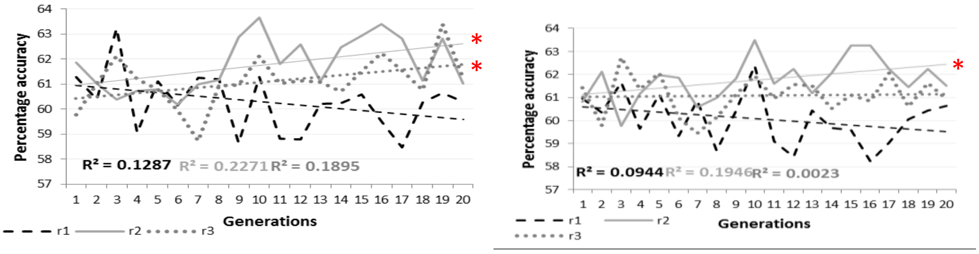

Figure 3: Mean generalisation accuracy per generation for breeding (left) and non-breeding (right) twin populations

We initially considered performance of application of the past tense rule, comparing the measures of regular verb performance against generalisation, 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 measures and breeding/non-breeding populations. Regular verb performance was reliably higher than generalisation ($F(1,108)=6288.30$, $p<.001$, $\eta_p^2=.983$).

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 to 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 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 are observed for regular verbs and irregulars (with generalisation patterning with regular verbs). When performance across generations is worsening for regular verbs, it is improving for irregular verbs, and vice versa. Because the learning domain of 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 optimising the same neuro-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 with regular mappings requiring steeper sigmoid functions and irregular mappings requiring more hidden units [46]. The combination of (a) selection by mean performance that could either be driven by stronger regular or irregular verb performance, and (b) parameters that favour 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 specialised for regular verbs at the expense of irregular verbs, while other lineages may become specialised 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.

This phenomenon is similar to Waddington's epigenetic landscape, an idea proposed by Conrad Waddington ([11], pp R459). In his model, Waddington associated 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 [48]. This is what makes an initial pluripotent cell to become a specialised cell, and reversing this process is impossible under normal circumstances.

Similarly, when selection is applied on a quasi-regular task, different aspects of the task may be optimised depending on the genetic propensities of initial populations. The trend then continues throughout the lineage because of genetic inheritance. Thus, if, as shown in lineage 1 (replication 1) in Figures 1 and 2, the first few generations improve their learning of irregular verbs at the expense of regular verb performance, the lineage is committed to this pathway. Genes for good learning of regular verbs have been lost from the gene pool. Evolution cannot go into reverse gear and find a pathway that combines good learning on both verb types. Replication 3 represents the opposite case of optimisation on regulars, while replication 2 shows improvement in both verb types across generations.

Changes in the frequency of different gene variants (here, binary values of 0 or 1) in the gene pool should alter the range of genetic variation across generations. Given that the range of environmental variation (EF of 0.6 to 1.0) remained consistent across generations, any changes in genetic variation should be reflected in changes in heritability. To explore this idea, we examined correlations in performance between MZ and DZ network twin pairs, using Falconer's equations to derive estimates of heritability [32]. 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. Although in our results the correlations sometimes violated this condition, we continue to plot heritability estimates according to the same formulae for consistency. Therefore, the plotted data should be seen as proportion to the heritability and environmentability observed in populations, rather than direct estimates under an additive model. Thus, the values sometimes range outside of the range 0 to 1, as the assumptions of the additive model become violated.

Figure 4 shows the estimates of heritability (variance due to genetic factors) for regular (4a) and irregular verbs (4b). 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 measure or replication population. Though replication 2 showed the steepest reduction in heritability, the difference in the pattern across replications was not reliable ($p=.107$).

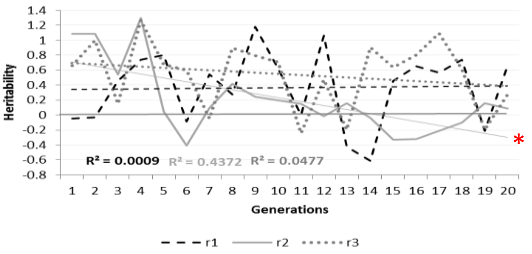

Figure 4(a): Heritability or proportion of variance due to genetic (or structural) factors for Regular Verbs. The red star indicates a replication wherein patterns emerging were most statistically significant.

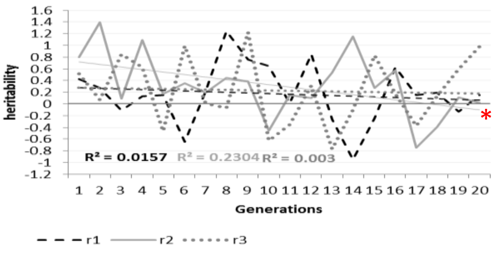

Figure 4(b): Heritability or proportion of variance due to genetic (or structural) factors for Irregular Verbs. A red star indicates a replication wherein patterns emerging were most statistically significant.

If a lineage becomes increasingly optimised on a task (or a specific aspect of the task domain), the range of its domain-relevant intrinsic parameters should decrease across generations, as only the genes producing the best parameter values are retained. For example, if populations are improving on irregular verbs, which require more capacity to hold non-systematic mappings, then across generations, networks with larger number of hidden units have a greater chance to get selected in the breeding pool. Across generations, the variability in the range of number of hidden units will reduce. By contrast, the range of variation in other less relevant parameters may be less affected. *Optimisation and heritability should therefore have an inverse relationship*.

In line with this expectation, in replication/lineage 1, regular verb performance and rule generalisation dropped across generations while irregular verb performance improved. Heritability for regular verbs was initially higher than that for irregular verbs, centred on 0.4 and it then increased across generations, implying lack of selection for parameter sets specialised for regularity. By contrast, heritability of irregular verbs was lower, centred on 0.2, and decreased with generations, implying selection for, and a narrowing of the range of, parameter sets specialised for irregularity. Note that this process of specialisation causes *overall* accuracy to drop, because irregular verbs form a minority of the dataset (there are only 98 irregular verbs compared to 410 regular verbs).

In replication/lineage 2, regular verb performance, irregular verb performance, and generalisation all increased across generations. Heritability of regular verbs dropped from high values of around 0.8 to around zero. A similar pattern was observed for irregular verbs, with heritability dropping from high values to almost nil. In this lineage, optimisation 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 generalisation 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, but at a lower value, centred on 0.2. These two observations suggest that the range of intrinsic parameters being targeted by selection works well for both regular 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 variance in performance is less due to genetic factors and more due to shared and non-shared environmental factors. Figures 5(a) and 5(b) display the variance due to shared environmental factors, in this case the filtered training datasets. 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 X generation (F(2,54)=3.65, $p=.033$, $\eta_p^2=.119$). The pattern was common across regular and irregular verbs. Figures 5(c) - 5(f) confirm that the range and mean level of EF was constant across generations for all lineages.

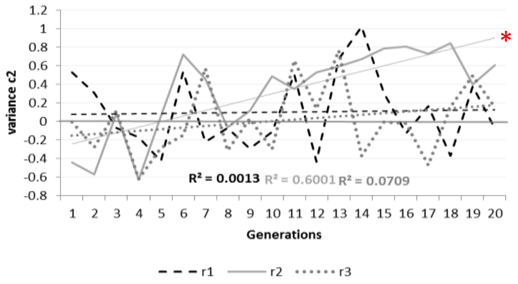

Figure 5(a): Proportion of variance due to shared environmental factors - Regular Verbs

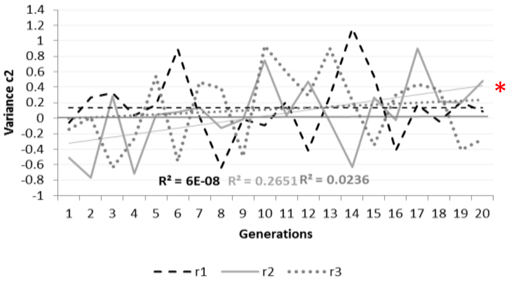

Figure 5(b): Proportion of variance due to shared environmental factors - Irregular Verbs

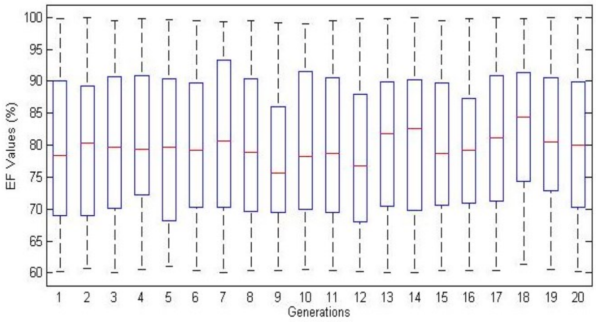

Figure 5(c): Percentage of EF per generation replication1

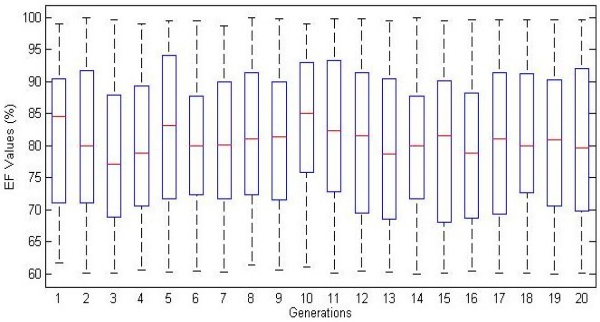

Figure 5(d): Percentage of EF per generation replication2

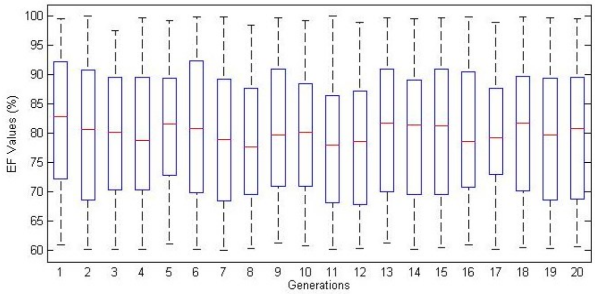

Figure 5(e): Percentage of EF per generation replication3

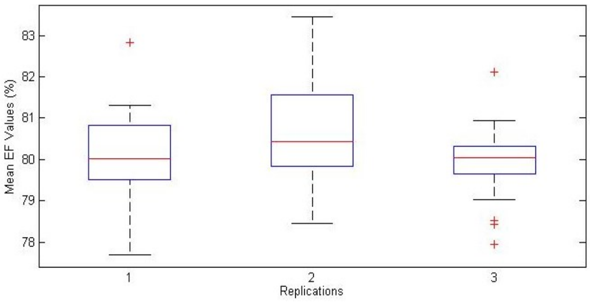

Figure 5(f): Percentage of EF per replication

Figures 6(a) and 6(b) represent 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 modulation neither by measure type nor by replication population. The figures show that the differences in initial weights led to large variability in behavioural outcomes. In cases when intrinsic factors were not very suitable to the task domain, having good initial weights might give networks a fighting chance, i.e. training could be biased towards non-shared environmental factors to enhance behavioural performance.

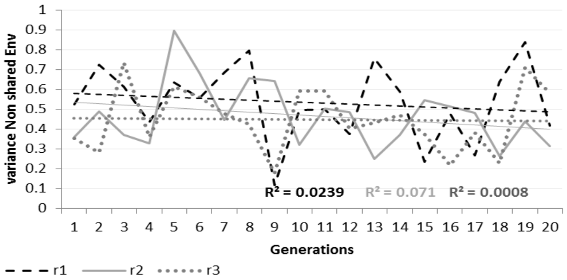

Figure 6(a): Proportion of variance due to non-shared environmental factors - Regular Verbs

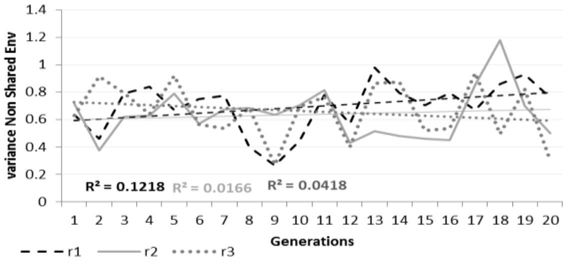

Figure 6(b): Proportion of variance due to non-shared environmental factors - Irregular Verbs

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 behaviours, 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 therefore in turn reveal the drivers behind changes in heritability.

Figure 7 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 generation (F(1,54)=190.55, p<.001, η_p^2=.779), with the reduction occurring at different rates across the three replication populations (F(2,54)=33.79, p<.001, η_p^2=.556).

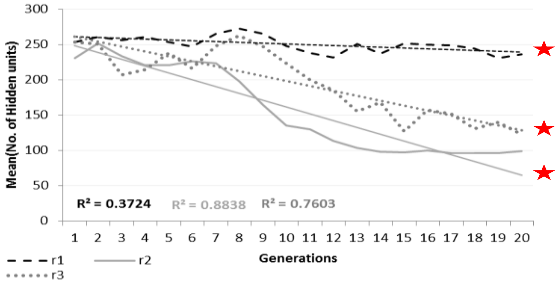

Figure 7 (a): Change in the mean value of the number of hidden units per generation

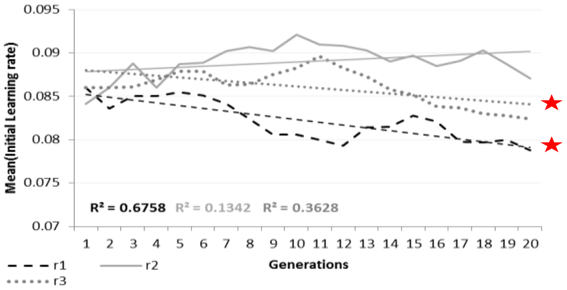

Figure 7 (b): Change in the mean value of the initial learning rate per generation

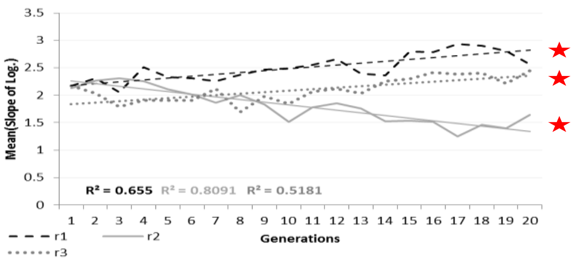

Figure 7(c): Change in the mean value of the slope of logistic activation per generation

For learning rate, the same pattern was observed, with an overall pattern of reduction across generations (F(1,54)=22.69, p<.001, η_p^2=.296) modulated by replication, with the reduction appearing in only two of the three replications (F(2,54)=12.22, p<.001, η_p^2=.312). Lastly, for slope of logistic activation, a differential pattern also emerged, this time with 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, η_p^2=.325; interaction of generation*replication: F(2,54)=61.06, p<.001, η_p^2=.693). Overall, replication 1 and 3 showed a common pattern of reduction in hidden units, reduction of learning rate, and increase in temperature. For replication 1, the reduction in hidden was milder, the learning rate fell lower, and the temperate rose higher. Replication 2 showed a different pattern of a greater fall in hidden units, no change in learning rate, and a drop in temperature.

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 category of non-systematic mappings, which are more demanding on computational capacity. Figure 8 depicts the variation in the ranges of the three parameters across generations. It thus reflects the parameters being targeted by selection in each lineage.

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

Regular verbs have systematic input-output mappings, which are less demanding on computational capacity. Lineage/replication 2 improved regular performances at the expense of irregular verbs, 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 more sympathetic to accommodating irregular mappings.

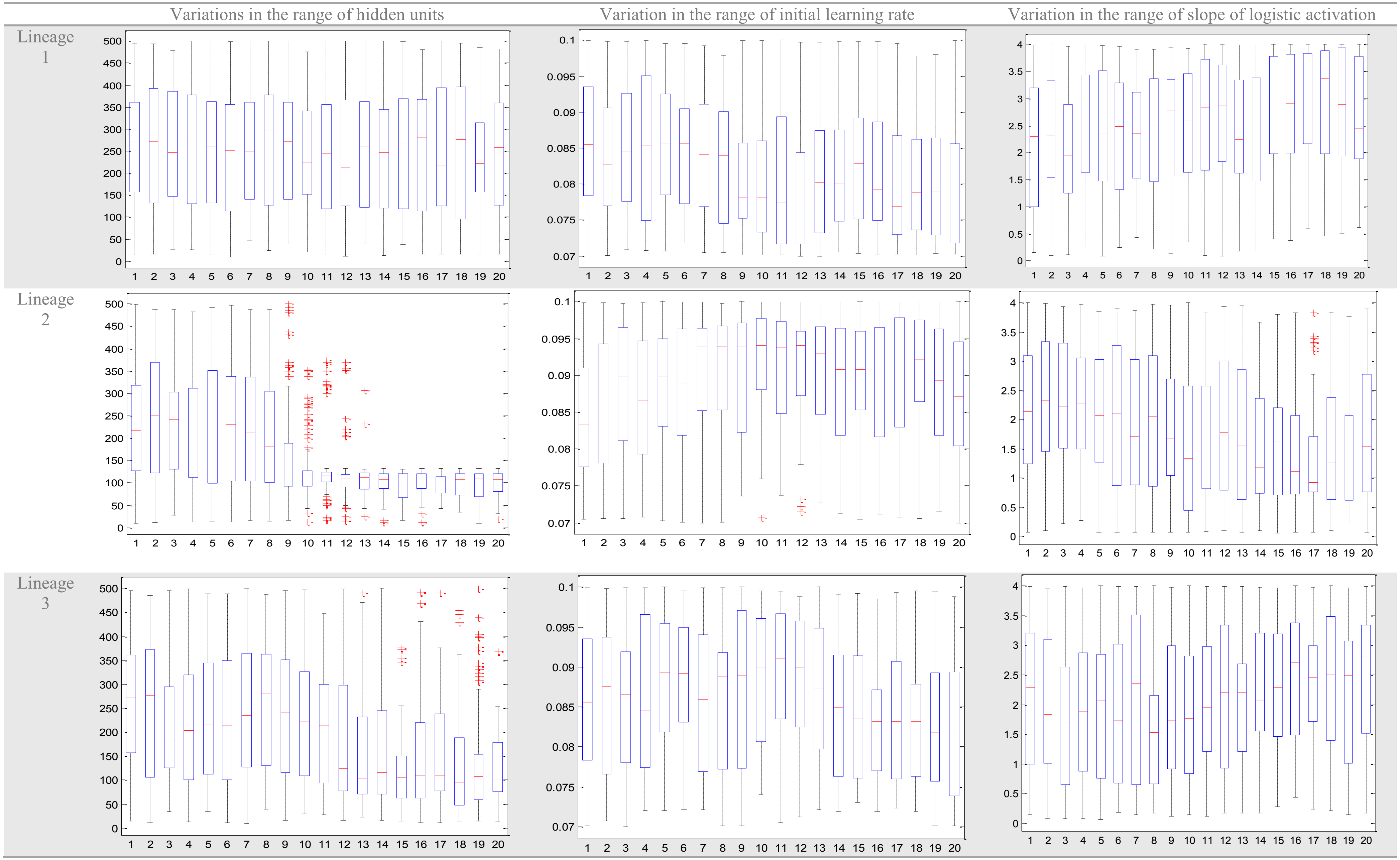

Figure 8: Range of Variation of Intrinsic parameters across Generations

7. Findings

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 behaviours depending on initial conditions – both genetic and environmental. (ii) Once selection starts targeting a particular aspect of task domain, it starts behaving similar to Waddington's epigenetic landscape. That is, from an initial pluripotent state, the developmental (or learning) pathway of populations (in that lineage/replication) becomes more specialised in the particular targeted aspect. Reversing this trend is difficult, if not impossible. (iii) Selection based on a stochastic method such as roulette-wheel, when combined with sexual reproduction method for population generation, has a limiting effect on final behavioural (or performance) levels achieved. Performance is affected in two ways: first, since roulette-wheel selection has a stochastic nature, there is a possibility of not-so-fit members being selected in the breeding pool. Secondly, the sexual reproduction method used to generate offspring prevents reliable transfer of best properties from parents to offspring. (iv) Heritability acts as an identifier of the aspect of the quasi-regular task being targeted by selection. Highly heritable behaviour indicates that the trait is not being selected for, whereas behaviour with low heritability implies selection and optimisation. Thus an inverse relationship exists between heritability and optimisation. (v) A higher proportion of variance caused due to shared environmental factors (filtered training sets) is an indicator of good or gifted learners. In other words, it shows that the particular population members have greater genetic predisposition of successfully acquiring the desired behaviour or task. (vi) Non-shared environmental factors (initial weights) lead to significant proportions of behavioural variance. This effect becomes magnified when intrinsic properties are not particularly suitable. In such cases, having good initial weights can provide networks with the extra support needed to acquire a task. Hence, training could be biased towards non-shared factors to improve performance.

8. Conclusion

In this paper, we introduced a novel computational approach inspired from principles of behavioural genetics to model the performance of 6-year-old children on English past tense acquisition. We analysed the proportion of variance accounted for by ANN computational parameters (or 'genes') and filtered training sets and initial weights (or 'environment') and also highlighted the importance of selection and sexual reproduction methods. Our model was able to identify the causal factors leading to behavioural or performance variability within that population and the factors responsible for that variability. Most importantly, our model showed that divergent behavioural outcomes can emerge when selection is applied on a quasi-regular task. Learning pathways start to behave like Waddington's landscape, becoming more specialised in one aspect of the task across generations. The model also showed that heritability and optimisation have an inverse relationship, with heritability identifying which aspects of the task domain are being targeted by selection. ANN training can therefore be biased towards specific aspects based on selection results to get desired performance (behavioural) outcomes. There are several avenues requiring further investigation to establish long-term trends and emerging behaviour as the populations evolve. More complex genome representations, for example, may allow encoding more computational parameters and increasing genetic variability. Also, it is necessary to understand the impact of employing alternative selection schemes on population divergence, as well as the implications for non-random assignment of environments to genotypes implied by gene-environment correlations believed to hold in human populations [33].

REFERENCES

1. 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, (pp. 479 – 490). doi:10.1007/s10519-004-1834-7
2. Bottou, L.Y. (1988): Reconnaissance de la Parole par Reseaux Multi-Couches. In Neuro-Nimes '88; Proceedings of the International Workshop on Neural Networks and Their Applications, (pp. 197–217). ISBN: 2-906899-14-3
3. Chater, N., Christiansen, M.H. (2008). Computational models of psycholinguistics. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 477-504). Cambridge University Press. Doi:10.1017/CBO9780511816772.021
4. Chomsky, N. (1965). Aspects of the theory of syntax. MIT Press, Cambridge.
5. Chomsky, N. (1980). Rules and representations. Behavioural and Brain Sciences (3). (pp. 1-61).
6. Cooper, G.M. (2000). The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates. Meiosis and Fertilization. Available: http://www.ncbi.nlm.nih.gov/books/NBK9901/
7. Cottrell, G. W. & Plunkett, K. (1991): Learning the past tense in a recurrent network: Acquiring the mapping from meanings to sounds. In Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society (pp. 328-333). Hillsdale NJ: Lawrence Erlbaum Associates.
8. 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, Volume 8, Issue 1 (pp 3-13).
9. Daugherty, K., Seidenberg, M.S. (1992): Rules or connections? The past tense revisited. Proceedings of the 14th Annual Meeting of the Cognitive Science Society (pp. 259-264). Hillsdale, N.J.: Erlbaum.
10. Falconer, D.S., Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. 4th Edition, Longman, Esex.
11. Ferrell, J.E., (2012). Bistability, Bifurcations, and Waddington's Epigenetic Landscape, In Current Biology , Volume 22 , Issue 11, (pp R458-R466).
12. Griffiths, P. The Distinction Between Innate and Acquired Characteristics (2009). *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/innate-acquired/>.
13. Gong, T., Shuai, L. (2013). Computer simulation as a scientific approach in evolutionary linguistics. *Language Sciences,* 40, (pp. 12-23).
14. Karaminis,T., Thomas, M.S.C. (2010): A cross-linguistic model of the acquisition of inflectional morphology in English and modern Greek. Proceedings of the 32ndAnnual Meeting of the Cognitive Science Society (pp. 730-735).
15. Karaminis, T., & Thomas, M. S. C. (2015). The Multiple Inflection Generator: A generalized connectionist model for cross-linguistic morphological development. *Manuscript submitted for publication.*
16. 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, (pp. 969-990).

17. Kohli, M., Magoulas, G.D., and Thomas, M.S.C. (2012): Hybrid Computational Model for Producing English Past Tense Verbs, In Proc of the 13th International Conference on Engineering Applications of Neural Networks (EANN), 20-23 September 2012, London, C. Jayne, S. Yue, and L. Iliadis (eds.), Springer CCIS 311 (pp. 315-324).
18. Kohli, M., Magoulas, G.D., Thomas, M.S.C. (2013): Transfer learning across heterogeneous tasks using behavioural genetic principles, Proceedings of the 13th UK Workshop on Computational Intelligence (pp. 151-158).
19. Lupyan, G., McClelland, J.L. (2003): Did, Made, Had, Said: Capturing quasi – regularity in exceptions. 25th Annual Meeting of the Cognitive Science Society.
20. MacWhinney, B. (1998). Models of the emergence of language. Annual review of Psychology 49 (pp. 199-227).
21. MacWhinney, B. (2004b). A unified model of language acquisition. In: Kroll, J., DeGroot, A. (eds.) Handbook of bilingualism: Psycholinguistic approaches. Oxford University Press, Oxford.
22. Mareshal, D., Thomas, M.S.C. (2007). Computational modelling in developmental psychology. *IEEE transactions on Evolutionary Computation*,11 (pp. 137-150). Doi: 10.1109/TEVC.2006.890232
23. 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, (pp. 311-339).
24. Pinker, S. (1984): Language learnability and language development. Cambridge, MA: Harvard University Press.
25. Pinker, S. (1994). *The language instinct: The new science of language and mind* (Vol. 7529). Penguin UK.
26. Plagianakos, V. P., Magoulas G. D. and Vrahatis M. N. (2006). Evolutionary training of hardware realizable multilayer perceptrons, Neural Computing & Applications, vol. 15(1) (pp. 33-40).
27. 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, vol. 81(3) (pp. 228-235).
28. 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 (pp.56-115).
29. Plomin, R. & DeFries, J.C. (1980). Genetics and Intelligence: recent Data. Intelligence (4). (pp.15-24).
30. Plomin, R. (1990). The role of inheritance in behaviour. Science, 248 (pp.183-248).
31. Plomin, R., & Spinath, F. M.: Intelligence: Genetics, genes, and genomics. (2004). Journal of Personality and Social Psychology, 86, (pp. 112-129).
32. Plomin, R., DeFries, J. C., & McClearn, G. E. (2008): Behavioral Genetics: A primer. San Francisco: Worth and Freeman.
33. Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2013). *Behavioral genetics* (6th ed.). New York, NY: Worth.
34. Plunkett, K., Marchman, V. (1991): U-shaped learning and frequency effects in a multilayered perceptron: Implications for child language acquisition. Cognition, 38 (pp.43-102).
35. Plunkett, K., & Marchman, V. (1993): From rote learning to system building: Acquiring verb morphology in children and connectionist nets. Cognition, 48 (pp.21-69).

36. 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 (ICNN). San Francisco. (pp.586-591).
37. Rumelhart, D. E., & McClelland, J. L. (1986): On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models (pp.216-271). Cambridge, MA: MIT Press.
38. Simon, H.A. (1962). The architecture of complexity. Proceedings of American Philosophical Society, 106 (pp.467-482). (Reprinted in Simon, H.A. (1969). The sciences of the artificial. Cambridge, MA: MIT Press)
39. Smith, A. D. M. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science,* **5**:10.1002/wcs.2014.5.issue-3, (pp.281-293).
40. Thain, D., Tannenbaum, T., Livny, M. (2005). Distributed Computing in Practice: the Condor Experience. Concurrency and Computation: Practice and Experience 17(2-4) (pp.323-356). Doi: 10./1002/cpe.938
41. Thimm, G., Fiesler, E. (1995): Neural network initialization. From Natural to Artificial Neural Computation. Lecture Notes in Computer Science Volume 930 (pp.535-542).
42. Thomas, M.S.C., McClelland, J.L. (2008): Connectionist models of cognition. In R. Sun (Ed.), Cambridge handbook of computational cognitive modelling (pp.23-58). Cambridge University Press.
43. 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 Tech report 2009-1.
44. 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.
45. Thomas, M.S.C., Forrester, N.A., Ronald, A. (2013): Modeling socioeconomic status effects on language development. Developmental Psychology, Vol 49(12) (pp.2325-2343).
46. Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2015). Multi-scale modeling of gene-behavior associations in an artificial neural network model of cognitive development. *Manuscript submitted for publication.*
47. van der Lely, H. K. J., & Ullman, M. (2001). Past tense morphology in specifically language impaired children and normally developing children. *Language and Cognitive Processes, 16,* 177-217.
48. Waddington, C. H. *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.* (1957).London: George Allen & Unwin, Ltd.
49. Wintner, S. (2010). Computational models of language acquisition. In *Computational Linguistics and Intelligent Text Processing* (pp.86-99). Springer Berlin Heidelberg.
50. Yam, J.Y.F, Chow, T.W.S. (2000): A weight initialization method for improving training speed in feedforward neural network. Neurocomputing 30. (pp.219-232)