The Difficulty of the Baldwinian Account of Linguistic Innateness

Hajime Yamauchi

Language Evolution and Computation Research Unit,
Department of Theoretical and Applied Linguistics,
The University of Edinburgh, Edinburgh, UK
hoplite@usa.net

Abstract. Turkel [16] studies a computational model in which agents try to establish communication. It is observed that over the course of evolution, initial plasticity is significantly nativised. This result supports the idea that innate language knowledge is explained by the Baldwin effect [2][14]. A more biologically plausible computational model, however, reveals the result is unsatisfactory. Implications of this new representation system in language evolution are discussed with a consideration of the Baldwin effect.

1 Introduction

For decades, the innate capacity of language acquisition has been one of the central issues of the study of language. How heavily does language acquisition rely on innate linguistic properties? This question, often called the *'nature & nurture problem'*, brings endless debates in linguistics and its adjacent fields. Indeed, a number of phenomena that occur during language acquisition are quite puzzling when one tries to determine what parts of language acquisition are innate or attributed to postnatal learning. An intensive array of studies has gradually revealed that this twofold structure of language acquisition never appears as a clear dichotomy. Rather, the intriguing interaction between innate and learning properties of language acquisition seems to require a new avenue of linguistic studies.

1.1 The Baldwin Effect and Language Evolution

James Mark Baldwin [2] assumed that if an individual is capable of acquiring an adaptive behavior postnatally, addition of such a learning process in the context of evolutionary search potentially changes the profile of populational evolution; the learning paves the path of the evolutionary search so that evolution can ease its burden of search. In addition, this special synergy of learning and evolutionary searches has a further effect, known as 'genetic assimilation' [18]. This is a phenomenon in which "a behavior that was once learned may eventually become instinctive" [17].

J. Kelemen and P. Sosík (Eds.): ECAL 2001, LNAI 2159, pp. 391–400, 2001.

Then this learning-guided evolution scenario, known as *the Baldwin effect*, possibly provides a strikingly attractive perspective to the *nature-nurture* problem in linguistics. It has been attested by a number of computer simulations in the field of computer science that if an environment surrounding the population is prone to shift to a new environment, some part of the behavior is better preserved for learning. If those environments do not share any commonality, an individual who relies in every aspect of behavior on learning will be the most adaptive. However, if those environments hold some universality, an individual who has partially nativised and partially learned behavior will be the most adaptive; for example, the nativised part of the behavior covers the universality and the learned part of the behavior covers the differences. Consider this in the case of language evolution. The whole human population is well divided into a number of sub-populations in many aspects; races, cultures, and so forth. Boundaries of language diversities often coincide with those of the sub-populations. Then, for children, it is a great advantage to keep some part of the linguistic knowledge for learning while the other is innately specified. This helps the child even if he is reared in a different linguistic society from his parents; he still may acquire the society's language. Therefore, the *nature-nurture* problem in linguistics can now be considered in the context of the evolution of language. Universality of the world's languages may correlate to the evolution of nativised linguistic knowledge while linguistic diversities are correlated to learning. Since this universality-*nature*, diversity-*nurture* correlations are perfectly compatible with Chomsky's *L*anguage *A*cquisition *D*evice theory [4], and as the Baldwin effect and the LAD theory both involve genetics, the study of the Baldwin effect in the domain of LAD becomes particularly appealing.

The Baldwin effect in linguistics may also provide an attractive solution for a long-standing problem. Preliminary studies suggest that language evolution is out of the scope of natural selection mainly because of its dysfunctional nature. For those researchers, language evolution is a consequence of exaptation or a big leap in evolution [13]. This no-intermediate scenario would be, however, explicable by natural selection when it is guided by learning since learning can smooth the no-intermediate landscape. Subsequently, it has been a popular idea that the Baldwin effect is a crucial factor in the evolution of language (e.g., [14][16]).

1.2 The Principles and Parameters Approach

Given its logical complexity, researchers agree that linguistic input is the most important ingredient of language acquisition. Counter-intuitively, however, such vital linguistic input employed to construct knowledge of a language is importantly often insufficient [3]. In other words, children have to acquire their target languages under qualitatively and quantitatively insufficient circumstances. Absence of "Negative Evidence" in language acquisition is one of the clearest examples of this. As a part of the insufficiency, usually children are not provided negative feedback for their grammatical mistakes while such information is vital for any second language learners.

To reduce this complication, Chomsky has claimed some special synergy of innate linguistic knowledge and the acquisition mechanism is required. The basic concept of his original formulation of the nature of language acquisition, called Principles & Parameters theory, [6] is as follows. In the P&P approach, two types of limited innate linguistic knowledge are accessible, called ‘principles’ and ‘parameters’. Principles are universal among all natural languages and considered as genetically endowed. Parameters are partially specified knowledge which are encoded in binary parametric values. Setting of each parameter is triggered by post-natal linguistic experiences. We can conceive the possible mechanism of the LAD as an incomplete learning device in which certain binary information is missing

2 Implementation of the LAD in Dynamic Systems

The combination of genetically hardwired features and postnatal learning processes in the Baldwin effect is perfectly compatible with Chomsky’s P&P theory of the LAD. Together with its “genetic assimilation” process [18], the Baldwin effect may shed light on the nature of the current relationship between innateness and postnatal learning in language acquisition.

Precisely because of this compatibility it is crucial to pay careful attention to the implementation of the P&P approach in a genetic search. Given an assumption that the LAD is one of the most elaborated cognitive abilities, it is highly unlikely that such ability is DIRECTLY coded in the genes. Rather it is more plausible to assume that linguistic innateness relies on some degree of polygenic inheritance [1].

More specifically, principles and parameters are not coded by a simple concatenation of genes. Rather a *combination* of those genes expresses one principle/parameter. This genetic mechanism is called “epistasis”. Epistasis is a situation in which the phenotypic expression of a gene at one locus is affected by the alleles of a gene at other loci. Pleiotropy, in a very crude form, means that one gene contributes to express more than one phenotypic character. Thus, one gene in the model will affect an expression of one phenotypic trait, but also will determine other traits.

In the next section, we examine the effect of the two phenomena in the study of the evolution of the LAD.

3 The Experiments

To test the effect of epistasis and pleiotropy on the Baldwin effect, we conducted two different types of simulations. The basic part of our model is adapted from the study of Turkel [16] to appear). First, an exact replication of Turkel’s simulation was tested. Then modified versions were tested. In those modified simulations, Stuart Kauffman’s [11] NK-Landscape model was introduced to implement epistasis and pleiotropy. The specific explanation of NK-Landscape in these simulations is given later. Here the basic structure of the model is explained. In Kauffman’s NK-Landscape model, unlike ordinary GA models where one gene expresses one trait of a phenotype, a SET of genes determines one trait of a phenotype. In other words, one specific part of the phenotype (a phenotype consists of 12 traits in this simulation) may be decided by two or more distinctive genes. How many genes are required to express one trait is specified in the value of K. The values of K are always between 0 and N-1 where N designates the number of the genes. Dependency of genes is either contiguous or non-contiguous. In the case of contiguous dependency, a gene forms a concatenation with other adjacent genes. Note that in the contiguous dependency case, which we employ in this paper, both ends of a chromosome are considered as neighbors of each other so that K-dependency of phenotypes is available in all loci.

In terms of evolutionary search, the increase of the value of K toward N means that the fitness landscape becomes increasingly rugged. In a rugged landscape, evolutionary search tends to be trapped in local optima. The correlation between the fitness and similarity of genotypes (typically measured by Hamming distance) is also kept low in the landscape. Therefore, an identical phenotype of two agents does not guarantee for them to have an identical genotype. In a simulation using this model, a look-up table is created at the beginning of the simulation. The size of tables corresponds to N times 2^K since each allele is affected by 2^K possible combinations of other genes.

In the next section, we look at the result of Turkel's original study, then make a comparison to our obscured phenotype model. All results of these simulations are averages of 100 runs.

3.1 Simulation1: Replication of Turkel

Based on Hinton & Nowlan's simulation [10], Turkel conducts an experiment that holds a populationally dynamic communication system. While Turkel mostly adopts Hinton & Nowlan's genetic encoding method (fixed, and plastic genes), he provides an external motivation for it according to P&P approach. Turkel considers those fixed genes —0s and 1s— as '*principles*', and the plastic genes —?s— as '*parameters*'.

The algorithm of Turkel's simulation is quite straightforward and mostly intuitive. Most parts of the algorithm are quantitatively the same as Hinton & Nowlan; initially 200 agents are prepared. The ratio of 0:1:? In Turkel is different in his four different configurations of simulations —2:2:8 (High-plasticity), 4:4:4 (Equal ratio), 3:3:6 (Original), and 6:6:0 (No-plasticity)— respectively. Distribution of these genes in an individual agent is randomly decided initially. In the initial population, generally there is no case that two agents hold the same genotype. The reproduction process includes one-point crossover with 20% probability. Considering the spirit of GA, it is somewhat odd but mutation is not included [10] mutation was not included also). Two agents (one is selected from 1st agent to 200th in order, and its partner is randomly selected) compare their genotype. If those two agents' genotypes are exactly the same pattern including loci of ?s, the first-chosen one is assigned 2 fitness points. If the agents do not exactly match but those no-matching alleles have 0-? Or 1-? Combinations, they are considered as potentially communicable. Then they are sent to learning trials. By changing all ?s into either 1s or 0s randomly, the two agents attempt to establish communication within 10 trials. If the agents succeed to possess exactly the same phenotype within ten trials, communication is considered to be established. In each trial, the agents reset their phenotype and express new phenotypes from their genotypes. During the learning process, learning cost is introduced implicitly. The size of decrement per trial is 1 from the highest fitness value of 12. The range of the fitness values is, thus, from 12 (immediate success) to 1 (complete failure). If two agents have any 1s and 0s combination in the same allele, they are assigned the fitness value of 1 since it would be impossible to establish communication.

In our replication experiment, we choose Turkel's "Original" configuration where the number of ?s is 6 and the number of both 1s and 0s are 3 each.

The result obtained from our simulation was, as expected, almost identical to Turkel's original simulation. Fig. 1 shows the average number of 0s, 1s, and ?s in the evolved population.

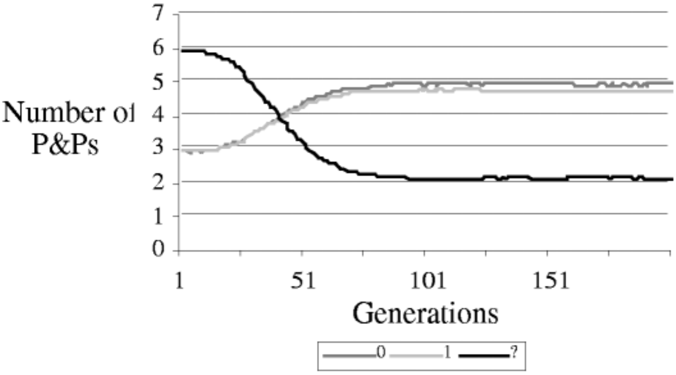

Fig. 1.

In the figure, a steep descent of ?s is observable in an early period. Once the population reaches the "plateau" condition, no further change takes place. On the plateau, virtually all agents have one unified genotype. The reason for this is the lack of mutation; the one-point-crossover reproduction process does not produce any turbulence under the unified situation.

It is often the case that before the Baldwin effect eliminates all plastic genes, a population reaches this plateau. This was especially salient in his preliminary studies where populations were more plastic. In those situations, the Baldwin effect did not have enough space to enjoy its power; before doing so, the populations typically converged to one genotype. Thus, at the end of each run, a comparatively large number of plastic genes remained, although the number of plastic genes was fewer than in the initial populations in almost all cases. To make it clearer, consider the following points. First, when the entropy of genotypes in a population is high (as in an initial period), high plasticity is advantageous; the more plastic, the more chance an agent has of proceeding to the learning trials. On the contrary, a fixed agent suffers great difficulty in this kind of situation; the fitness value is most likely 1 since the chance of exact match is extremely slim.

Too much plasticity cannot increase the actual fitness value either. Although highly plastic agents can often potentially communicate with other agents, the actual probability of establishing communication is quite low as the number of possible 0 and 1 combinations increases exponentially.

If the agent fails to establish a communication, the fitness value is 1. Thus, although it is somewhat contradictory, the best strategy to maximise fitness value is to keep the number of parameters as small as possible. It effectively means increasing the chance of establishing communication within 10 learning trials. To do this, it is necessary to reduce the number of plastic genes —genetic assimilation. Genetic assimilation, however, increases the number of fixed genes. Since the penalty for discrepancy of fixed genes on the same locus is most fatal (one Hamming distance is enough), this elimination process has to be done by increasing the identical genotype except in the loci of plastic genes. In other words, low plastic agents have to make sure that they meet either agents who have exactly the same genotype or all-the-same-but-partially-plastic agents. This turns out to be a selective pressure toward a uniform genotype. Therefore, genetic assimilation must intrinsically go hand in hand with convergence to identical genotypes. Importantly, however, these two processes are quasi-independent processes; although the force of both pressures comes from natural selection through the reproduction process, genetic assimilation is required from the learning trial *per se* while the convergence pressure comes into the place by more general requirement, “parity”. As noted above, when two agents are compared their pre-learning phenotype (= genotypes), discrepancy of principles is strongly malign —even with one discrepancy in their principles, the two agents have no possibility of establishing communication— while parameters always match with any principles or other parameters. As long as any loci that have principle-principle pairs match, an agent can have any number of parameters on any locus; although a lot of parameters indeed decrease the chance of communication but never reduce the chance completely while discrepancy between principles extinguishes it. In this regard, parameters are more benign than principles. Thus the pressure of convergence is generally greater than that of genetic assimilation. Since the pressure of convergence drives the agents to align their genotypes, consequently the population typically converges into a single genotype before complete genetic assimilation takes place. This is the reason why when the population is highly plastic, the absolute number of plastic gene remains higher than in a population.

3.2 Simulation2: Implementation of NK-Landscape Model

Our next simulations incorporate the NK-Landscape models while most of Turkel’s algorithms are untouched. A brief description of the simulation is given.

First, we determine the number of gene dependency regarding the expression of the phenotype. K designates the number. The value of K is fixed within a simulation; the same value is always applicable to any locus (this means that at any locus, the degree of gene dependency is not affected), any agents, and any generation. Since the range of K is from 2 to N-1, the maximum value is 11 (N=12) in these simulations.

Then, we prepare 200 agents. All agents consist of 12 genes. This time, instead of the three types of genes —0, 1, ?— only two types of genes exist, namely 0 and 1. Thus, at this level, there is no plasticity. These genes are equally shuffled into the 12 loci. The number of the two types of genes are the same in one agent, 6 each. These 12 genes are randomly distributed into 12 loci.

Thirdly, a look-up table is generated. This table correlates a genotype and phenotype. Below, an example is provided (Table 1).

Table 1.

	000	001	010	100	011	101	110	111
Locus1	0	**?**	1	1	?	?	0	?
Locus2	?	?	?	0	1	1	0	?
Locus12	?	1	1	0	?	?	0	?

The number of rows corresponds to the number of loci —12. The number of column corresponds to the number of possible combinations of genes. If K=3, the number of column is 2^3. To project a principle/parameter in the first position of a phenotype, we have to check the first row –"Locus1". If three genes from the first locus are 0, 0, 1, respectively in the genotype, we put ? in the first position of the phenotype (the cell in the table is emphasized). To project a principle/parameter in the second position of the phenotype, the second row is referred to. At the end of this projection process, the phenotype contains 12 principles/parameters in total. This is compatible with Turkel's genes. To make the simulations comparable to the former simulation, the ratio of 0, 1, and ? is set as 1:1:2. This is done by controlling the ratio of 0, 1, ? in look-up tables. Once this process is done, the rest of the simulation is exactly the same as Turkel's.

Although all possible values of K are tested, here we pick up three of the results; K=2, K=7, and K=11. All are in Fig. 2. First, we look at the result of K=2. The graph shows that genetic assimilation is still saliently observed.

6 parameters at the initial population are eliminated up to 2.9 (recall all results are an average of 100 runs) around 90th generation. This is one parameter more than the original simulation. Correspondingly, the position of the "plateau" shifts slightly to the right hand side. This means that slightly more generations are required to reach a single genotype. Secondly, K=7 is tested. The decrescent curve of the parameters is much shallower than that of K=2. As a consequence, the left edge of the plateau shifts more to the right. At this point, no decrement is observed. Rather, a small increase of plasticity is observed. This is because the increase of plasticity may improve the chance to obtain the fitness value of 2 or more. On the other hand, decrease of parameters is a tougher demand since it has to come with genetic convergence; a parameter cannot be replaced with 1-principle or 0-principle randomly; it must be par with other agents.

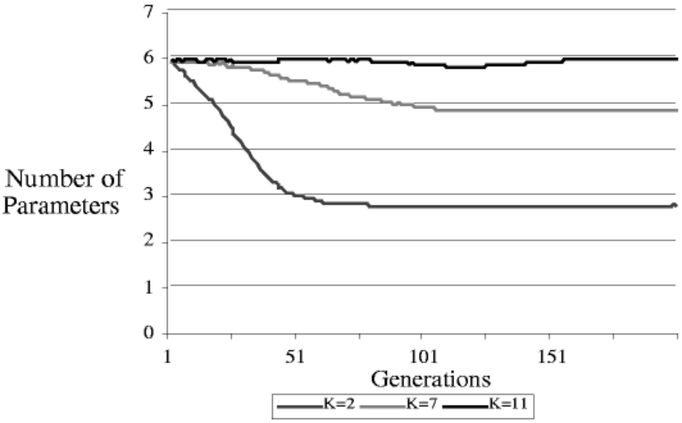

Fig. 2.

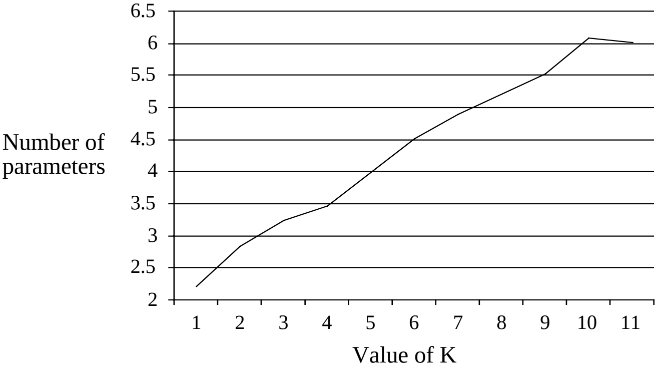

Fig. 3.

Fig. 3 shows the relationship between the value of K and the number of parameters at the end of each simulation. The consequence is crystal clear —as the value of K increases, more parameters remain in the population.

From these, it is apparent that the Baldwin effect is progressively weakened as the genetic dependency increases. In other words, the Baldwin effect is highly sensitive to epistasis and pleiotropy.

The results shown above beautifully reveal how epistasis and pleiotropy affect the Baldwin effect in populational dynamic communication. These results strongly suggest that parameters are hardly eliminated, even if keeping high plasticity may be a costly option. From these, it is now clear that under these circumstances, the scenario of the evolution of the LAD may severely undermine its elimination of parameters.

4 Conclusion

The experiments show that pleiotropy and epistasis effectively dampen the emergence of the Baldwin effect in the dynamic communication system. Although the modification is simple and quite straightforward regarding its technical complexity, the actual outputs are radically different. This has to be taken as a serious caution for our future studies. In sum, epistasis and pleiotropy in genes for the LAD, thus, may require a radical re-interpretation of the scenario of the evolution of the LAD.

However, there are some points we should improve the models to make a firmer claim. For example, in the simulations presented here, during the communication period, agents convert their ? characters to either 0 or 1 characters. We interpret this attempt to establish communication as learning. Strictly speaking, it is difficult to consider it as learning in a linguistic sense. In the simulations, learning takes place without any input from previous generations or even from the same generation. Usually, language acquisition takes place with linguistic inputs in a linguistic community. Adults' utterances are learners' primary linguistic inputs. When the learners become adults, their speeches become the next generation's inputs. Thus, linguistic inputs generally come down from previous generations to next generations. Such inputs are independent from genetic inheritance. Furthermore, the process does not include any update process of an agent's internal state.

Recently, more and more scholars have begun to reconsider the exact mechanism of the Baldwin effect. Most of the studies of the Baldwin effect itself share their roots in either Waddington's studies *in vivo* or Hinton & Nowlan's computer simulation *in silico*. Although the Baldwin effect is alleged to be observed in both studies, it is also true that the actual mechanisms for the Baldwin effect working in these studies are quite different. As Simpson [15] and Depew [9] argue, the Baldwin effect is easily dissected into its parts, and possibly the effect is simply just the sum of these parts. If we strictly follow this point of view, there is no need to invoke the sum as "a new factor in evolution [2]". In his exploration of language evolution, a biologist T. Deacon [8], however, has recently proposed a new type of mechanism of the Baldwin effect. This new mechanism, called "niche construction" has a self-organizing, emergent aspect in its core. This self-organizing, emergent type of mechanism seems to be particularly attractive for the case of language evolution, as it might provide a solution by which language evolution can circumvent the problem of pleiotropy and epistasis raised here.

References

1. Atmar, W. (1994). Notes on the simulation of evolution. *IEEE Transactions on Neural Networks* 5(1).
2. Baldwin, J. M. (1896). A New Factor in Evolution. *The American Naturalist, 30,* 441-451, 536-553.
3. Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

4. Chomsky, N. (1972). *Language and mind*, enlarged edition. New York: Harcourt Brace Jovanovich.
5. Chomsky, N. (1981a). *Lectures on government and binding*. Dordrecht: Foris.
6. Chomsky, N. (1981b). Principles and parameters in syntactic theory. In N. Hornstein & D. Lightfoot (Eds.), *Explanations in Linguistics*. London: Longman.
7. Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.
8. Deacon, T. (1997) *The Symbolic Species*. New York: W.W. Norton.
9. Depew, D. (2000) The Baldwin Effect: An Archaeology. *Cybernetics And Human Knowing*, *7*, (1), 7-20.
10. Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Computer Systems, 1*, 495-502.
11. Kauffman, S. A. (1989). Adaptation on rugged fitness landscapes. In D. L. Stein (Ed.), *Lectures in the science of complexity, 1*. Redwood City, CA: Addison-Wesley.
12. Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces. *Cognition*, *61*, 161-193.
13. Piatelli-Palmarini, M. (1989). Evolution, Selection and Cognition: From “learning” to parameter setting in biology and the study of language. *Cognition*, *31*, 1-44.
14. Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences, 13*, 707-784.
15. Simpson, G. G. (1953) The Baldwin Effect. *Evolution*, *7*, 110-117.
16. Turkel, J. W. (to appear) The Learning Guided Evolution of Natural Language. In T Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. New York: Cambridge University Press.
17. Turney, P., Whitley, D., & Anderson, R.W. (1996). Evolution, learning, and instinct: 100 years of the Baldwin effect, *Evolutionary Computation*, *4*, (3), iv-viii.
18. Waddington, C. H.: *The evolution of an evolutionist*. Edinburgh: Edinburgh University Press.