THE BALDWIN EFFECT WORKS FOR FUNCTIONAL, BUT NOT ARBITRARY, FEATURES OF LANGUAGE

MORTEN H. CHRISTIANSEN & FLORENCIA REALI
Department of Psychology, Cornell University, Uris Hall, Ithaca, NY 14853, USA

NICK CHATER
Department of Psychology, University College London, 26 Bedford Way London, WC1H 0AP, UK

Human languages are characterized by a number of universal patterns of structure and use. Theories differ on whether such linguistic universals are best understood as arbitrary features of an innate language acquisition device or functional features deriving from cognitive and communicative constraints. From the viewpoint of language evolution, it is important to explain how such features may have originated. We use computational simulations to investigate the circumstances under which universal linguistic constraints might get genetically fixed in a population of language learning agents. Specifically, we focus on the Baldwin effect as an evolutionary mechanism by which previously learned linguistic features might become innate through natural selection across many generations of language learners. The results indicate that under assumptions of linguistic change, only functional, but not arbitrary, features of language can become genetically fixed.

1. Introduction

Although the world's languages differ considerably from one another, they nonetheless share many systematic constraints on how they are structured and used. Explaining how such universal linguistic constraints evolved in the hominid lineage is the focus of much debate in language evolution research. One view suggests that linguistic universals are best viewed as *arbitrary* features of language with no functional explanation, but instead deriving from an innate Universal Grammar (UG; Chomsky, 1965). This abstract body of linguistic knowledge is proposed, by some theorists, to have evolved gradually through biological adaptations for increasingly complex grammars (e.g., Briscoe, 2003; Pinker & Bloom, 1990). An alternative view seeks to explain linguistic universals as *functional* features of language, emerging due to communicative and cognitive factors outside of grammatical knowledge (e.g., Bybee, 1998). These features are seen as by-products of linguistic adaptation, in which language itself has been adapted through cultural transmission across many generations of language learners (e.g., Tomasello, 2003).

The Baldwin effect (1896) is the primary evolutionary mechanism by which the arbitrary features of UG are envisioned to have been genetically fixed in the human population. Although a Darwinian mechanism, the Baldwin effect resembles Lamarckian inheritance of acquired characteristics in that traits that are learned or developed over the life span of an individual become gradually encoded in the genome over many generations (see Weber & Depew, 2003). That is, if a trait increases fitness, then individuals that, due to random genetic variation, require less exposure to the environment to develop that trait will have a selective advantage. Over generations, the amount of environmental exposure needed to develop this trait decreases, as individuals evolve increasingly better initial conditions for its rapid development. Eventually, no environmental exposure may be needed; the trait has become genetically encoded. A frequently cited example of the Baldwin effect (e.g., Briscoe, 2003) is the ability to develop hard skin on certain areas of the body with relatively little environmental exposure. Over time, natural selection would have favored individuals that could develop hard skin more rapidly (because it aids in mobility, prevents infection, etc.) until it became fixed in the genome, requiring little environmental stimulation to develop. Similarly, it has been suggested that arbitrary linguistic features, which would originally have had to be learned, gradually became genetically fixed in UG via the Baldwin Effect (Pinker & Bloom, 1990).

In this paper, we use computer simulations[a] to investigate the circumstances under which the Baldwin effect may operate, for arbitrary and functional features of language. Building on previous work (Chater, Christiansen & Reali, 2004), Simulation 1 indicates that arbitrary linguistic features *cannot* be genetically fixed via the Baldwin effect when linguistic change is incorporated — even when this change is driven in part by the genes themselves. In Simulation 2, we show how functional features of language can come to be genetically fixed in the population when they promote better communicative abilities. Finally, we discuss the implications of the simulations for theories of language evolution.

2. Simulation 1: Arbitrary Language Features

Following recent work on the possible evolution of UG (e.g., Briscoe, 2003; Nowak, Komarova & Nyogi, 2001), we model language and learners as a set of binary vectors. Specifically, we adopt the framework of the pioneering

[a] All simulations were replicated several times due to their stochastic nature.

simulations of Hinton & Nowlan (1987), used by Pinker & Bloom (1990) to support their suggestion that the Baldwin effect underlies the gradual genetic fixing of arbitrary grammatical features in UG. Our previous work indicated that although the Baldwin effect can occur within this framework in the context of arbitrary linguistic features, the effect disappears when language is allowed to change (Chater et al., 2004). However, these simulations were limited in scope; we therefore conducted a new series of simulations to determine whether our original results would replicate after addressing the limitations.

In our earlier simulations, a language was defined as a set of arbitrary binary features, $F_1 \ldots F_n$, taking the values 0 or 1. The n "genes" of the learners correspond to each of the n features of the language. The genes can take three values, representing an innate bias (0, 1) for a feature being 0 or 1 in the language; or neutrality (represented as '?'). For example, if $n = 3$ the language may correspond to [0, 1, 1] and the genes of a random agent to [?, 1, 0]. At the beginning of each generation, an initial language (phenotype) is expressed for each agent based on its genes (genotype). The innate bias toward a particular feature value will in most cases result in that value being expressed in the phenotype (in most of the simulations the 'stickiness' of the bias is 95% in the direction of the designated value), but on occasion it will be expressed in the opposite direction. For the neutral (learning) genes there is a 50% change of either setting (1 or 0). Thus, in our previous example, the initial language of the agent could be [1, 1, 1]. If the initial language does not match the target language, the agent begins a process of trial and error learning, in which learners randomly sample features using the biases in their genes. Once a feature is 'guessed' correctly, it is not changed. The learner keeps guessing until all the features in its language match those of the target language, with the fastest learners being selected to form the basis for the next generation. Some mutations would occur across generations, with an equal probability of randomly reassigning a gene to 0, 1, or ? (mutation rate varied between simulations). Although the neutral bits initially speeds learning, agents that are genetically biased toward a feature F_i will guess it faster. Thus the Baldwin effect should gradually ensure that all the arbitrary features of the language become genetically encoded.

Chater et al. (2004) found a Baldwin effect for arbitrary linguistic features, for the case where the language is fixed. In these simulations, reproduction was implemented as simple duplications of the top 50% of the learners subject to a 1% mutation rate. Does the same result hold, given a more realistic model of genetic transmission? To better approximate hominid evolutionary dynamics,

Table 1. Number of generations needed to reach the success criterion for the Baldwin effect (parameter value : number of generations)

Genome Size	Population Size	% Initial Neutral Bits	Stickiness of Innate Bias	% Survivors	% Mutation Rate
10 : 25	24 : 369	0 : 23	100 : 152	26 : 52	0.1 : 232
20 : 51	100 : 51	25 : 69	95 : 51	50 : 51	1 : 51
50 : 201	250 : 47	75 : 137	90 : 85	74 : 195	2.5 : 104
80 : 1045		100 : 147	80 : 88		

the current simulations use a simple model of sexual reproduction, instantiated as random cross-over between two sets of learner genes.

We first replicated our original results in which the language/genome size was set to 20, the population size to 100, the number of initially neutral bits to 50%, the 'stickiness' of the innate genetic bias to 95%, the number of surviving agents to the top 50%, and a 1% mutation rate. Using a success criterion that more than 95% of the initial bits in the top 50% of the learners' genomes should correctly match the target language, we found that a robust Baldwin effect occurred after 51 generations. We then varied the simulation parameters and found that a robust Baldwin effect occurred in all circumstances, with parameter variations only affecting the speed with which it emerged (see Table 1). These results show that our earlier results generalize to sexual reproduction, and show that the Baldwin effect is highly robust, with a fixed language. If such a robust effect disappears under when the language is allowed to change, this cannot easily be dismissed.

An important limitation of our original simulations is that language change was completely independent of the genes. It seems reasonable to assume that if the genes control language learnability then they should also influence the direction of language change in a process similar to Baldwinian niche construction (e.g., Odling-Smee, Laland & Feldman, 2003). To explore this, we carried out a set of simulations in which language at time $t+1$ was determined by a combination of genes and language at time t. Specifically, p percent of the change would be determined by the most frequent gene values in the previous population and the remaining $1-p$ percent of change by the previous language. Given that other pressures than learnability also affects language change (such as cognitive/communicative constraints, parsability, language contact, linguistic drift, etc.), we also incorporated random language change at a rate of ten times faster than the mutation rate (i.e., 10%). The faster rate of linguistic change reflects the fact that cultural evolution is much faster than biological evolution (Dawkins, 1976). Whereas linguistic change is measured in thousands of years, biological evolution is measured in hundreds of thousands of years. Other

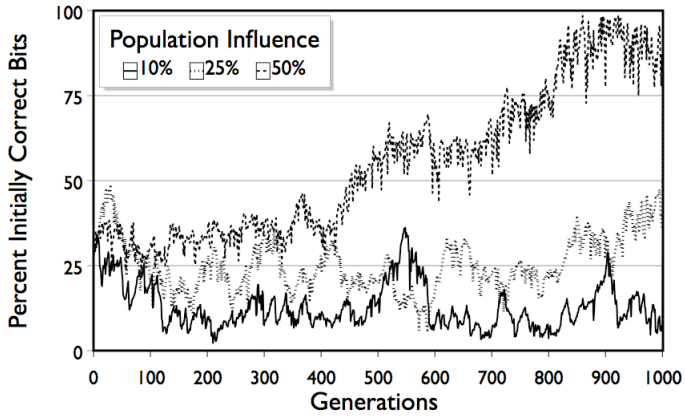

Figure 1. The effect of population influence on the emergence of the Baldwin effect.

simulation parameters were the same as in our initial replication above.

The results of these simulations (Figure 1) show that only when there is a very high degree of population influence does the Baldwin effect emerge. Only when the direction of linguistic change is at least 50% determined by the previous generations genes do we observe a robust Baldwin effect after 835 generations. This suggests that arbitrary features of language would have to be predetermined strongly by the genes from the very beginning, thus leaving little room for subsequent evolution of the kind envisioned by Pinker & Bloom (1990). This corroborates our previous findings that under reasonable assumptions about language change, the Baldwin effect does not occur for arbitrary linguistic features. Unlike the example of hard skin, where the environment provides a stable target for the Baldwin effect, language change is too fast for genetic commitments to arbitrary features to be worthwhile. However, it is possible that non-arbitrary features of language could become genetically fixed in the population if they facilitated communication in some manner; e.g., improved abilities for word learning, increased working memory capacity for language, vocal apparatus optimizations for speech, and so on.

3. Simulation 2: Functional Language Features

Because the arbitrary features of language by definition do not affect communicative function (e.g., Pinker & Bloom, 1990), Simulation 1 did not need to incorporate communication between agents. However, to explore the degree to which functional features of language could have become genetically

fixed via the Baldwin effect, it is necessary to take communication into account to provide a context within which the non-arbitrary features can be functional.

We used the same representation of language and genes as before, with the initial language expressed in the same way. However, learning was implemented differently, now mediated by communicative interactions. Communication was only possible between agents who had a majority of the same kinds on language features (either 0 or 1). Thus, an agent, a_1, whose language is [0, 0, 0, 0, 1], would be able to communicate with an agent, a_2, with a [0, 0, 0, 0, 0] language but not with agent a_3 that has a [0, 1, 1, 1, 0] language. Agents benefit mutually from successful communication in proportion to the overlap in their features. The successful *two-way* interaction between a_1 and a_2 would result in an increase in both agents' communication scores by 9 (the combined number of 0s in their two languages). The simulations also integrate the developmental trend that comprehension precedes production: even though a_1 can only "produce" four 0s, it can "comprehend" a_2's five 0s. However, if the difference between the productive abilities of two agents is more than one unit, then lesser competent "speaker" will not be able to understand its more proficient communication partner, resulting in a *one-way* interaction. In this case, the proficient speaker received the combined communication score (as before), whereas the less competent agent would only receive its own contribution to that score. Hence, if a_2 interacted with a_4, whose language is [0, 1, 0, 1, 0], a_2 would increase its communication score by 8 while a_4's score would only increase by 3.

In this framework, less competent agents are able to learn from more competent agents (with stronger bias towards 0s or 1s); this is meant to reflect the tendency for children to learn much of their language from others with greater language skills than themselves (e.g., adults or older children). Learning can only happen when two-way communication is possible (as described above), and consists in a process in which the less competent agent, based on the biases in its genome, re-samples the first bit in its language that differs from the more competent agent's language. For example, in a communicative interaction between a_1 and a_4, the latter would resample its second language bit. If a_4's genome encoded an innate bias (0 or 1), then there would be a 95% chance of getting this bit expressed; but if the genome encoded a neutral bit, the chance of either value would be 50%. Thus, genes constrain learning as in Simulation 1.

To further mirror the learning conditions from the previous simulations, we introduced noise into the learning process at a rate ten times higher than the mutation rate. During 10% of the learning opportunities a random bit in the

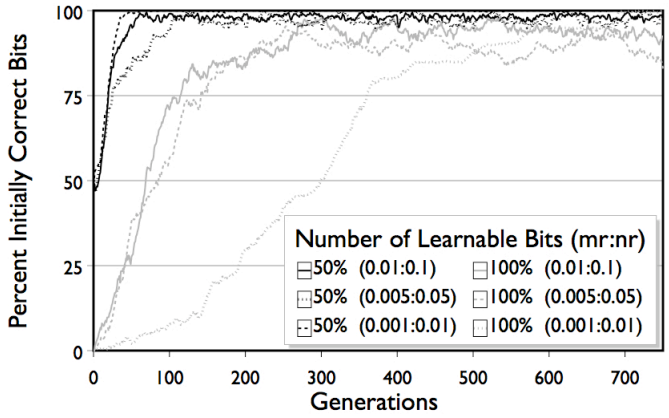

Figure 2. The influence of variations in the number of initial learnable bits on the Baldwin effect for different mutation rates (mr) and noise rates (nr).

learner's language would be chosen for potential reassignment (given the learner's genetic bias for that bit) instead of the first bit that deviated from the competent speaker's language. This paralleled the 10% random change in the target language in Simulation 1.

From each generation of 100 learners, pairs of agents were randomly picked for 500 interactions. The 50 agents with the highest communication scores were selected, and cross-over sexual reproduction used to create the next generation (combined with a 1% mutation rate). The results (Figure 2) show that a robust Baldwin effect emerges across several different variations in mutation rate and number of neutral bits in the first generation. Even when the first generation has all neutral (learnable) bits, a robust Baldwin effect emerges after 33-269 generations. Thus, functional features that improve communication abilities may become genetically fixed in the population. For example, vocabulary learning is likely to rely on innate domain-general abilities for establishing reliable mappings between forms and meanings (e.g., Bloom, 2002). As such, the ability to acquire a large vocabulary may have become gradually innate by way of the Baldwin effect because it would have increased communicative abilities.

4. General Discussion

These results indicate that the Baldwin effect may not provide a suitable evolutionary mechanism for explaining the emergence of arbitrary features of language. Rather, the results suggest that functional features that facilitate communication may be a better candidate for aspects of language that have come

to be genetically fixed over evolutionary time. For a trait to be amenable to the Baldwin effect, it needs to be stable over a period of many generations. Functional features are stable in that they facilitate communication on a continuous basis and thus are likely to become 'Baldwinized' when communicative abilities affect selective fitness in a population. In contrast, abstract linguistic features are free to change randomly exactly because they are non-functional and not subject to direct selective pressures. More generally, the simulations raise doubts about the gradual evolutionary emergence of a UG, as proposed by Pinker & Bloom (1990), and instead support a cultural transmission model of language evolution in which the Baldwin effect has enabled certain cognitive/functional features to become genetically encoded.

References

Baldwin, J.M. (1896). A new factor in evolution. *American Naturalist, 30,* 441-451.

Bloom, P. (2002). *How children learn the meanings of words.* New York: OUP.

Briscoe, T. (2003). Grammatical assimilation. In M.H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 295-316). New York: OUP.

Chater, N., Christiansen, M.H. & Reali, F. (2004). *Is coevolution of language and language genes possible?* Paper presented at the Fifth International Conference on the Evolution of Language, Leipzig, Germany.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Bybee, Joan (1998). A functionalist approach to grammar and its evolution. *Evolution of Communication 2*. 249-278.

Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.

Hinton, G.E. & Nowlan, S.J. (1987). How learning can guide evolution. *Complex Systems, 1,* 495-502.

Nowak, M.A., Komarova, N.L. & Nyogi, P. (2001). Evolution of universal grammar. *Science, 291,* 114-118.

Odling-Smee, F.J., Laland, K.N. & Feldman, M.W. (2003). *Niche construction: The neglected process in evolution.* Princeton, NJ: Princeton University Press.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*, 707-784.

Tomasello, M. (2003). On the different origin of symbols and grammar. In M.H. Christiansen and S. Kirby (Eds.), *Language evolution* (pp. 94-110). New York: OUP.

Weber, B.H. & Depew, D.J. (Eds.) (2003). *Evolution and learning: The Baldwin effect reconsidered.* Cambridge, MA: MIT Press.