Syntax 2:1, April 1999, 28–37

THE 'SHELL GAME': WHY CHILDREN NEVER LOSE

Teresa Satterfield

Abstract. This paper articulates a formal solution to the puzzle of child language learnability within the Principles and Parameters–based framework. The language learning (parameter-setting) task requires, in principle, that the selection of syntactic knowledge be sufficiently constrained for the child to arrive at the appropriate target grammar, expending a minimum of computational effort and time. Since previous L1 analyses impose very strict requirements on the learner, solutions are achieved only at a very high cost. Further, not only do the standard accounts frequently contradict fundamental empirical facts of child linguistic development, such as the degree of variability observed in the production of early grammatical structures regardless of input; the accounts also fail to consider a crucial aspect which impacts selection: the young child's innate potential to efficiently acquire multiple languages simultaneously. The primary aim of the current paper is to provide a computational model that demonstrates a ''bilingual universals'' (in the spirit of Roeper 1996) stage of development based on real world data. The proposed model actually reflects a more precise UG-based representation within early monolingual grammars, as well as plausibly accounting for variability found in child L1 grammars.

1. Introduction

Genetic evolution as a computational technique was proposed and advanced by Holland (1975). It has since been refined and elaborated by numerous researchers, and applied to various domains. Clark (1990, 1992) actually led the way in the promising direction of facilitating solutions to the language learning and language change problems with a type of simple Genetic Algorithm (GA). From this work, it does seem that language acquisition can at least abstractly be patterned after the behavior of other adaptive (intelligent) systems in nature.

This paper begins with a brief discussion of GA parameterization techniques, taking the behavior of the Null Subject Parameter (NSP) across languages as illustrative of the language acquisition process confronting different L1 learners. With the formulation of bilingual learnability as a point of departure, the question of monolingual acquisition is then addressed. Drawing on compelling evidence which shows the child to be at once conservative and non-conservative with respect to the selection of particular hypotheses (Pinker 1984, among others), I advance the idea that, as a consequence born in part from the capability for multilingualism that is initially inherent to all children, and in part from frequently ambiguous input data, a child's early grammar contains a degree of variation which is essentially more ''bilingual'' than monolingual. The paper concludes with highlights of the GA's performance on a monolingual L1 parameter-setting task.

2. Genetic Algorithms

GAs are computer programs designed to efficiently search a complex problem space and obtain near-optimal solutions. Since this algorithm is a search procedure based on the mechanics of Darwin's natural selection, it uniformly combines survival-of-the-fittest tactics with randomized, yet structured, exchanges of information in order to form a search technique that mirrors some of the creativity reflected in human search operations. In the current study, research is presented that lends additional support to the effectiveness of GAs, as based on two syntactic parameter-setting models. By implementing well understood biological (and mathematical) formulations, it becomes possible to formalize a theory of language learning which generates a given parameter's options on analogy with natural selection.

The basic premise is the following: genetic search can be used to optimize a function over a discrete parameter space, so that any point in the parameter space can be represented as a *n*-bit vector. The technique manipulates a set of such vectors to record information gained about the function. The pool of vectors is called the *population*, an individual bit vector in the population is called a *genotype*, and bit values at each position of a genotype are called *alleles*. The function value of a genotype is called the genotype's *fitness* or *merit* score.

- *n*-bit vectors or strings are represented much like DNA chains which hold genetic information (i.e., 100111101...).
- A pool of strings (10011101, 00101110, 11011001, etc.) = the population.
- Each individual string is a genotype.
- The bit values at each position of the genotype are alleles. (1_____)
- The value of genotype is the genotype's fitness or merit score. (1.0=fit)

Figure 1: GA Terminology

Two primary operations apply to the population in a standard GA. Reproduction changes the contents of the population by adding copies of genotypes with above-average fitness. No new genotypes are introduced; however, by changing the distribution in this way, the average fitness of the population tends to rise to that of the most-fit of the existing genotypes.

Along with ''fitness-based reproduction,'' it is also necessary to generate new genotypes and add them to the population. The primary means for generating plausible new genotypes is with crossover. In a simple GA, crossover entails the selection of two random genotypes, taking some alleles from both ''parents,'' and recombining these to produce a complete genotype. The offspring are added to the population, where they have the opportunity to survive or die depending on their own fitness measures.

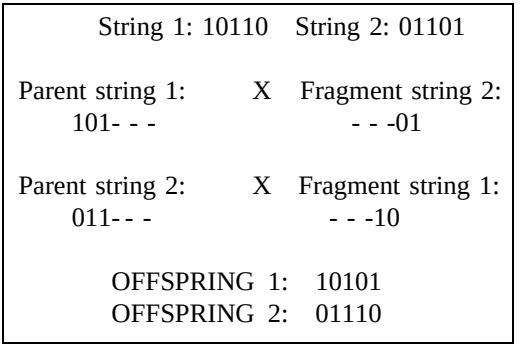

Figure 2: Schematic of Crossover

To perform a search, reproduction and crossover operations are iterated. Eventually a maximally-valued genotype will come to dominate the population, and convergence is attained. When the population has converged to a single genotype, reproduction and crossover will no longer change the makeup of the population. For this reason, most GAs also include a mutation operator, which provides a chance for any allele to be changed to another randomly chosen value. Hence, mutation guarantees that every value in every position of a genotype has a chance of occurring.

3. A Metaphor for Language Learning

As proposed by Clark (1990, 1992), the example of the GA can be successfully invoked to represent the basic non-deductive and ''automatic'' nature of parameters as a search method. In this model, the child is abstractly formalized as a GA to the extent that s/he, like other simple GAs, maintains and operates on a population. In this case, the GA works on a population of hypothesis strings which is made up of sequences from a binary alphabet that expresses truth value expressions where 1 equals true and 0 equals false to represent features or characteristics of a given parameter of variation. Given a grammar, a truth value can be assigned to each of the parametric

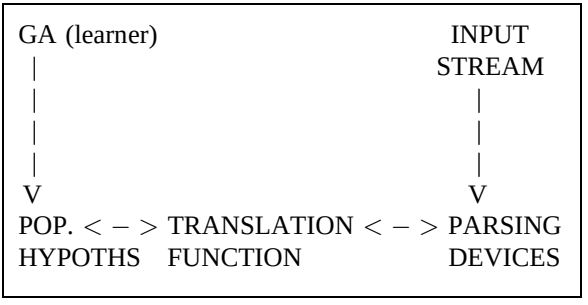

Figure 3: Clark's GA

propositions, based on the properties assumed to be held in a particular target grammar, making up a large string of 0's and 1's which represent a grammar's entire parametric sequence. These individuals are then mapped onto parsing devices through a translation function. The output of each parser is then judged against a linguistic input item. The subsequent ratings are then mapped back through the translation function to the population, where genetic operators of the GA component utilize the information to produce a new generation of strings.

The GA model seems to deal with many of the issues that made language learning problematic in principle for the earlier accounts: (i) GAs avoid the costliness of computations because the process of deductive reasoning is simulated by the evolutionary cycle, so the child is not directly using knowledge-based reasoning strategies to choose between competing hypotheses, thus minimizing computational effort over time; (ii) the GA learner works from a population-based strategy to ensure the arrival at the correct parameter value when the input data conflicts with the population hypotheses, since s/he would have access to sets of hypotheses in different locations of the search space, one of which would be likely to correspond to the target grammar, at least superficially. Also by implementing a population-based technique, it is possible to maintain a range of solutions from bad, good, better, or best, giving the learner more explicit selection information; (iii) whereas the previous accounts posit certain mechanisms that, in the strictest forms, disallow too general hypotheses, the GA can guide hypothesis selection in a less rigid manner, so that it allows for overgeneralization, yet gives preference for more conservative hypotheses as time goes on. This feature is quite elegant, as it depicts the typical behavior of child learners, who can and do change from a current hypothesis even in the absence of errors.

From the current knowledge of real world languages, the main criticism of the Clark models may be that the author simply stopped too soon. However, there are limitations that have surfaced in nonlinguistic applications of simple genetic algorithms that prompt us to also modify and extend the GA with respect to language learning. For instance, the simple GA is a reliable method for discovering the defining characteristics of one single grammar, but if the learner is simultaneously faced with more than one input in the environment, as is the case for the bilingual child presumably (and perhaps the young child who is developing two registers within one language such as Standard American English and African American English), then in these scenarios, the GA model described thus far requires that a learning system converge to the one-best solution, giving way to a performance that is not always desirable. The Clark GA will treat the distinct hypotheses as competing for one category, thereby producing an average population that may possibly have good performance in only one of the necessary target languages. This process is often known as *genetic drift*. Due to the averaging effect, it is more than likely that the resulting solution will be sub-optimal with respect to any of the linguistic environments.

4. An Alternative GA Approach

To represent the requirements on language speakers in a more realistic manner, the concepts of *niche* and *speciation* are incorporated into the model (Goldberg 1989, Forrest 1993). The need for niche development not only stems from the notion that a distributed formulation can give better results with less total work; it also allows for the attainment of an optimal *set* of solutions. A very natural way to distribute a GA is to partition the population of hypotheses, since it is not computationally economical to repeatedly conduct genetic search operations on a population-wide basis.

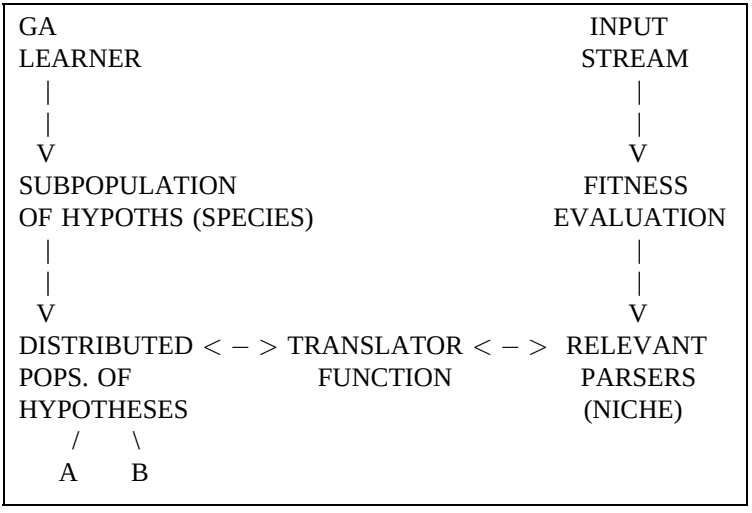

Figure 4: An Alternative GA

Along these lines, *niche* can be thought of in terms of a string's specialized behavior or attribute(s) in a given habitat. For the purposes of this investigation, it will be claimed that a hypothesis string has a niche to the extent that it is relevant to one target grammar. For language learning, the definition of *species* can include a set of strings that share common truth value sequences or have defining characteristics such as proximity in the search space. In nature, it is a well-known fact that species typically thrive under different combinations of environmental factors, or niches, which would prove relatively unappealing to other species.

For the computational model, the notion of niche is formalized by the following procedure: when an input datum is received and decoded, the relevancy of all strings to that specific environment is preliminarily measured via a simple ''match scoring'' mechanism (Booker 1985). Genotypes become relevant only if they reach the minimum match threshold for that particular iteration as measured against the input text. The niche rate determines how many strings will be useful during each cycle, by assigning values that allow only those hypotheses with the best scores to be categorized as relevant for the

search. For instance, given input string 1001*, hypothesis string 1110* would not be activated, whereas strings 1011* and 1001* are closer matches and, depending on the threshold requirement for that particular generation, could both be selected. This niche mechanism makes a sound contribution to the GA approach, not only because as it acts as a ''hardware accelerator'' by speeding the learning toward convergence; it also ensures that the input text will be minimally parseable by enlisting the most relevant hypotheses of that population for the task. This fairly efficient operation further reduces the possibility of any type of random walk through the problem space. Since the goal is ultimately to model a child who actually ''learns,'' in the sense that s/he becomes increasingly more accurate at distinguishing and representing linguistic input over time, the niche application with its gradually rising threshold does that by making the GA progressively sensitive to the similarities and differences between certain genotypes in a way that it was not previously.

After the species is determined, each potential solution is mapped to its corresponding parsing device and measured for fitness relative to the local population. Drawing largely from Clark (1990, 1992), a string's merit is defined by its ability to account for linguistic evidence as a function of the number of core violations returned by the parse, and a weighted economy feature which signals the number of syntactic chains covered by each parse. Within current generative linguistic theory, there should not be a discrepancy of elegance or economy among successful parses: theoretically, the string is always covered in the least number of nodes (Chomsky 1995). In the case of a failed parse, then, it makes no difference if the node count is fifteen or fifty, it is unacceptable if the string does not represent the minimal structure available to the learner. Contrary to the stipulation given for the alternate GA model, there does not appear to be a fitness continuum of bad to best in current linguistic formulations. Still the issue awaits compelling proof one way or the other; although data in child language studies appears to attest that for some constructions such as null arguments, it seems to be the case that certain structures are simply more economical than their counterparts in other languages. For this reason, the decision was made to retain the economy feature in this GA, reflecting the intuition that there may be an acceptable variation in node counts given two exact sentences in two respective grammars.

For recently evaluated hypotheses, the current fitness of each member is then stored and the genotypes are mapped from the parsing devices to the genetic process locale. In an attempt to prevent any distortion to the balanced search strategy that has been carefully initiated, recombination is limited by a ''likes-mate-likes'' policy, so that parents are normally chosen from within the same niche membership. In short, the motivation for this strategy is that there is little advantage in performing crossover between different sets when the recombination of distinct members is not likely to assist in the search for better hypotheses. The best choice seems to be to implement a mating restriction, since the GA is more likely to combine the bits of genotypes that are already in a family of proven ''winners,'' relative to their particular

environment. Following crossover, the fitness values of the new offspring are calculated. The stop criteria chosen in this model regulate the resources and space among the hypotheses in distributed populations, and places an upper limit on the number of cycles. I assume that the language learner has converged when genotypes directly correspond to the target grammar(s). That is, about 90% of the hypothesis strings match their respective target at every bit position. The other 10% should match at almost every bit, but not be identical. The learner is said to strongly converge only when the subpopulations are at an equilibrium and the most genotypes of each subgroup remain identical to the target strings' sequences.

All in all, these functions were created and organized to embody the characteristics of language learning in general. As is no doubt obvious, this current model lacks a mutation operator. Originally, the probability of mutation was suppressed to 0 in order to more rigorously test the other operations. Intuitively, it was thought that if stable groups of genotypes could be maintained without mutation, then performance would be even more enhanced when the mutation application were present. In the final analysis, it seemed logical to eliminate mutation in general during the language learning task, since any such continuous, random alternation goes against the deterministic bent of a parameter-setting concept (Satterfield 1995).

5. Simulations and Discussion

An application of the alternative GA will be highlighted for Chinese in this section. Nevertheless, taking the behavior of the Null Subject Parameter (NSP) across languages the process confronting the very young language learner can be illustrated through a variety of target genotypes, such as: 1001*, 0110*, 0000*, and 1111* encoded for Spanish, English, Chinese, and German languages respectively. The five-bit strings abstractly express features of the target grammars which are found in real child-caregiver language data, such as ''this grammar projects TP: yes (1) or no (0)''; ''this grammar has strong D-features in its functional heads: yes or no.'' When the proposition to be specified is not crucial, the bit is listed with ''*,'' the

- Spanish Target = 1001* [data base includes ambiguous variations: (1001*)]
- English Target = 0110* [data base includes ambiguous variations: (0111*)]
- Chinese Target = 0000* [data base includes ambiguous variations: (0010*)]
- German Target = 1111* [data base includes ambiguous variations: (1101*)]

Figure 5: Grammars as Genotypes

"wildcard" marking. In Spanish and Chinese, null subjects are licensed, in English, they are not. German is a partial case.

The intent is to show how the GA demonstrably fits the child's linguistic development process. With respect to a monolingual environment, GA findings successfully utilize the initial capacity for multilingualism which is inherent to all very young children. Since it is widely held that every child has the innate potential to acquire multiple languages simultaneously, it is reasonable to represent a large pool of hypotheses that the child must sort through in a quick and efficient manner. As a consequence of not only this multilingual capability, but also of the effects of ambiguous input data, the L1 learner maintains different species of strings within the total population. These strings represent the variations that surface concurrently in every child's grammar, and which cause developing grammars to be essentially more "bilingual" than monolingual in nature during the earliest phases. As the monolingual child locks onto the "appropriate" values for the target grammar, the seemingly free variation of target structures with "inappropriate" constructions, (i.e., the production of both overt and thematic null subjects in a very young child's English) will incrementally disappear, giving rise to the correct forms for that particular grammar.

Given the nature of this particular GA model, L1 parameterization proceeds systematically to reduce the hypotheses which are incompatible with the input. The monolingual child will lose this early access to a range of variations, fully maintainable only by a child receiving constant bilingual input matching the dual parameter values. These assumptions have broad implications, not only

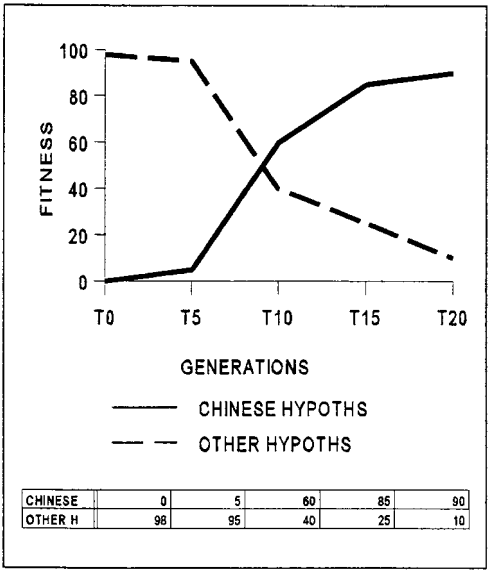

Figure 6: Results of a GA Simulation

for a formal GA model, which can be economically extended to other applications such as L2 acquisition, but it also provides more conceptual force and adaptability for the Principles and Parameters framework, which now receives more formal motivation for the ability to constrain acquisition without inhibiting the child's capacity to obtain multiple grammars.

6. Conclusion

An assessment of any computational model is incomplete without some cost comparison or evaluation of performance. It is a very reasonable practice to evaluate the performance of methods in order to choose the most efficient strategy for solving the class of problem. As a point of comparison, the results of Clark's GA model which also simulates parameter-setting was checked against the current model. A main point is that while the Clark version is quite simple, it still requires much memory and time, as it operates on a global level. Given genotypes of the same length as those implemented in the alternative GA, the Clark model takes an upper limit of thirty-five generations for the learner to converge, as opposed to the twenty generations averaged in the modified GA. Moreover, his application can only model the purely monolingual speaker, if such an individual indeed exists, since all hypotheses must be averaged to one single solution.

Several monolingual theories of parameter-setting have been proposed throughout the years; however, the introduction of bilingualism and bilectalism into the mix has always presented difficulties for other approaches. The main advantage with the current formulation is that it is the only model to date which characterizes a single, universal parametric system that economically interprets learnability (via parameters) in a wider context to include many types of speakers. For this reason, a direct comparison between previous models and this latest alternative is no longer valid, considering the substantially different approaches which have now evolved.

It is important to understand that as with any model of this nature, this GA application is only a representation: it is merely a metaphor for the actual process of L1 language learning. It is not a comment in any form on the actual cognitive or physical mechanisms that may be involved in language acquisition. Indeed, the Principles and Parameters framework in and of itself is an abstraction of learnability; thus, it is impossible to even decide at this point on the reality of the existence of a successful parameter model, or what that model would necessarily contain. Given more empirical backing of the type mentioned in the current investigation, then to the extent that the model fits additional patterns of child grammatical development, I simply hope to demonstrate that the representation offered with this particular GA model is more or less feasible.

It is also important to understand that, for good or for bad, this model will never allow the learner to reduce the genotypes to one singly optimal solution. As a consequence, it may sacrifice some flexibility at a certain level.

Finally, there are those who might reject any computational modeling of language out of hand; in this case, the GA does not often escape criticism. It has been noted that genetic algorithms are extremely powerful search tools, to the degree that they can solve problems in several domains, just so long as the fitness metric and the operators have been fine-tuned to exactingly specific settings. Still, it must be emphasized that GA techniques are far from perfect; they also impose a trade-off of sorts. They sacrifice peak performance in order to quickly achieve relatively high-quality solutions or levels of performance. Nevertheless, this ''satisficing'' strategy might actually be closer to what occurs in natural language learning. Another criticism leveled at the GA model is that it is too intricate to represent a child learner. Recall that the young learner is never conscious of this language acquisition process; it can be likened in this sense to an involuntary reflex, such as simple respiration. On parallel with language, if one attempts to model respiratory mechanisms, the account will be much more complicated than the physical outcome displays. Just as the youngest of children breathe quite well, with no conscious effort of the process, they also obtain grammar(s).

References

BOOKER, L. 1985. Improving the performance of genetic algorithms in classifier systems. In *Proceedings of an international conference on genetic algorithms and their applications*, ed. J. Grefenstette. Hillsdale, NJ: Lawrence Erlbaum.

CHOMSKY, N. 1995. *The minimalist program.* Cambridge, Mass.: MIT Press.

CLARK, R. 1990 a, b. Papers on learnability and natural language selection. In *Technical reports in formal and computational linguistics 1*, Université de Genève.

CLARK, R. 1992. The selection of syntactic knowledge. *Language Acquisition* 2:85–149.

FORREST, S. ed. 1993. *Proceedings of the fifth international conference on genetic algorithms*. San Mateo, CA: Morgan Kaufman.

GOLDBERG, D. 1989. *Genetic algorithms in search, optimization, and machine learning*. Reading, Mass.: Addison-Wesley Publishers.

HOLLAND, J. 1975. *Adaptation in natural and artificial systems*. Ann Arbor: The University of Michigan Press.

PINKER, S. 1984. *Language learnability and language development.* Cambridge, Mass.: Harvard University Press.

ROEPER, T. 1996. Is bilingualism universal? A view from L1 acquisition. Ms., University of Massachusetts.

SATTERFIELD, T. 1995. Bilingual selection of syntactic knowledge: The extended parameterization hypothesis. Ph.D. dissertation, University of Iowa.

Teresa Satterfield
Program in Linguistics
University of Michigan
1076 Frieze Building
105 South State Street
Ann Arbor, MI 48109-1285

tsatter@umich.edu
www.umich.edu/~tsatter/