Evolutionary Model for Child Language Acquisition

Emily Jones, Scott Dubinsky, Richard Higgins
University of Maryland, College Park

Abstract

A major open question in linguistics is how to model child language acquisition in a robust yet accurate manner. We propose an evolutionary computation model for language acquisition based on the parameter theory of universal grammar. To find the ideal genetic algorithm parameters, we tested variables including parameter count, parameters left unset in a language, number of grammars, how many parameters are set per sentence, mutation rate, recombination rate, survivor selection strategy, and survivor count. We also tested variants of the script which implemented bilingualism, negative reinforcement, and errors in sentences. In all variants, we found that mutation was deleterious, generational selection was best, and that a complete language with no unset parameters did not hinder language acquisition. Furthermore, our model predicts that bilingual learning requires knowledge of which language is which, negative reinforcement is ineffective, and errors in sentences do not hinder language acquisition. Overall, our results match current linguistic literature, but do not match evolutionary computation hypotheses.

Introduction

In the linguistics field, there is a lack of valid computational models for child language acquisition. Several models have been proposed, all of which fail in certain respects. Evolutionary computation has been used successfully to teach a computer to parse a language, primarily through the use of finite state machines but has never before been used to study language acquisition. We propose a model that uses evolutionary computation, a technique that has never before been used to study language acquisition, and attempts to resolve several related open questions.

Modern linguistics is almost entirely based on Chomskyan theories of Universal Grammar (Yang, 2002). In brief, this states that there is something innate in humans that allows us to use language. Furthermore, language varies along specific, limited dimensions known as parameters, which can usually be set in a binary manner. Language acquisition, then, is the process of setting these parameters appropriately for the target language based on input sentences the child hears. To simplify, we assume that the child is able to unambiguously parse sentences and retrieve the relevant parameters.

Of alternative models (Yang, 2002), the most popular is the trigger model. In this model, sentences are discarded unless they conflict with the current grammar hypothesis. However, there are problems with this approach. First, this algorithm is a hill-climbing algorithm, and thus liable to getting stuck in a local maximum. Second, the trigger model is a brittle algorithm, prone to getting on the wrong track due to errata in the input data. Finally, trigger models predict sharp changes in which parameters are set in language production, which does not match actual data. Our own model uses triggers to start a change, but not to deterministically set the output.

Although nothing in linguistics is universally accepted, Chomsky's theory of Universal Grammar comes closest. The most common refinement to UG is parameterization, the idea that languages vary along certain limited dimensions, which are called parameters. Pearl and Lidz provide what is probably the best explanation of parameters, as well as propose their own version of the statistical model mentioned above (2011). One of our simplifying assumptions, that all

parameters come pre-set to a certain value, is based on (Hyams, 1989). This greatly simplifies the application of evolutionary computation to the problem.

One of the trickier parts to model is the presence of ambiguous triggers. An unambiguous trigger is a sentence or a clause that is only grammatical in grammars that have a parameter set a certain way. An ambiguous trigger is one that can be set in different ways in different grammars. Fodor(1998) explains that to deal with unambiguous triggers, the child must first be able to detect the ambiguity. However, her model is the brittle trigger model described above. In our model, any incorrectly-parsed sentences due to ambiguity can be subsumed under the error checking we do. Alternatively, Sakas/Fodor(2012) say that there are enough unambiguous triggers to learn from without using ambiguous triggers, in which case we can simply discard ambiguous sentences.

We here test various objections to computational models of language acquisition by determining if our model can converge when these objections are addressed. First, our model should be sufficiently resilient as to allow errors in input sentences to not prevent convergence. Second, our model should be able to converge given memory limitations as demonstrated by a smaller number of agent grammars. Third, our model should demonstrate the relative rarity of certain parameters across all sentences by converging even when the number of parameters set per sentence is low. In addition, we here study the effect of partially unset parameters in language. For example, certain kinds of movement simply don't occur in Korean (Han, Musolino, & Lidz, 2007), and are therefore considered to be unset in the language. However, (Hyams, 1989) shows that these are still set within a person's internal grammar. Our model will allow grammars to have parameters set which are not set in the language.

There is not a definitive or dominant model in the grammar hypotheses space. Our model, given some interpretations as to its physical representation, does not contradict existing linguistics research. As such, it could be significant both in giving insight into relevant settings for optimal language acquisition rates, for introducing new computational methods into the field, and for competing in the theory pool of existing models.

Methods

We began our implementation with a canonical genetic algorithm adjusted to match linguistic theory. First, we define a language as a set of binary parameters, some of which may be unset, represented by a bit string with the unset parameters set to *. At initialization, the agents are each represented by their genotype, a grammar of randomly set bits of the same length as the language. During each round, we generate a sentence, represented by a bit string the same length as the language which has a subset of the language's parameters set. This models a sentence which has been correctly lexed and parsed to generate a parameter set list. The phenotype is the ability of an agent's grammar to match the input sentence's parameters, as measured by the reverse Hamming distance ([length of the language] – [Hamming distance]) between the sentence and the grammar. Based on this fitness value, parents are selected for reproduction using Roulette wheel selection, then mutation and recombination will be carried out at the frequency inputted followed by selection using the method and count inputted. This algorithm is diagrammed in Figure 1.

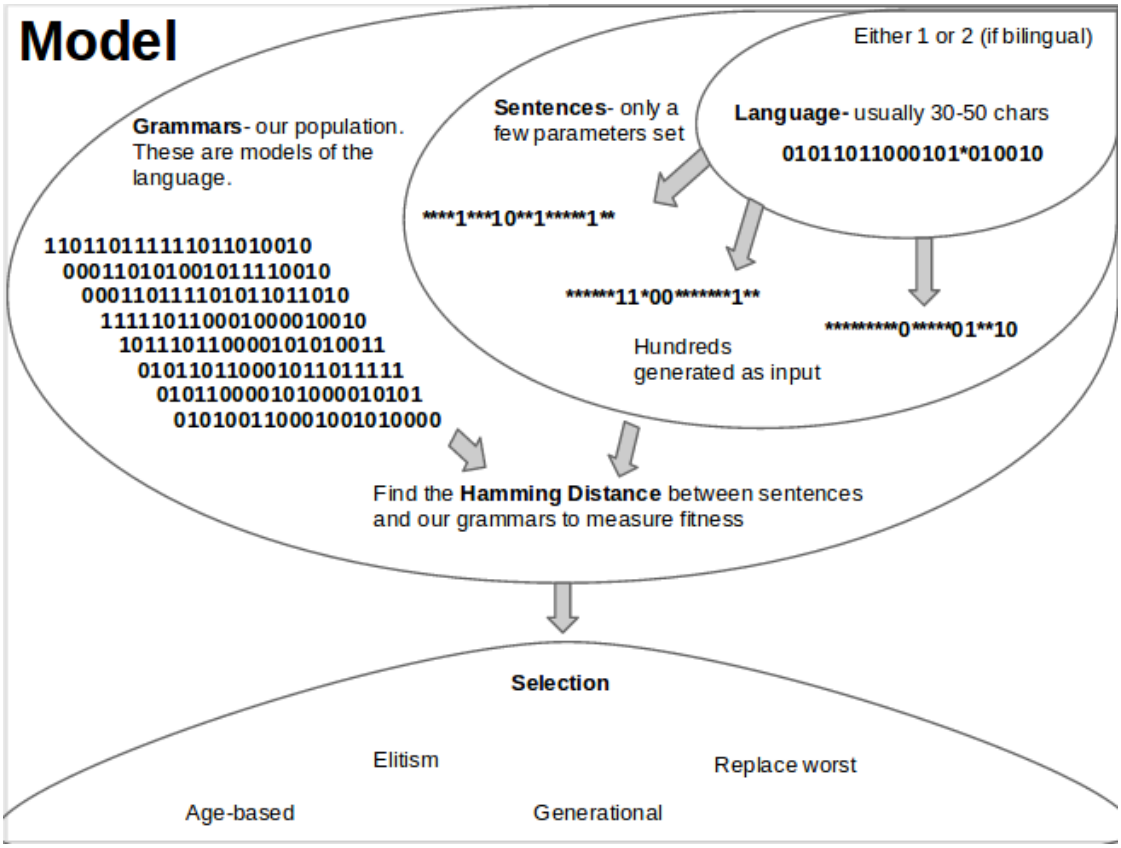

Figure 1. Schematic of our evolutionary algorithm.

The final fitness of a run was calculated by measuring the inverse hamming distance between the language and the consensus sequence of all final grammars. For termination, we did not use a final fitness threshold as our fitness function used for reproduction was different from the final fitness function. Thus, we instead ran the algorithm for a set number of generations and determined success or failure by the consensus sequence fitness at the end of that run. This has the added advantage of mimicking the critical period in child language development, where the amount of time a child has to learn language is independent of its success.

In order to determine the ideal inputs for language acquisition, we gave as inputs to the script the following settings: length of language, percent of language parameters unset, number of grammars, number of parameters set per sentence, number of recombination points, mutation probability, survivor selection method, and survivor replacement count. These inputs were based on current literature. While most studies claim there are between 30 and 50 parameters, there is no consensus on what they are (Roberts & Holmbert, 2005). Further, given the complexity of language, it's liable to be in the thousands (Newmeyer, 2004). It follows that the number of parameters set per sentence is also unknown, as is the number of parameters unset in individual languages. To simplify things, we examined language lengths from 30 to 50, sentence lengths from 3 to 12, and 0-2% unset parameters. Little is known about the number of internal grammars each person stores, and so we tested using 20-200 grammars as our population size. We also varied the mutation rate between 0-10% and the recombination rate between 0-20 points. We used elitism, generational, age-based, and replace-worst survivor selection methods, with a varying range for the number or parents to replace (age-based and replace-worst) or preserve (elitism).

In addition, variants of the script were run to test relevant linguistic phenomena. First, in bilingualism, 2 languages were created, and the grammars were compared to one of the two languages during each round. Final fitness was measured by forming a consensus sequence from the set of grammars whose inverse hamming distance was greater for one language than for the other. Second, in negative feedback, each sentence had a single error, and any match between the grammar and the sentence was awarded a fitness of zero. This modeled language acquisition through negative reinforcement, in which a child would be scolded for forming an incorrect sentence, a technique regarded by the linguistics community as ineffective as the child interprets the entire sentence as incorrect. Finally, in errors, a sentence was generated with incorrect parameters with a certain probability.

The genetic algorithm was coded in Ruby 1.8.6 entirely by team members and executed through a bash shell script. All code can be found on our public repository at https://bitbucket.org/relh/artificial-life-language-acquisition.

Results

Our first task was to determine whether our termination condition reflected the fitness of the grammars. We compared the values of the final consensus fitness to those of the best final fitness and the best overall fitness, or the grammar that best matched the language over all runs. We first observed that the best final and best overall fitness were correlated (Spearman, rho=0.6594, $p<10^{-64}$), but that best overall fitness was higher than best final fitness (Wilcox, $p<10^{-64}$), suggesting that the best grammar does not always remain until the final round, but instead is broken up by mutations and recombination (Figure 2). However, final consensus fitness was not correlated with either metric, but was higher than best overall fitness (Wilcox, $p<10^{-64}$) or best final fitness (Wilcox, $p<10^{-20}$). Despite this, we chose to use the consensus sequence as our measure of success of a run as it is a more linguistically relevant choice. Finally, using the consensus fitness, we found that our algorithm did not prematurely converge; when allowed to run for 5000 rounds, all runs achieved a maximum fitness score. This is to be expected, as our search space has no local maxima because our language is generated randomly and the entire language serves as input to the sentences. In addition, we achieved convergence given sufficient rounds with as few as 5 grammars.

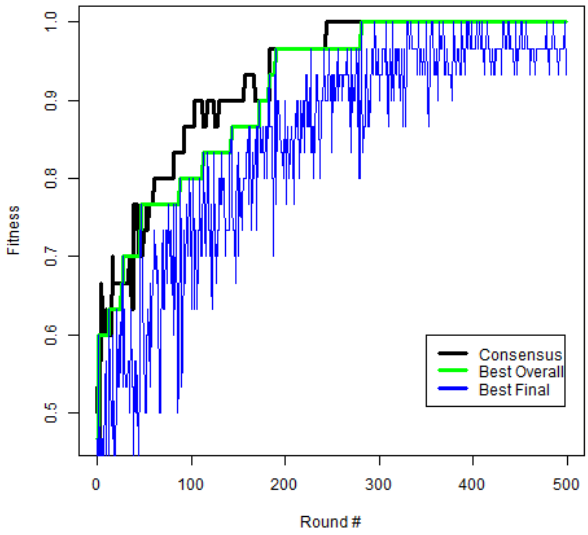

Figure 2. A sample trajectory of a single run. The final consensus fitness fares better than best overall fitness, which in turn fares better than best final fitness. Best final fitness fluctuates and does not necessarily improve as it reflects the state of individual grammars at each round rather than the average across multiple grammars.

Our next task was to determine the ideal inputs to produce the highest final consensus fitness. Language length was inversely correlated with fitness (Spearman, rho=-0.7581, $p<10^{-64}$) while the number of grammars was correlated with fitness (Spearman, rho=0.6423, $p<10^{-64}$). Although more unset parameters meant higher fitness (Spearman, rho=0.1214, $p<10^{-7}$), we still achieved maximum fitness even with all parameters set. Lower mutation rate was also correlated with fitness (Figure 3 and Table 1), but neither recombination rate nor number of parameters set in each sentence affected the final outcome, suggesting that measuring and maintaining longer blocks of grammars together did not convey a fitness advantage. In addition, we observed that crossover was a better genetic operator when there were fewer agents and more time. When increasing the number of grammars from 20 to 120 with a fixed language length of 180, we found that the number of recombination points was only correlated with performance when there were 20 grammars (Table 2).

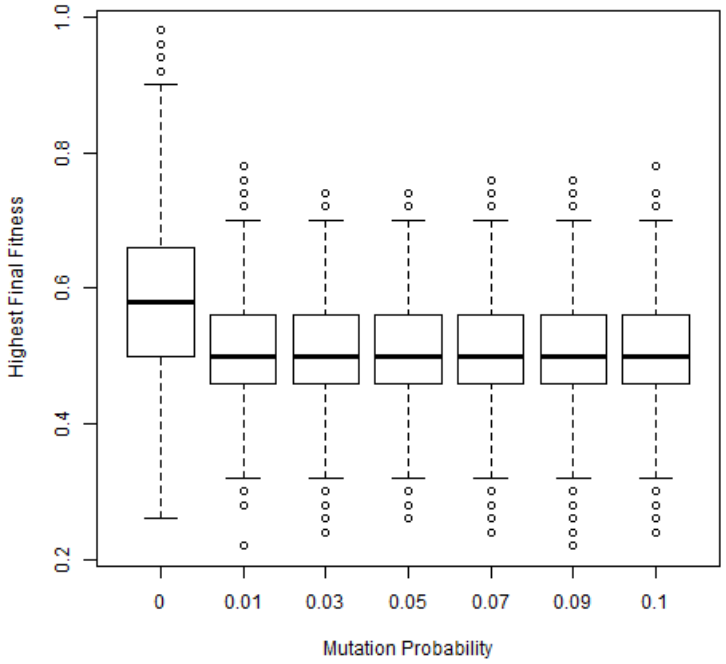

Figure 3. Box plots of final fitness scores across various mutation probabilities.

Table 1. Spearman correlation between mutation rate and final consensus fitness over different script variants.

	Regular	**Bilingual**	**Negative**	**Error**
rho	-0.2909	-0.2787	-0.0119	-0.3146
p-value	10^-31	10^-114	0.4	10^-147

Table 2. Spearman correlation between number of recombination points or mutation rate and final consensus fitness over different population sizes.

# Grammars	**120**	**100**	**80**	**60**	**40**	**20**
Recombination points rho	0.0101	0.0020	0.0100	0.0397	0.0473	0.0426
Recombination points p-value	0.6866	0.9347	0.6880	0.1160	0.0886	10^-13
Mutation rate rho	-0.3498	-0.2403	-0.1756	-0.1829	-0.1239	-0.1458
Mutation rate p-value	10^-47	10^-22	10^-12	10^-13	10^-7	10^-9

Finally, we examined survivor selection methods, and found that generational selection was the most successful. Specifically, given a language of length 20 and a population size of 30, generational of any size (all parents were replaced regardless of the inputted selection size) outperformed most other strategies (Wilcox, $p<0.05$, Figure 4). Other methods very similar to generational, including elitism with 1 parent retained and age-based with 20 children replaced, performed equally well. This trend held regardless of whether the best final fitness or the consensus fitness was examined, thus the effect of generational selection was not just due to the consensus metric. This suggests that the most effective way for our model to learn is to apply genetic operators to all parents and replace all the children.

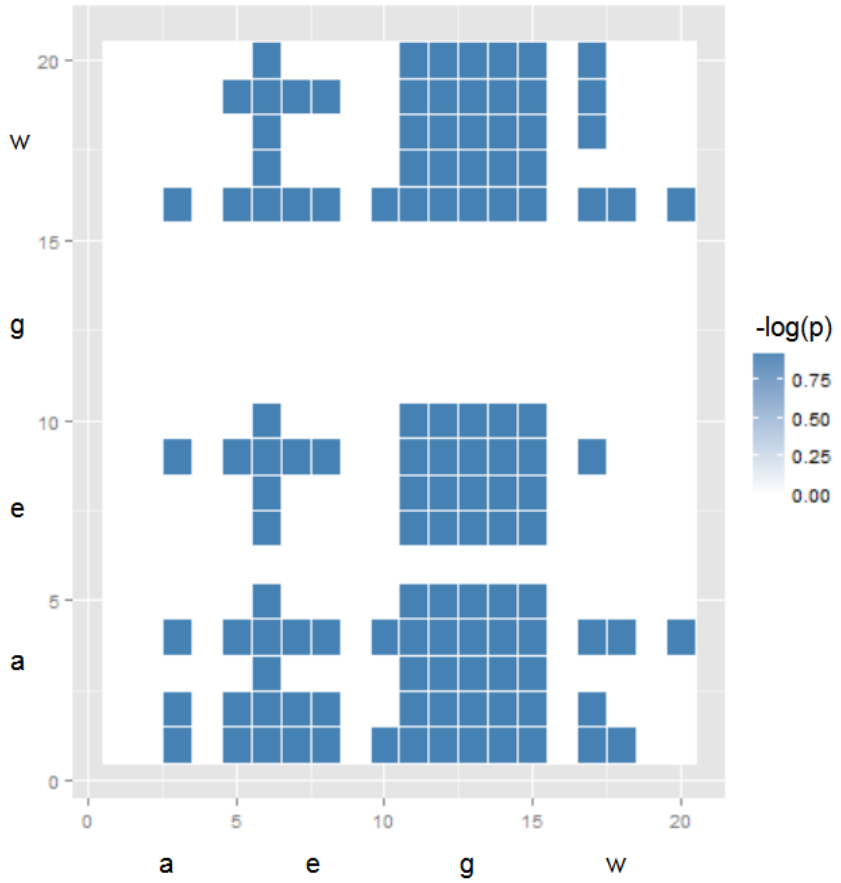

Figure 4. Heat map of survival selection strategies. Each set of 5 columns or rows is a single selection method (from left to right, age-based, elitism, generational, and worst), while each row within each set of 5 is a selection count (increasing order). A colored block means that the column is significantly better (via Wilcox test, $p<0.05$). E.g., block 5,9 is colored, meaning that age-based selection with 20 survivors is significantly better than elitism with 15 survivors.

Our final task was to compare variants of our script to the performance of the original (Figure 5). For all variants, performance was worse than the original (Wilcox, $p<10^{-31}$ for bilingualism, $p<0.0003$ for negative feedback, and $p<10^{-19}$ for error). For bilingualism, because all grammars are compared against both languages, improvements are lost and gained over and over again, leading to improvement followed by premature convergence. Interestingly, making each language the XOR of the other did not improve performance; this is because regardless of genetic operators and survival strategy, we still could not cluster the grammars according to the language they most resembled. Instead, fitness eventually converged on all grammars having the same fitness for both languages. When comparing whether certain inputs improved performance, we found that the same trends as were present for the original script held with the exception of negative feedback. For negative feedback, no parental selection strategy was more successful than the others and mutation rate not correlated (Table 1). This may have been because negative feedback performed the worst of all the variants, thus no genetic operator or selection strategy could rescue fitness.

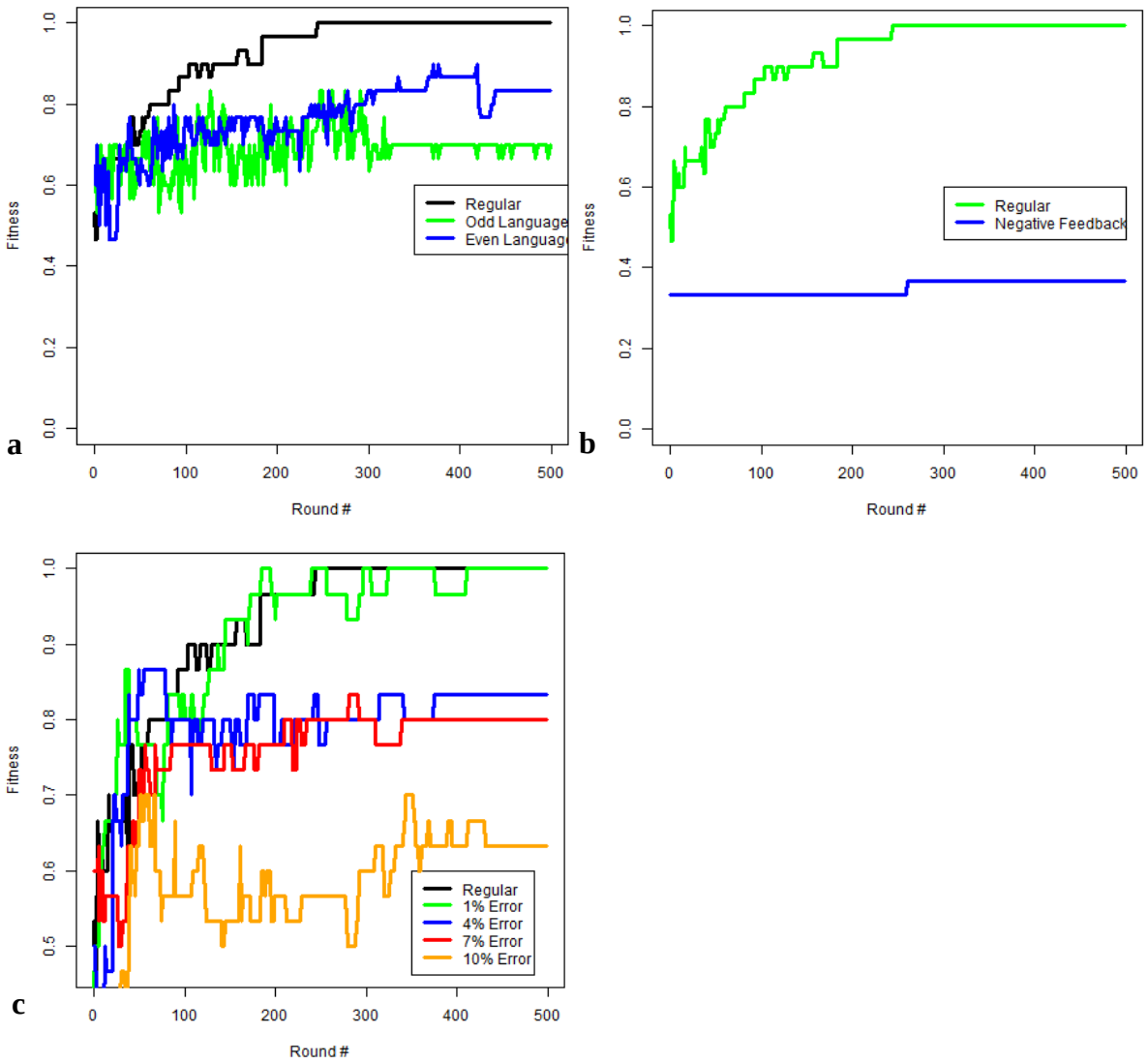

Figure 5. Sample trajectories of consensus fitness for bilingualism (a), negative feedback (b), and error (c) given the same inputs of language length 30, 0 parameters unset, 20 grammars, 9 parameters set by each sentence, 5 recombination points, no mutation, and generational selection.

Discussion

We here demonstrate an evolutionary model of language acquisition which avoids some of the common pitfalls of such computational models while still matching current linguistic theory. First, we were able to use a consensus sequence of the final grammars to achieve maximal fitness without premature convergence, even with as few as 5 internal grammars, with as many as 180 parameters, with all parameters set, and with as few as 3 parameters set per sentence. Our use of the consensus sequence to measure fitness better matches linguistic reality than most computational models because it suggests that children merge multiple internal grammars to produce sentences rather than selecting one based on prior knowledge of its fitness. Convergence on maximal fitness with few grammars models memory limitations and demonstrates that our model is realistic, as using a number of internal grammars near the low end of the proposed actual range would allow the child to have more memory available for other

tasks such as developing a lexicon. Convergence with a language above the proposed actual range of lengths and with all parameters set provides strong support for our model's ability to learn language effectively. Finally, we used few parameters per sentence to model the relative rarity of certain parameters in real language and we still able to achieve good language acquisition. Future research should extend our algorithm to implement parameters which are dependent upon each other, parameters which appear in sentences with a certain individual probability, and much larger language lengths up to thousands of parameters as proposed by Newmeyer (2004).

Next, we found that the best operators and selection strategies matched current linguistic theory but not evolutionary computation hypotheses. No mutation was best, which matches linguistic reality by suggesting that changes to internal grammars are based on evidence, not just random. Interestingly, neither recombination nor number of parameters set in each sentence affected the final outcome, suggesting that measuring and maintaining longer blocks of grammars together did not convey a fitness advantage, which appears to contradict the building block hypothesis. In addition, we observed that the evolutionary computation hypothesis that crossover is a better genetic operator when there are more agents and less time while mutation is the more effective operator when there are fewer agents and more time also did not apply here. This is likely because mutation was always a detrimental force, thus recombination rate was used as the primary genetic operator, and the performance improvement from increased recombination was only observed when it was needed most when fewer grammars were available. Last, we found that generational selection was the most effective, suggesting that children alter all of their grammars when learning language rather than the few that performed the best at parsing the sentence. Due to limited computer resources, it was impossible for us to test all possible input setting combinations. Instead, we assumed that inputs were independent and tested them across a range of values, so it is possible that we were unable to find the ideal settings.

Finally, we examined variants of our basic script to demonstrate whether our model is robust in modeling alternative forms of language acquisition. First, the reduced performance of the variants as compared to the original script matches with current linguistic theory. Bilingual children are less fluent in either language than monolingual children, there is no evidence that negative feedback assists language acquisition whatsoever, and children exposed to high error rates are less fluent than those who are not. For bilingualism, we were unable to cluster grammars according to the language they most resembled and so fitness converged on all grammars having the same fitness for both languages. This suggests that children must be able to tell two languages are distinct from each other, which agrees with current linguistic knowledge, or else they will develop 1 “meta-grammar” that works somewhat for both languages. Future research should attempt to implement a model for bilingualism which achieves clustering of languages, perhaps through age-based selection in order to alternate which language's grammar is being updated and through a marker which tags a grammar for a specific language after it is sufficiently close to that language. For negative feedback, our model was unable to learn using this reinforcement technique, which matches current linguistic theory. For error, our model was able to recovery from error rates up to 10% and still converge on maximal fitness given enough rounds. Thus, our model avoids some of the common objections to computational models of language acquisition, matches current literature, and proposes potential answers to current open linguistic questions.

References

Fodor, Janet 1998. Unambiguous Triggers. Linguistic Inquiry, vol. 9, number 1. Massachusetts Institute of Technology.

Chung-hye Han, Julien Musolino, Jeffrey Lidz. 2007. V-raising and grammar competition in korean: Evidence from negation and quantifier scope. Linguistic Inquiry.

Hyams, Nina 1989. The Null Subject Parameter In Language Acquisition. The Null Subject Parameter. Kluwer Academic Publishers, O. Jaeggli and K Safir editors.

Newmeyer, Frederick J. 2004. Against a parameter setting approach to typological variation. Linguistic Variation Yearbook.

Lisa Pearl and Jeffrey Lidz 2011. Parameters in Language Acquisition. The Cambridge Handbook of Biolinguistics. Cambridge, UK. Cambridge University Press, 129-159.

Ian Roberts, Anders Holmberg. 2005. On the role of parameters in universal grammar: A reply to Newmeyer. Organizing Grammar: Linguistics Studies in Honor of Henk van Riemsdijk.

William Sakas, Janet Fodor 2012. Disambiguating Syntactic Triggers. Language Acquisition. Psychology Press.

Yang, Charles. 2002. Knowledge and learning in natural language. Oxford University Press.