A Stochastic Model for the Vocabulary Explosion

Colleen C. Mitchell (colleen-mitchell@uiowa.edu)
Department of Mathematics, 225E MLH
Iowa City, IA 52242 USA

Bob McMurray (bob-mcmurray@uiowa.edu)
Department of Psychology, E11 SSH
Iowa City, IA 52242 USA

Abstract

During the second year of life, the rate at which most children learn words accelerates dramatically, the so–called "vocabulary explosion". Most accounts posit changes in the child or specialized learning mechanisms to account for this sudden change. However, recently McMurray (2007) demonstrated that acceleration is a mathematical consequence of parallel learning and the statistical distribution of word difficulty across the language. We generalize this model by developing a stochastic version. It demonstrates that the gradual nature of learning is critical for producing acceleration, and given sufficient gradualness, virtually any distribution of word difficulty can yield acceleration. Thus, the vocabulary explosion may be even more mathematically robust than previously thought.

Keywords: Vocabulary Explosion; Stochastic Model.

Introduction

One of the most dramatic changes during language acquisition is the spurt of word learning that typically occurs in the second year of life. During this time, the rate of word learning accelerates dramatically (Figure 1). Lexical acquisition norms suggest that between 11 and 15 months of age, children acquire on average 2.7 new words/month, while from 17-21 months, they average 27.9 (Dale & Fenson, 1996).

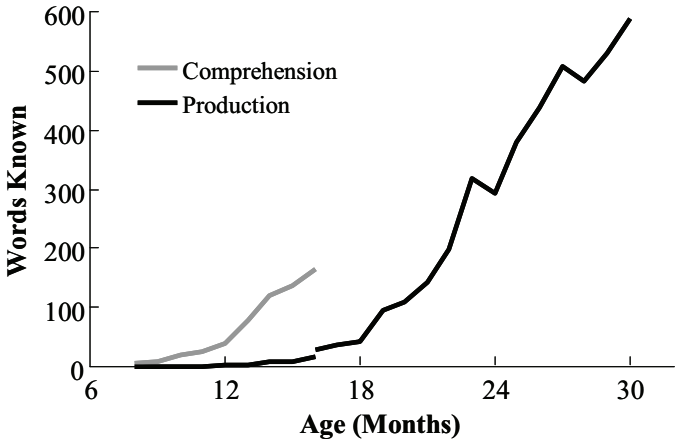

Figure 1: Lexical acquisition norms from the MacArthur-Bates Communicative Development Inventory (MCDI).

The cause of this so-called vocabulary explosion is the subject of much debate. Initial explanations posited a unitary change in the child: the sudden realization that things have names (Reznick & Goldfield, 1992), the onset of categorization abilities (Gopnik & Meltzoff, 1987), or the acquisition of word learning constraints (Mervis & Bertrand, 1994).

Such things would suggest a one-time, stage-like increase in the rate of learning. This was ruled out by Ganger and Brent (2004) who analyzed the lexical acquisition functions of 38 children. All but five showed smooth acceleration and no sudden spurt. While this rules out a one-time developmental event as the cause of the vocabulary explosion, it nevertheless raises the question as to what causes the more general acceleration seen throughout childhood.

A second class of explanations can be termed bootstapping or leveraged learning approaches. These include mechanisms like segmentation (Plunkett, 1993), mutual exclusivity (Markman, Wasow & Hanson, 2003) or syntactic bootstrapping (Gleitman & Gleitman, 1992). In all three cases, existing words in the lexicon are used to help acquire new words. Thus, as each word is learned, these specialized mechanisms become more powerful, enabling faster word learning.

Such approaches are supported by computational work (Van Geert, 1991). Simple exponential growth systems can model the vocabulary spurt, but only if the rate of change is a function of the number of words already learned.

However, recent work by McMurray (2007) suggests an alternative: that acceleration is a mathematical by-product of known properties of word learning. Acceleration will occur as long as words are learned in parallel and there are fewer easy words than moderately difficult words. This model suggests that specialized word learning mechanisms are not needed to explain acceleration in vocabulary growth (although they may exist to solve other problems).

We extend this model to examine the role of learning history in creating acceleration. We derive a generalized stochastic version of the deterministic model presented in McMurray (2007). This model takes as a parameter r, the amount of history with a word that is required to learn it. Using this model, then, we ask about the relationship between learning history and acceleration. We demonstrate that given sufficient required history, it is possible to relax the criterion in McMurray (2007): acceleration in vocabulary growth is even more guaranteed.

The Model

The original McMurray (2007) model employed a simple deterministic system to model word learning. Each word in the model is assigned a degree of difficulty D_i (measured in time-to-acquisition). Then on each time–step, the model accumu-

lates one point for each unlearned word. When a word crosses its threshold, it is considered learned.

This simplistic approach to learning was intended as a minimal instantiation of parallel learning in which the mathematics of learning itself could not cause acceleration – each word accumulates points at a constant rate. This simple learner implements no bootstrapping type mechanisms – the current contents of the lexicon cannot influence the rate at which unlearned words accumulate points. While this may be implausible theoretically, it is a useful null-model – if acceleration can be seen in this model (without specialized learning devices), it will certainly appear in more nonlinear approaches.

A critical factor in the McMurray (2007) model was the distribution of difficulty, g, across the lexicon. This describes the relative number of easy, moderate and hard words. Given, g, the number of words known at any given time, T, is simply the integral of that distribution from 0 to T.

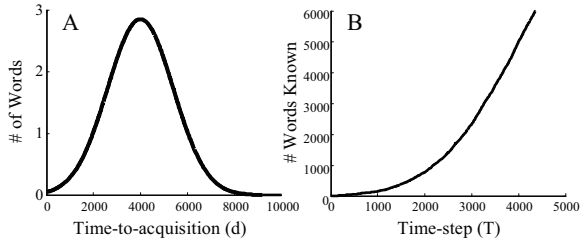

Figure 2: The model presented in McMurray (2007). A) The Gaussian Distribution of time-to-acquisition. B) number of words known as a function of time-step.

The most likely distribution of difficulty is Gaussian (Figure 2A). Word-difficulty is likely to be the sum of many factors (e.g. phonological complexity, frequency in the child's environment, syntactic category, etc). Since these factors will be largely independent, their individual distributions sum to a Gaussian (by the Central Limit Theorem). However it is important to note that any distribution of difficulty which includes only few easy (small D_i) and more moderate (mid range D_i) will show acceleration in learning.

This model demonstrates acceleration in word learning (Figure 2B), despite a constant accumulation of points. Thus, it was concluded, as long as words are learned in parallel and vary in difficulty in this way, acceleration is guaranteed.

A Stochastic Version

One could reasonably argue that this effect is due to the determinism of the model. A model incorporating some degree of randomness would be more theoretically valid and perhaps more generalizable. We therefore introduce a stochastic version of the model which shows not only that the acceleration in word learning is extremely robust but also reveals the importance of learning history.

We start by describing a discrete version of the deterministic model and then generalizing it to a stochastic version. We number each word $i = 1$ to N where N is the number of words to be learned. Each word is assigned a difficulty D_i so that the distribution of difficulties approximates g. In the deterministic case, at each time–step one point is added to each word. When a word reaches its threshold, D_i, it has been learned.

To make this model stochastic, we again number each word $i = 1$ to N. Next, each word is assigned a probability, p_i. We think of p_i as the probability that the child is exposed to the i^{th} word in any one time step. Thus p_i is proportional to the frequency of the word. Each word is also assigned a threshold, r_i, the number of repetitions required before a word is learned. Then at each time step, a word has probability p_i of gaining a point and it is learned once it has accumulated r_i points. We can think of r_i as the difficulty *independent of its frequency of occurrence*. Thus, the original deterministic model can be described as the special case in which all p_i are 1 and r_i corresponds to word difficulty, D_i.

The analysis of the stochastic model is done by first investigating the time to acquisition of a single word. We compute the cumulative distribution function or c.d.f, $F_i(T)$, for the time of acquisition of that word. So $F_i(T)$ is the probability that the i^{th} word has been learned by time T. Next we compute the expected number of words learned as a function of time, $L(T)$. Since the acquisition of the words is independent, $L(T)$ is the sum of these distributions, F_i. That is

$$\begin{aligned} L(T) &= \text{expected number of words learned by time } T \\ &= \sum_i P(\text{the } i^{th} \text{ word was learned by time } T) \\ &= \sum_i F_i(T) \end{aligned}$$

The analysis and discussion of this stochastic model is done in two cases. In both cases, r (the amount of required exposure) is constant across all words. We adopt this simplistic assumption for ease of analysis–in real language words may vary dramatically in difficulty but currently there are no existing metrics. Work in progress is undertaking analysis of cases in which r varies.

In the first case, $r = 1$, the word is learned the first time it is heard and the learner has no need to keep any history of past experience with the word. We therefore call this the history-free case and it instantiates a somewhat implausible one-shot learning. In the second case, $r > 1$, the word is learned only after r points are accumulated. In this case, the accumulated history of the word is important. The cases are treated separately since they have fundamentally different behavior which highlights the importance of gradual learning in the vocabulary explosion.

History-Free Case: $r = 1$

The first simulations and analysis contrast the deterministic model with the stochastic ($r = 1$) model. In this simplest case, at each step there is a probability p_i of acquiring a point and hence learning the word and no dependence on previous time steps.

The number of time steps until the word is learned, X, is a Geometric Random Variable with parameter p_i. This means

that the expected time to acquisition is $E(X) = \frac{1}{p_i}$. Since we know the expected time to acquisition, we can chose the values of p_i such that the time to acquisition fits any distribution we would like, and allows us to compare this model to the deterministic model.

Simulations A series of simulations implemented the deterministic and stochastic models to compare their performance. Each model was run 10 times using the representative parameters of McMurray (2007). For each model, a 10,000 word lexicon was initialized. Each word was given a difficulty D_i (time-to-acquisition) randomly chosen from a Gaussian distribution, g, with a mean of 4000 and standard deviation of 1400. For the stochastic model, these were converted into probabilities ($p_i = \frac{1}{D_i}$). Then at each time-step 10,000 random numbers (one for each word) were selected from a uniform distribution ranging from 0 to 1. Any word, i, whose random number was less than its p_i , was deemed learned and removed from further consideration. The deterministic model was identical to McMurray (2007).

Results are displayed in Figure 2. All of the deterministic models showed a period of slow growth followed by acceleration. In contrast, the stochastic models showed initially rapid learning which gradually tapered off.

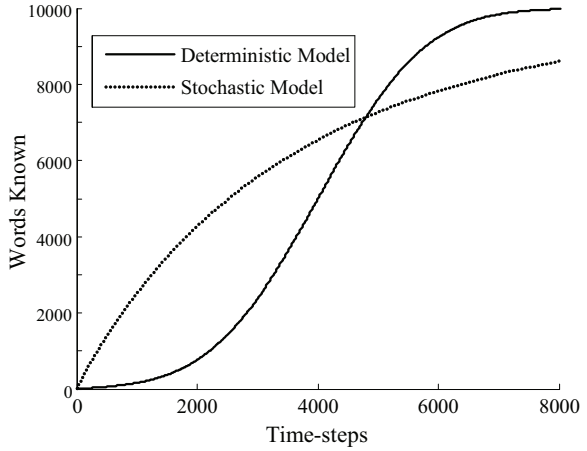

Figure 3: Number of words learned by each model as a function of time-step.

Analysis We begin with the analysis of a single word. For ease of notation, we temporarily drop the subscripts on F, D, p and r. As we've described, $E(X) = \frac{1}{p}$ is equal to D which is equivalent in the two models. What, then explains the discrepancies between them?

First, while the first moments of time-to-acquisition are equivalent, the second moments are not. In the stochastic model, the variance of the time to acquisition is $\mathrm{Var}(X) = \frac{1-p}{p^2}$ (it is zero in the deterministic model). Additionally, the coefficient of variation (a non-dimensional measure of the spread of a random variable) is $\sqrt{1-p}$. Since p must necessarily be small (otherwise the model will acquire the bulk of the lexicon within the first measurement period), the variance in time-to-acquisition (X) will necessarily be high. Thus, any single run of the model is quite unlikely to approximate the expected values of X (see Figure 4). In a sense, then, the extreme variance of this model prevents it from modeling difficulty to the same degree as the deterministic model.

The probability that the word is learned at time T is simply the probability that that word was not learned on the previous $T-1$ steps times the probability that it is learned on time T.

$$p^{(T)} = (1-p)^{T-1}p$$

Thus, the probability that a word is learned *by* time T is

$$F(T) = \sum_{i=1}^{T} p^{(i)} = 1-(1-p)^T$$

Intuitively, the rate of change of this function F at time T is the probability that the word is learned at time T. That is $\frac{\Delta F}{\Delta T} = p^{(T)}$. Since $p^{(T)}$ decreases as a function of T, $F(T)$ is concave down. Thus, the likelihood of learning any given word continually decreases over the life of the model. This then explains the failure of this model to show acceleration.

We have shown that the c.d.f. for each word is concave down. In order to obtain the function, $L(T)$ for the expected number of words learned up to time T, we sum the c.d.f.s over all the words in the lexicon. So

$$L(T) = \sum_{i=1}^{N} F_i(T).$$

Since the sum of concave down functions is concave down, we know that $L(T)$ is concave down. This can also be understood by looking at the expected number of new words learned on a particular step. At any time step, the expected number of new words learned is the sum of the probabilities p_i of the remaining words. As more words are learned, this sum must decrease. Thus the number of new words learned is expected to decrease at each time–step. Note that this is true regardless of how the probabilities p_i are chosen.

Discussion The stochastic model with $r = 1$ seems to argue against the findings of McMurray (2007). The vocabulary explosion is not guaranteed despite a graded distribution of difficulty (instantiated as p) and parallel learning. However, this case ($r = 1$) does not reasonably capture either of these important theoretical constructs.

First, the history free case does not clearly instantiate word difficulty. As we discussed, p_i in this model can be mapped onto difficulty in the deterministic model as it's reciprocal ($p_i = \frac{1}{D_i}$), making the expected time to acquisition of any word is equivalent across both models. However, the variance of the time to acquisition in the history free case is enormous. If the mean time-to-acquisition is 4000 time–steps this yields a variance of nearly 16,000,000 time-steps. As a result, the observed time to acquisition for any word is unlikely to be near the expected time to acquisition (Figure 4).

Second, this model builds in explicit deceleration. Since the probability of acquiring a word is independent of previous

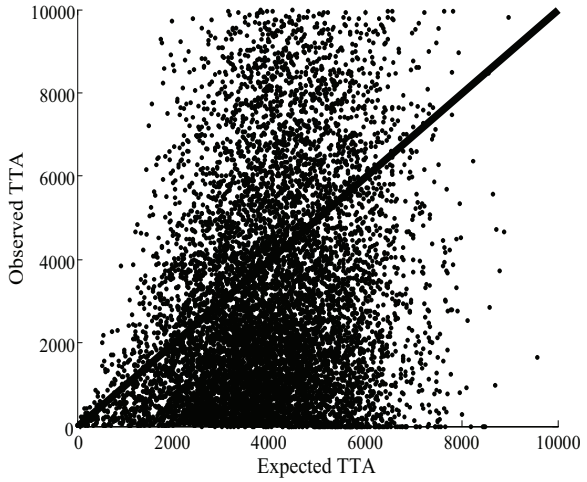

Figure 4: The observed time-to-acquisition as a function of the a priori expected time-to-acquisition for a representative run of the $r = 1$ model.

time-steps, the expected time to acquisition for an unlearned word increases over time. If a word starts with an expected time to acquisition of 4000 time-steps and is not learned after 1000, then its expected time to acquisition is now 5000.

Finally, while it can be construed as parallel, this model fails to model learning, since learning is a gradual process. While events can be construed as occurring at the same time, this does not constitute parallel learning because learning requires the gradual accumulation of material. Parallel learning requires maintaining this material across multiple items.

This is true in basic learning principles (Rescorla & Wagner, 1972) and connectionist accounts (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996). It also appears in word learning, where children learn the sound-pattern of a word before its meaning (Graf Estes, Alibali & Saffran, 2007), and much of learning takes the form of a slow elaboration (Capone & McGregor, 2005) or gradual improvement of skills (Fernald, Perfors & Marchman, 2006). Even so–called fast-mapping does not support this. Horst and Samuelson (in press) have demonstrated that novel names used in fast-mapping situations are not retained, even five minutes later. Parallel learning cannot be instantiated as a series of independent acquire-or-not events. By decoupling acquisition of a word from its developmental history in the system, the history free case fails to capture parallel learning.

Thus, the $r = 1$ case does not seem to instantiate any realistic developmental processes. However, we now contrast it with the model in which $r > 1$, to reveal the importance of gradualness in dictating the form of growth.

General Case: $r > 1$

Gradual learning can easily be incorporated into this stochastic model by setting r to any value greater than 1. In this case, the value of r represents the degree of gradualism, or the amount of history that the child must have with a word to acquire it. Once this property is built into the stochastic model, acceleration returns under virtually all circumstances.

In this model, at each step there is a probability p of acquiring a point and the word is learned once r points have been acquired. The number of time steps until the word is learned (X) is a Negative Binomial Random Variable with parameters (r,p). The expected time to acquisition is $E(X) = \frac{r}{p}$ and again, we can compare with the deterministic model by setting r and p such that $E(X)$ is the time to acquisition.

Simulations To test this, the prior stochastic simulations were repeated, this time with a range of rs (2-20), which was constant for all words within a simulation. Difficulties were selected from an identical Gaussian distribution, and converted to the probability of receiving a point ($p_i = \frac{r}{D_i}$). This ensured that the mean time-to-acquisition was constant across simulations despite the change in r. 10 simulations were conducted for each r.

Figure 4 shows average vocabulary size as a function of time. It is clear that higher values of r show long periods of slow growth followed by acceleration. However, even $r = 2$ shows a period of slow growth followed by acceleration. Thus, a vocabulary explosion can be seen as long as $r > 1$.

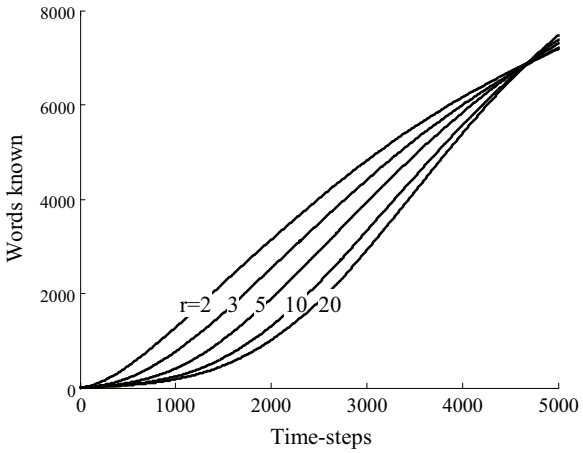

Figure 5: Vocabulary acquisition in the stochastic model as a function of time and r.

Analysis Our analysis begins, as before, with the analysis of a single word. Similarly to the previous analysis, the variance of the time to acquisition is $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$, giving a coefficient of variation of $\sqrt{\frac{1-p}{r}}$ Here, an increase in r (increasing the amount of required history) or an increase in p (forcing the model closer to the deterministic model) will reduce the variability. As a result, incorporating gradual learning makes any given instantiation of the model more likely to show effects of word difficulty.

The probability that a word is learned at time T corresponds to (1) the $T - r$ steps in which no points are earned each with probability $(1 - p)$ and (2) the r steps where points

are earned. This gives a probability:

$$p^{(T)} = \frac{(T-1)!}{(T-r)!(r-1)!}(1-p)^{T-r}p^r.$$

Thus the probability that a word is learned by time T is $F(T) = \sum_{i=1}^{T} p^{(i)}$. The rate of change of this function F at T is $\frac{\Delta F}{\Delta T} = p^{(T)}$. A calculation shows that $p^{(T)}$ is increasing for $T \leq \frac{r-1}{p}$ and decreasing for $T > \frac{r-1}{p}$. Therefore, $F(T)$ is concave up for $T \leq \frac{r-1}{p}$ and concave down for $T > \frac{r-1}{p}$. Any value of r greater than 1 will this yield some portion of initial acceleration (concave-up).

Now that we have computed the c.d.f. for a single word, we can sum over the lexicon to obtain the function $L(T)$ for the expected number of words learned by time T, $L(T) = \sum_{i=1}^{N} F_i(T)$. This sum is in general very difficult and except in simple cases must be computed numerically.

Discussion This demonstrates that even when $r = 2$ (minimal history), acceleration is guaranteed for some range of time. If $r > 1$, the probability of having learned a word at any given time accelerates. Larger values of r increase the suddenness of the acceleration; the length of the acceleration phase (Figure 5); and the likelihood that the observed time to acquisition matches the expected. Thus, instantiating history creates a non-deterministic model that captures both parallel learning and difficulty, and hence, shows acceleration. A few special cases deserve attention.

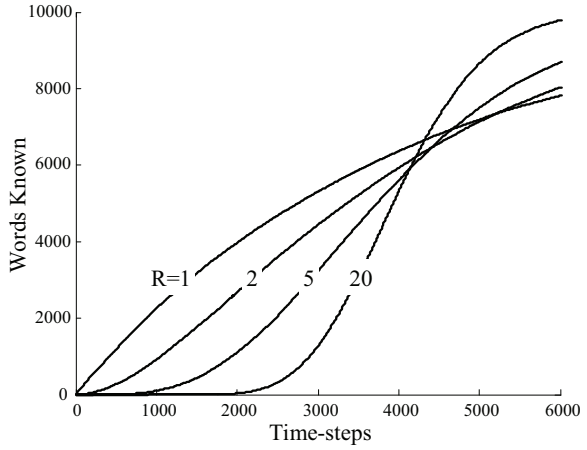

Figure 6: Results of a stochastic model with a constant p (and by consequence time-to-acquisition, D) across all words.

An implausible but theoretically illuminating case occurs in a model in which all words have the same difficulty and the same frequency. Then $L(T) = N \cdot F(T)$ and we know it has a nice period of acceleration. Figure 6 shows simulations of a stochastic model in which p_i was constant across the $10,000$ word lexicon, and is chosen so that all cases have the same expected time to acquisition as previous simulations. It is clear that at $r = 1$, no acceleration can be seen. However, at all $r > 1$, there is acceleration, and at higher rs, the spurt is quite dramatic. This explicitly violates one of the two assumptions

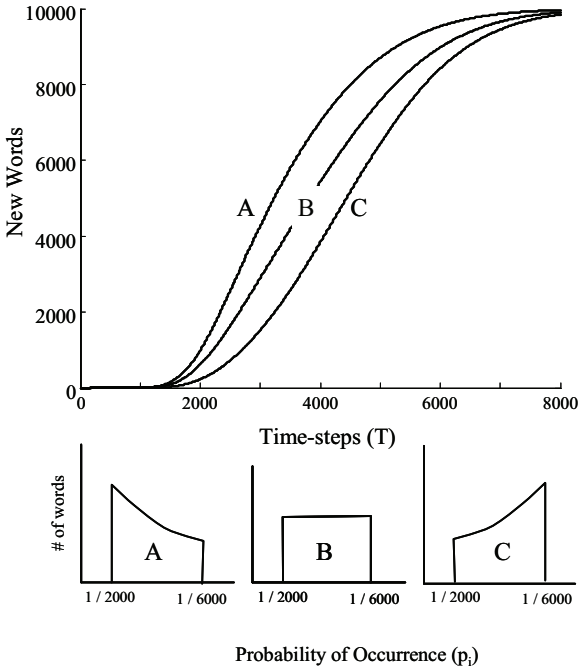

Figure 7: A comparison of the stochastic model with $r = 20$ when the distribution of probability, p_i, is A) a violation of McMurray, 2007, B) uniform and C) consistent with McMurray, 2007. The acceleration phase is longer for consistent distributions.

made by McMurray (2007) since the distribution of word difficulty does not monotonically increase.

This powerful example suggests that the vocabulary explosion may be more robust than McMurray (2007) hypothesized. It shows that while the precise shape of $L(T)$ depends in complicated ways on the distributions of frequencies (p) and required history (r), the vocabulary spurt will exist for a wide range of reasonable choices, *even if the difficulty distribution is not monotonically increasing.*

Simulations have also been conducted for various distributions of p_i with r assumed to be the same across all words. Figure 7, for example, demonstrates the acquisition of a model when the distribution of p_is (frequency) was either in accordance with McMurray, 2007 (Figure 7, curve C), a violation (curve A), or flat (curve B). When $r = 20$, all three models show acceleration, even the model that directly violates these assumptions. However, the shape of the distribution plays an important role in the length of the slow period preceding the acceleration.

We can also show that in the limit as r and p increase with the distribution of expected time to acquisition fixed, the deterministic model is obtained, and the spurt is dependent on the distribution of difficulty.

Discussion

These simulations and analysis confirm that the broad framework laid out by McMurray (2007) is correct: acceleration in

word learning arises out of the mathematical regularities of parallel learning and variation in difficulty. The present work extends these findings by demonstrating that this was not due to the determinism of the original model. Any stochastic model that incorporates even the minimal amount of gradual learning will show acceleration in word learning.

The only case in which this was not true was the stochastic model in which learning occurred after a single exposure ($r = 1$). However, this model does not meaningfully capture learning, and it is not unambiguously parallel. The same results could be achieved by a model which randomly sampled each word sequentially. Thus, this does not appear to be a disconfirmatory case.

More importantly, our analysis of this stochastic model suggest that the degree of history required to learn a word is itself a factor in determining the acceleration observed in word learning. Specifically, when learning is very gradual, a substantial spurt can be observed, even when the distribution of easy and hard words does not conform to the assumptions of McMurray (2007) (e.g., if there are more hard words than easy words). Of course, gradualness interacts in complex ways with the distribution of difficulty, and word frequency, and future work must examine both the empirical and computational underpinnings of this interaction. However, the bottom line is that specialized mechanisms are even less necessary to account for acceleration than previously thought. There are multiple routes to the same end.

This model is sufficiently general that it can be applied to virtually any parallel learning system. While the issue of acceleration is theoretically important in vocabulary acquisition (Bloom, 2000), the mathematics presented here will also underlie many other domains of learning.

The vocabulary explosion is an incredibly dramatic developmental process. This has led to a large number of theories positing equally dramatic changes or learning devices on the part of the child. However, such things are not necessary to explain acceleration. Apparent acceleration will always appear in parallel learning systems, even when the fundamental learning processes are perfectly constant. The so-called vocabulary explosion is a mathematically robust phenomenon that will arise under virtually any parallel learning circumstances. There is no need to invoke more complex mechanisms to explain it.

Acknowledgments

The authors would like to thank Dan McEchron for assistance with figure preparation, and Larissa Samuelson, Prahlad Gupta, John Spencer and Gregg Oden for helpful comments on the development of this project.

References

Bloom, P. (2000) *How children learn the meanings of words.* Cambridge: The MIT Press.

Capone, N., and McGregor, K.K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech, Language, and Hearing Research*, 48, 1468-1480.

Dale, P., and Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments,& Computers*, 28, 125-127.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. and Plunkett, K. (1996) *Rethinking Innateness: a Connectionist Perspective on Development.* Cambridge, MA: The MIT Press

Fernald, A., Perfors, A., and Marchman, V. (2006) Picking up Speed in Understanding: Speech Processing Efficiency and Vocabulary Growth across the 2nd Year. *Developmental Psychology*, 42(1), 98-116

Ganger, J., and Brent, M. (2004) Reexamining the Vocabulary Spurt. *Developmental Psychology*, 40(4), 621-632.

Gleitman, L. R. and Gleitman, H. (1992). A picture is worth a thousand words, but thats the problem: the role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, 1, 31-5.

Gopnik, A., and Meltzoff, A. N. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development*, 58, 1523-1531.

Graf Estes, K.M., Evans, J., Alibali, M.W., and Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, 18, 254-260.

Horst, J.S., and Samuelson, L.S. (in press) Fast Mapping but Poor Retention in 24-month-old Infants. *Infancy*.

Markman, E.M., Wasow, J.L., and Hanson, M.B. (2003) Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, 47, 241-275.

McMurray, B. (2007) Defusing the childhood vocabulary explosion. *Science*, 317(5838), 631.

Mervis, C. B., and Bertrand, J. (1994). Acquisition of the novel name/nameless category (N3C) principle. *Child Development*, 65, 16461662.

Plunkett, K. (1993). Lexical segmentation and vocabulary growth in early language acquisition. *Journal of Child Language*, 20, 4360.

Rescorla, R.A and Wagner, A.R. (1972) A theory of Pavlovian Conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. Black & W. Prokasy WF (Eds.) *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton Century Crofts.

Reznick, J. S., and Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology*,28, 406413.

Ross, S. (1997). *A First Course in Probability* (pp. 162-167). Upper Saddle River, NJ:Prentice Hall

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98, 353.