Cognitive Science 33 (2009) 1503–1523

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01071.x

On Leveraged Learning in Lexical Acquisition and Its Relationship to Acceleration

Colleen Mitchell,[a] Bob McMurray[b]

[a]*Department of Mathematics, and Delta Center, University of Iowa*
[b]*Department of Psychology, and Delta Center, University of Iowa*

Received 10 September 2008; received in revised form 6 January 2009; accepted 24 February 2009

Abstract

Children at about age 18 months experience acceleration in word learning. This vocabulary explosion is a robust phenomenon, although the exact shape and timing vary from child to child. One class of explanations, which we term collectively as *leveraged learning*, posits that knowledge of some words helps with the learning of others. In this framework, the child initially knows no words and so learning is slow. As more words are acquired, new words become easier and thus it is the acquisition of early words that fuels the explosion in learning. In this paper we examine the role of leveraged learning in the vocabulary spurt by proposing a simple model of leveraged learning. Our results show that leverage can change both the shape and timing of the acceleration, but that it cannot create acceleration if it did not exist in the corresponding model without leveraging. This model is then applied to the Zipfian distribution of word frequencies, which confirm that leveraging does not create acceleration, but that the relationship between frequency and the difficulty of learning a word may be complex.

Keywords: Vocabulary explosion; Word learning; Fast-mapping; Acceleration; Language development

1. Introduction

1.1. What causes learning to accelerate?

Classic approaches to learning are typically decelerative. The power law of learning (Anderson, 1982; Logan, 1992; Newell & Rosenbloom, 1981), for example, posits that during motor learning tasks, subjects' reaction times slow in proportion to the log of

Correspondence should be sent to Colleen Mitchell, Department of Mathematics, 225E MacLean Hall, University of Iowa, Iowa City, IA 52242. E-mail: colleen-mitchell@uiowa.edu

time. That is, as time increases, there is a smaller effect of learning. Mathematical approaches like Rescorla–Wagner (Rescorla & Wagner, 1972) and connectionist approaches (Rumelhart, Hinton, & Williams, 1986) both model this explicitly: The amount of change due to learning is a function of the distance between the current state and some goal state or behavior. As the system approaches the goal, learning necessarily decreases. In both of these examples, the maximal learning rate occurs early and tapers off.

However, in a few cases learning accelerates. This is particularly apparent in language acquisition. A classic example of this is the so-called *vocabulary explosion.* Also known as the word burst, the naming explosion, or the word spurt, this typically occurs during the middle of the second year of life when toddlers transition from acquiring only a few words per week, to rates upwards of 20 words per week (Bloom, 1973; Fenson, Dale, Reznick, Bates, & Thal, 1994; Ganger & Brent, 2004; Reznick & Goldfield, 1992). While there is debate about the timing and universality of this phenomenon (Mervis & Bertrand, 1995; Rescorla, Mirak, & Singh, 2000), there is agreement that learning accelerates for most children around this time, and that it must eventually accelerate for all children if they are to reach their adult vocabularies in time (Bloom, 2000).

This contrast with classic learning theory suggests that word learning may operate by unique principles and their own specialized mechanisms. The purpose of the present study is to challenge this with a mathematical analysis of one class of such mechanisms, which we term *leveraged learning*. In a leveraged learning system, the acquisition of one or more items (e.g., words) can be leveraged to improve the learning of future material. This idea has been extensively applied to word learning, a domain in which multiple sources of leverage have been posited. Thus, our analyses are framed in terms of word learning, although our findings concern leveraged learning broadly.

We start with an overview of the empirical work on the vocabulary explosion and the proposed leveraged learning mechanisms. We then review a recent model of the vocabulary explosion (McMurray, 2007), which suggests that such mechanisms may not be needed and provides a framework for our analysis. Finally, we present a set of analyses in which we develop that model to account for leveraged learning. They uniformly demonstrate that while leveraged learning can change the rate of acquisition, it has little to do with the presence or absence of acceleration.

1.2. Acceleration in word learning

The vocabulary explosion is one of the most robustly observed phenomena in language acquisition (e.g., Bloom, 1973; Fenson et al., 1994; Ganger & Brent, 2004; see Fig. 1A). One class of explanations for it suggests that learning accelerates due to some unary change in the child. Examples of such events include the realization that objects have names (Reznick & Goldfield, 1992; Stern, 1924) or a reorganization of the conceptual space to permit a more robust representation of objects (Gopnik & Meltzoff, 1987, 1992). Other researchers posit a change in the learning mechanism as constraints on learning suddenly appear (Golinkoff, Mervis, & Hirsh-Pasek, 1994) or word learning shifts from a primarily

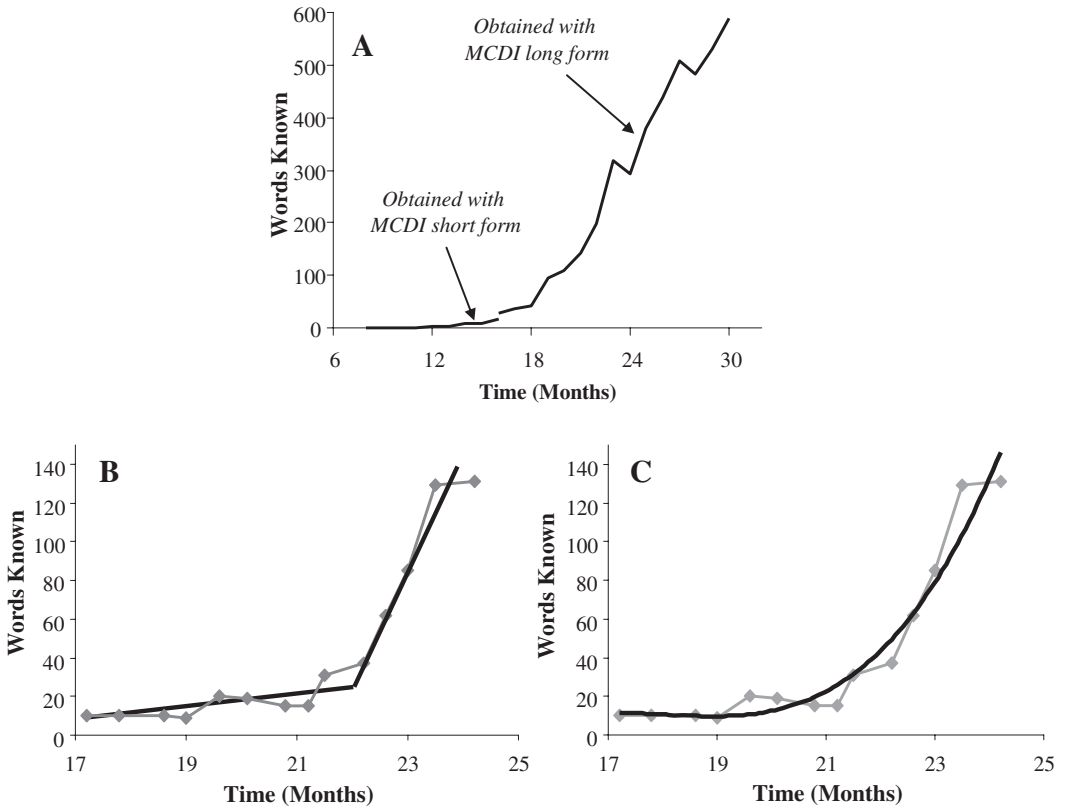

Fig. 1. (A) Number of words known as a function of time adapted from production norms of the MacArthur Bates Communicative Development Inventory (MCDI). Note that this uses separate assessments for 0–15 months and 16–32 months. Thus, the lines are shown broken. (B) Data for a typical child (adapted from Plunkett, 1993) fit with two lines. (C) Data for that same child fit with third-order polynomial (as in Ganger & Brent, 2004).

associative process to a more social/inferential one (Golinkoff & Hirsh-Pasek, 2006; Nazzi & Bertoncini, 2003).

Such explanations essentially model the rate of vocabulary acquisition as two lines (e.g., Fig. 1B), an early period of growth with a shallow slope, followed by an abrupt transition to a much steeper slope. Ganger and Brent (2004) recently evaluated such a model against the growth curves of 38 children and found that a single polynomial function (featuring smooth acceleration) was the better fit in 33 of 38 children. This is more consistent with models of word learning in which acceleration is an ongoing and continuous process, not a two-stage process (Fig. 1C).

1.3. Leveraged learning

One candidate for such a process is a class of mechanisms that we term *leveraged learning*. Here, as each word is acquired it provides information that aids in the acquisition

of future words. As more words are acquired, the net amount of leverage increases, and learning accelerates. Indeed, such a process appears implicitly in models like that of van Geert (1991) and Elman et al. (1996), which describe vocabulary acquisition as a simple differential equation in which the rate of acquisition is a function of the current lexicon size.

There have been a number of proposed mechanisms that fall into the category of leveraged learning mechanisms. Perhaps the most well known is fast-mapping by mutual exclusivity (Carey & Bartlett, 1978; Markman, 1994; Mervis & Bertrand, 1994; though see Horst & Samuelson, 2008). In this situation, children are confronted with several objects for which they have names, and one for which they do not. After hearing a novel word, children typically select the unnamed object as a referent, using their knowledge of which objects already had names (Halberda, 2006). This offers an explanation for acceleration: As more names are known, the likelihood of being in the appropriate situation increases. In addition, fast-mapping is observed both before and after the explosion (Markman, Wasow, & Hanson, 2003), suggesting that acceleration due to fast-mapping arises out of leverage, not its sudden onset.

Leveraged learning is not restricted to mutual exclusivity; the process of segmenting the stream of speech into words is another candidate. Peters (1983) as well as Brent and Cartwright (1996) posit that children segment words from running speech using known words (or sequences of words) as access points. For example, in the phrase *cognitive wugness,* knowledge of the word *cognitive* can be used to segment and identify the word, *wugness*. In this way, each time a word is acquired this can be leveraged to support the acquisition of future words. This is very clearly shown by Yang's (2004) computational work on word segmentation. He compares a segmentation model in which words are learned independently of each other (using only transition probabilities and a simplified prosody) to one that also incorporates this sort of leverage. The leveraged model shows an increase in performance of 10–20% on a real corpus of English.

Empirically, there is substantial support for leverage in segmentation. Adults can use newly learned words as a basis for segmentation in an artificial grammar (Dahan & Brent, 1999). Moreover, Bortfeld, Morgan, Golinkoff, and Rathbun (2005) demonstrated that 6-month-old infants can use their own name as well as the word *mommy* in segmentation. Finally, Plunkett (1993) links the vocabulary growth functions of three children to their abilities to segment speech.

Finally, leveraged learning can also come from syntax. Formal approaches such as syntactic bootstrapping (Gleitman & Gleitman, 1992) and statistical approaches like ''frequent-frames'' (Mintz, 2003) posit that the distributional properties between words can be used to extract syntactic categories like nouns and verbs. For example, the class of objects that appear after *the* and *an* are typically nouns, while those that appear after *are*, *has*, and *is* are typically verbs (see also Christophe, Millotte, Bernal, & Lidz, 2008). While specifically syntactic processes like these are often thought to occur well after the vocabulary spurt (although work like Christophe et al., 2008 and Mintz, 2005 suggests that it can be at least close), these represent leveraged learning mechanism that may account for longer term acceleration.

Leveraged learning provides a compelling account of acceleration in word learning. However, there is little proof that it is causal. The bulk of evidence has attempted to tie the

onset of leveraged learning to the timing of the vocabulary spurt. For example, Plunkett's (1993) longitudinal study suggests that the onset of correct segmentation coincides with the vocabulary explosion. There have also been attempts to tie the onset of fast-mapping to the vocabulary explosion (Mervis & Bertrand, 1994), but this has not held up (Woodward, Markman, & Fitzsimmons, 1994). However, correlated timing alone cannot demonstrate causality. More importantly, this sort of analysis ignores the strength of leveraged learning. In leveraged learning, acceleration derives not from the onset of the mechanism but rather from the gradual accumulation of the knowledge that provides leverage.

Finally, recent work suggests that fast-mapping may not even be sufficient to promote short-term retention. Horst and Samuelson (2008) found that objects correctly named in fast-mapping situations do not appear to be retained even 5 min later. This suggests that leverage exerted by mutual exclusivity may be exerted primarily at the timescale of online referent selection and may provide little help for long-term acquisition. It also raises the question of retention for other forms of leverage. Thus, while leveraged learning mechanisms provide a compelling framework in which to think about acceleration in learning, it is not clear that they actually account for it.

1.4. Acceleration without specialized mechanisms

Recently, McMurray (2007) presents an alternative account of acceleration in word learning that requires no recourse to leveraged learning. This model points out that word learning is fundamentally a *parallel* process. Words accumulate knowledge at the same time. Thus, during the ''slow phase'' of word learning prior, children are doing significant learning (on words they have not acquired yet), and a spurt that happens at 18 months is partially the product of this knowledge that accumulated long before (see Bloom, 2004). Once we consider this, then the relative difficulty of acquiring words become a crucial factor in predicting the rate of acceleration.

To illustrate this, McMurray developed a simple model in which word learning was modeled as the most linear, least accelerative process imaginable. If acceleration could still be seen in such a simple system, then leveraged learning (or other specializations) may not be necessary to account for the vocabulary explosion.

In this model, learning is simulated as the accumulation of points. At each time step, each word accumulates a single point (and the rate of accumulation is fixed over time). Each word has a threshold for learning (its time-to-acquisition or TTA)—when a word's point total exceeds its TTA, it is learned. In this system, the distribution of thresholds predicts the time course of acquisition. If there are very few easy words (low thresholds), and many moderately difficult words, the model will go through a period of slow learning (when it gradually acquires the easy ones), followed by a sudden acceleration as the model suddenly reaches the threshold for many more words. In this model, the number of words known at any point in time, t, is simply the integral of the distribution of TTA from 0 to that time. Thus, as long as the number of words at a given difficulty level increases as a function of difficulty, acceleration will always be observed—there is no need for any explicitly accelerative mechanisms.

Thus, McMurray (2007) concludes that as long as two conditions hold, word learning will always appear to accelerate. First, learning should be somewhat gradual and words should be learned in parallel. Second, the distribution of word difficulty should vary such that there are fewer easy words than hard words. If these hold, learning will always accelerate.

This would seem to rule out the need for any leveraged learning to explain acceleration. To examine this, McMurray (2007) also models leveraged learning in this framework. Here, as each word is acquired, a point (or partial point) is added to the unlearned words. This model also demonstrates acceleration. Critically, though, when each learned word imposes a cost (the inverse of leveraged learning), acceleration is also seen. This suggests that leveraged learning mechanisms may have little to do with the underlying form of the function. However, little analysis was presented with this model, and it did not attempt to work out the mathematical relationship between leveraged learning and the form of the growth function.

1.5. Overview

Given the aforementioned evidence that children utilize leveraged learning, it is important to understand how such processes can impact the form of vocabulary growth. The purpose of the present paper is to address this explicitly. We start from the McMurray (2007) model of vocabulary acquisition and derive a general, continuous-time form of it. We then implement leveraged learning and present a complete analysis of this model demonstrating that leveraged learning cannot create acceleration when there is none.

2. The model

2.1. Vocabulary growth without leveraged learning

The McMurray (2007) model was a discrete-time model in which time was treated as a discrete series of steps. However, the analysis of our leveraged model is simplified by considering the analogous continuous-time model. Thus, we first formulate the continuous version of this model (in which no leveraged learning occurs).

In this model, $L(t)$ represents the number of words learned up to time t. Time is continuous, as are the number of points required to learn a word. Thus, the distribution of difficulty is continuous (as a word's difficulty is not limited to the set of integers). We will let a represent the accumulated knowledge thus far (in the discrete version it is the number of points accrued). At any given time $a(t)$ will be the same across all words, as all words acquire knowledge at the same rate. The distribution of word difficulty is defined by the function $g(a)$, which defines the number of words that can be acquired for any value of a.

Fig. 2 shows an example of one such $g(a)$, a Gaussian distribution with a mean of 4,000 and a standard deviation of 1,400. In this example, most of the words are acquired around $a = 4{,}000$, and a smaller number of words can be acquired at smaller or larger values. In addition to describing the distribution of easy and hard words, any specific value of $g(a)$

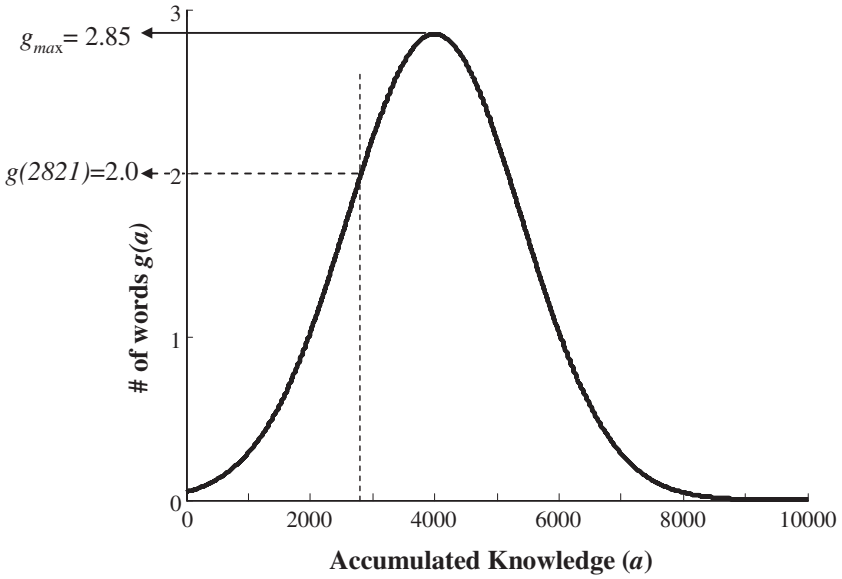

Fig. 2. An example of $g(a)$. Here, $g(a)$ is a Gaussian distribution of TTA with a mean of 4,000 and a standard deviation of 1,400. At 2,821 points of accumulated knowledge (dashed line), approximately two new words will be acquired. Also shown is $g_{\max}$, the maximum number of words that can be acquired at any point in accumulated knowledge. This occurs at a = 4,000 (the mean).

gives the number of *new* words that are learned for a small change in accumulated knowledge, a. Thus, in this example, there are about two *additional* words that can be acquired between a = 2,821 and 2,822. Note that by this definition, $g(a)$ is equal to $\mathrm{d}L/\mathrm{d}a$, the change in the number of words known for some change in a.

We can now formulate the problem as an ordinary differential equation. If there is no cost or benefit associated with previously learned words, then $\mathrm{d}a/\mathrm{d}t = 1$. That is, for each unit of time, the model gains one unit of accumulated knowledge. So, in general, the amount of accumulated knowledge at any given time t is simply $a(t) = t$. Thus, we can replace a with t in the relation $\mathrm{d}L/\mathrm{d}a = g(a)$ to give the differential equation for L:

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \frac{\mathrm{d}L}{\mathrm{d}a} = g(a) = g(t) \tag{1}$$

Therefore,the number of words learned up to time t is the area under the difficulty density $g(a)$ from zero to t. That is,

$$L(t) = \int_0^t g(s)\mathrm{d}s \tag{2}$$

(where s represents time, integrated from 0 to t, the current point in time) which we call $G(t)$. So $L(t)$ will be concave up wherever $g(a)$ is increasing. Thus, any distribution of difficulty with fewer easy words than moderately difficult words will show an initial phase of acceleration. This is the conclusion reached by McMurray (2007).

2.2. Vocabulary growth with leveraged learning

Now, we consider the case where there is a benefit (or cost) associated with learned words. That is, each time a word is learned, the accumulated knowledge of remaining words is adjusted slightly by adding the cost/benefit factor c. If c is positive, each time a word is learned, the accumulated knowledge of the remaining words increases to provide leveraged learning. If c is negative, this simulates interference, where each word slows down the acquisition of future words.

For each unit of time, a increases by one unit (as before), plus the benefit offered by each new word (c) multiplied by the number of new words learned at that time step. That is,

$$\frac{da}{dt} = 1 + c\frac{dL}{dt} \tag{3}$$

Integration of this function over time allows us to compute the accumulated knowledge at any given time, as the number of points for a given word is the number of time steps plus the benefit offered by all previously learned words.

$$a(t) = t + cL(t) \tag{4}$$

Given this function, we can now derive the differential equation describing the growth of L. We first use the chain rule to show that

$$\frac{dL}{dt} = \frac{dL}{da}\frac{da}{dt} = g(a)\frac{da}{dt} \tag{5}$$

Next, writing everything in terms of L and t (using Eq. 4 to define a and Eq. 3 to define da/dt) gives us:

$$\frac{dL}{dt} = g(t + cL(t))\left(1 + c\frac{dL}{dt}\right) \tag{6}$$

Note that this function includes the derivative, dL/dt, on both sides of the equation as the change in a is itself a function of the change in L (the number of new words acquired). Solving the equation for dL/dt we get

$$\frac{dL}{dt} = \frac{g(t + cL(t))}{1 - cg(t + cL(t))} \tag{7}$$

While this equation looks complicated, it is in fact an exact ODE and can therefore be solved. However, the solution is only given implicitly.

$$L(t) = G(t + cL) \tag{8}$$

Here, $t + cL$ is a (accumulated points) and $G(a)$ is the integral of the difficulty distribution $g(a)$. This solution makes intuitive sense. As in the no-cost/benefit case, $L = G(a)$, L is the integral of the distribution of TTA. While in the initial model we integrated from 0 to t, in this model, L is integrated from 0 to $t + cL$.

The fact that now the range of integration (for computing L) depends on L creates considerable complexity. Thus, we cannot solve explicitly for L—given t, we cannot compute the equation for the number of words known at that time (although we can simulate it iteratively). However, we can solve the inverse, solving for t as a function of L. That is, given lexicon size, we can know the time at which it is reached. This is shown in Eq. 9.

$$t = G^{-1}(L) - cL \tag{9}$$

This means that the time at which any given value of L is achieved is the original (no leveraging) time, $G^{-1}(L)$, minus a linear term, cL. Graphically, this means that a benefit will shift the graph left by an amount proportional to the number of words L. This is illustrated in Fig. 3.

Using this function, we can break down the relationship between c (the degree of leverage) and the form of learning. We do this in three cases: (a) when c is large, (b) when c is small, and (c) when c is negative (interference).

2.2.1. When leveraging is large

Eq. 7 shows the change in the number of words as a function of time given a leveraging benefit c (and the current lexicon size, L). We know (by the implicit function theorem) that this function, $L(t)$, is a function of t only as long as its derivative is well defined. That is, as long as the denominator in Eq. 7 $(1 - cg(t + cL))$ is non-zero. Thus, the solution will be well defined by Eq. 9 as long as the difficulty distribution, $g(a)$, stays below $1/c$. We

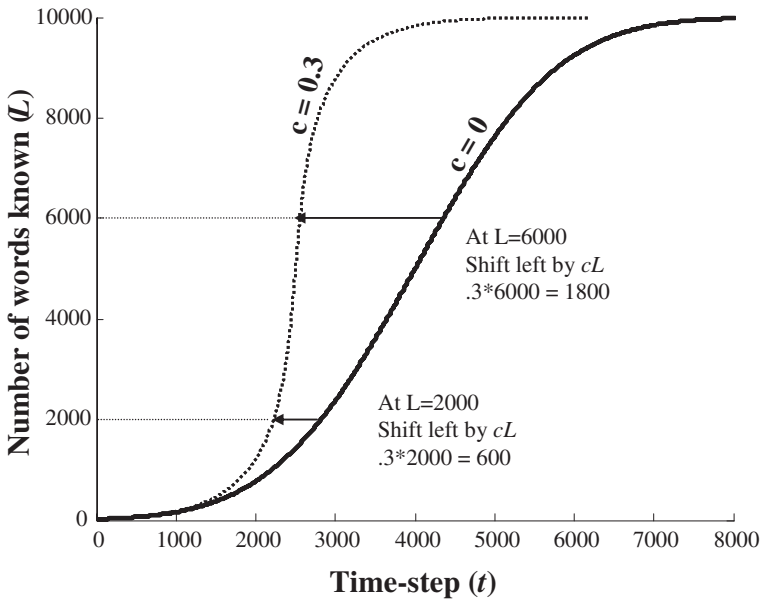

Fig. 3. Effect of a positive leverage term (c) on the rate of acquisition. At any given time, the function is shifted leftward by c multiplied by the number of words known at that point in time (L).

therefore require that c be less than one over the maximum value of g, which we term g_{max}, and which represents the number of words at the peak of the difficulty distribution (see Fig. 2). Thus, the solution is only well defined for small c and too much benefit will make the model unrealistic. For example, the distribution of difficulties used above was a Gaussian distribution of 10,000 words, with a mean difficulty 4,000, standard deviation 1,400. This has maximum of about $g_{max} = 2.85$ words (at its mean of 4,000). Therefore, any benefit, c, of less than approximately 0.35 will satisfy this condition.

If c exceeds this critical amount, $c > 1/g_{max}$, the solution $L(t)$ will be pushed left to such a great degree that it has an infinite slope. This means that the benefit from the words learned so far has given the child so much knowledge that he or she automatically knows all words. Fig. 4, illustrates this case. This figure displays Eq. 9 computed for a model with the above Gaussian distribution of TTA, when $c = 0.5$. After about 2,000 words, the function bends back on itself (achieving an infinite slope at a time of around 1,800). When this same simulation is computed iteratively, it appears that learning gradually accelerates from $t = 0$ at about time 1,800, at which point the entire remaining lexicon (8,000 words) is learned in one time step.

Thus, a small value of c is required. In the simulations above, each word takes an average of 4,000 time steps for learning, yet the benefit of learning a word must be less than one-third of a point! This suggests that in order to maintain realistic growth functions, whatever benefit is offered by leveraged learning must (a) be very small, (b) operate over a limited number of words, or (c) be counteracted by some cost of learning new words.

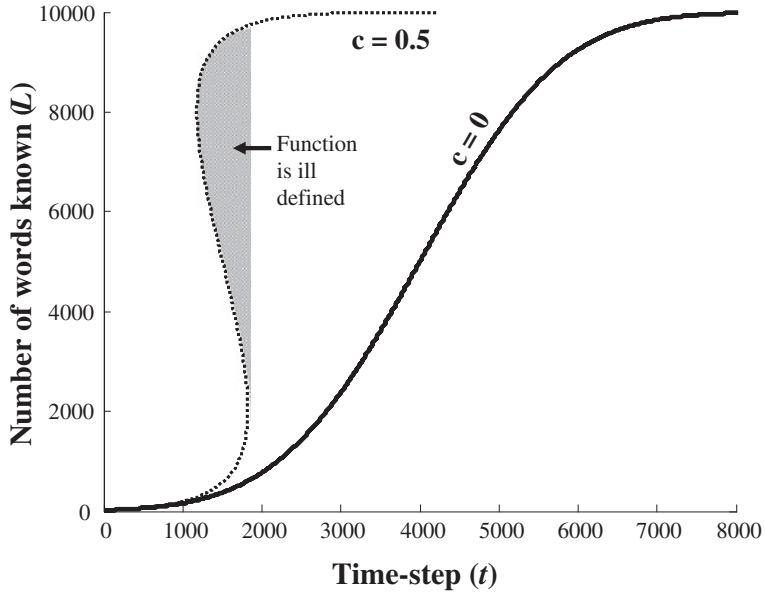

Fig. 4. Effect of large positive values of c (leveraging) on acquisition. Here, at $c = 0.05$, the function becomes ill defined sometime after the system has acquired 2,000 words. Note that this graph was generated by computing for each L, the t at which this many words would be acquired. Running time forward (iteratively) yields a function for $L(t)$ that jumps directly from 2,000 to 10,000 words known at around $t = 1{,}800$.

2.2.2. When leveraging is small

Given small benefits from leveraged learning (e.g., when $0 < c < 1/g_{\max}$), we can now ask what role leveraged learning has in shaping the overall learning function. If the first derivative of $L(t)$ describes the velocity of learning at any given time (Eq. 7), the second derivative allows us to compute its acceleration.

$$\frac{\mathrm{d}^2 L}{\mathrm{d}t^2} = \frac{g'(t + cL(t))}{(1 - cg(t + cL))^3} \tag{10}$$

If we have chosen c less than the critical amount, then the denominator is always positive. Therefore $L(t)$ will be concave up whenever g' is positive, that is, whenever the distribution of difficulty is increasing. This is exactly as in the no-benefit case. Again, a period of acceleration will be observed in this model if and only if there are fewer easy words than moderately difficult words. Thus, a benefit can increase the acceleration in learning, but it *cannot create acceleration where none existed before* (Fig. 5).

This formula for the second derivative can also be used to compute the inflection point of the growth function. Note that the literature on word learning typically defines the inflection point as the onset of acceleration. By contrast, we define the inflection point, mathematically, as the point at which acceleration switches to deceleration (as acceleration always occurs, it makes no sense to talk about its onset). Given Eq. 10, then, the inflection point occurs whenever the numerator is 0, that is, when $g'(t + cL) = 0$. As g describes the distribution of easy and hard words, the inflection will be at a maxima or minima (critical point) in the density function, g. For the Gaussian density used above, there is only the one critical point for g, its maximum (at a = 4,000, at which point L = 5,000—half the words are learned). Solving for t gives $t = 4{,}000 - 5{,}000c$. Thus, for this density of difficulties, the inflection point will always be at the time where half the words are learned. If there is a benefit ($c > 0$), this point will occur at an earlier time.

This illustrates that changes in c can alter the dynamics of learning, moving up the onset and the inflection point. Thus, this model is compatible with results like those of Mervis and Bertrand (1994) or Plunkett (1993), which suggest that variation in a leveraged learning (mutual exclusivity or segmentation in these cases) across children correlated with the onset of the vocabulary explosion across children. Accordingly, variation in c between children causes changes in the inflection point, although it is not causally related to acceleration itself.

2.2.3. Interference

The last interesting case is when $c < 0$. This represents a situation in which learning one word imposes a cost on future learning. Research has typically focused on cases where learning a word benefits future learning, motivated by the need to explain both the vocabulary explosion, as well as the ambiguity faced by the child in everyday naming situations (e.g., Quine, 1960). However, there are situations in which learning a word seems

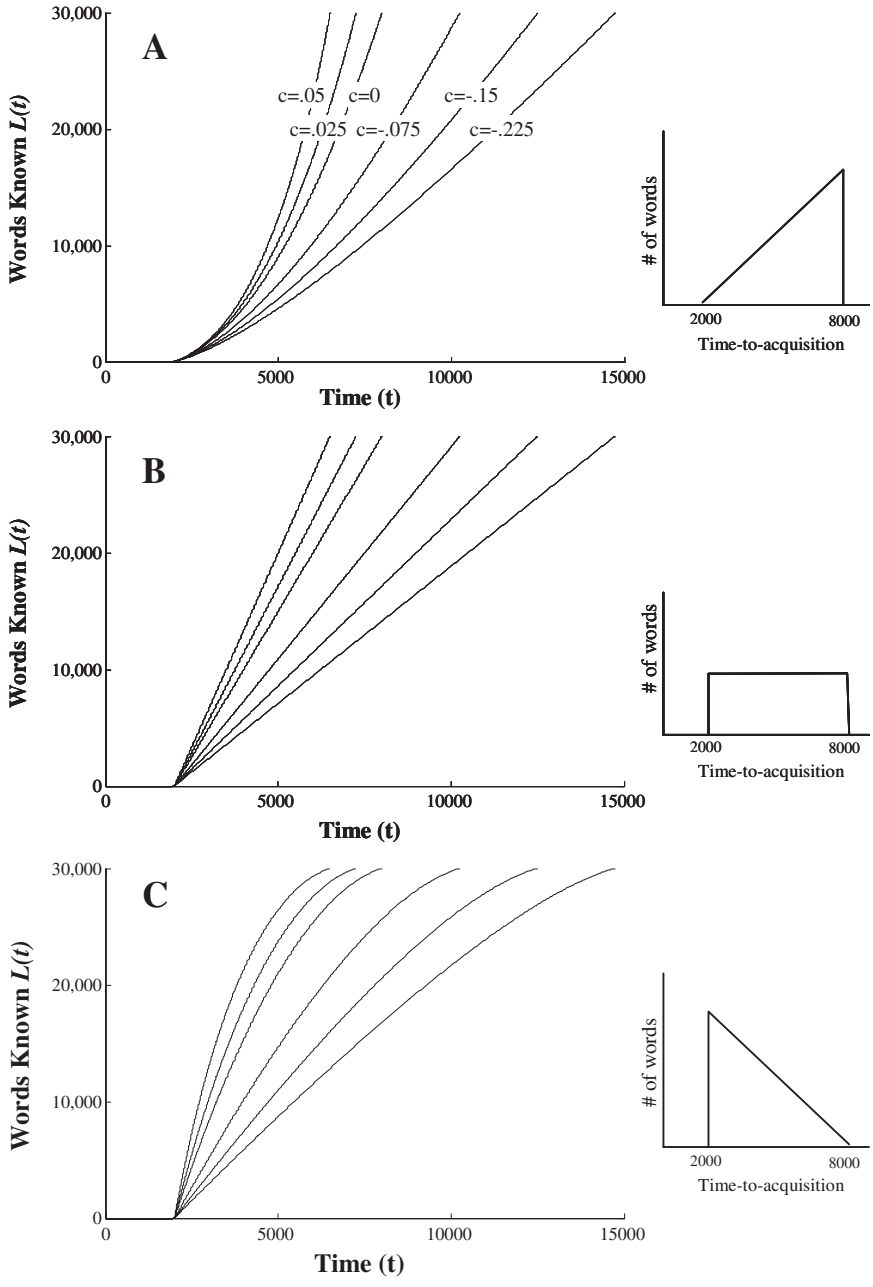

Fig. 5. Words known as a function of time and leveraging, c, for three different distributions of difficulty, $g(a)$. Lower panels show schematics of $g(a)$. Left panels show results when $g(a)$ is decreasing, a violation of McMurray (2007), which does not show acceleration. Center panels show a uniform distribution of $g(a)$, which is also a violation and shows no acceleration. In both of these cases, positive values of c do not create acceleration—they simply shift the curves to the left. The right panels shows results with an increasing $g(a)$. In this case, acceleration derives from the increasing $g(a)$, and c merely hastens or forestalls it.

to interfere with future learning. For example, Swingley and Aslin (2007) have demonstrated that 1.5-year-olds have difficulty learning words that sound similar to words they already known (e.g., *tog*) (see also Fennell & Werker, 2003; Stager & Werker, 1997; but see Rost & McMurray, 2009).

This is straightforward to model in this framework. If c is negative, each time a word is learned, it inflicts a cost on the remaining words. In this case, the solution for $L(t)$ will be pushed right, causing slower learning. However, as in the case where c is small, this can only affect the steepness and will show acceleration whenever the no-leverage model shows acceleration. Interestingly, there is no lower bound on c.

2.3. What is the nature of the difficulty distribution?

One question that is raised by the present study as well as by McMurray (2007) is, What is the nature of word difficulty? The distribution of easy and hard words is important in predicting the presence of acceleration and the shape of the growth function (even when learning is leveraged). However, despite decades of research, the notion of what makes a word easy or hard to learn is relatively poorly defined.

The prior analyses intentionally did not define the function, $g(a)$, specifying the difficulty distribution. Any $g(a)$ that shows acceleration will still show it for all values of leverage (c); any $g(a)$ that does not will not. McMurray (2007) assumed that $g(a)$ was Gaussian, and for good reason: If we assume that a word's difficulty is the sum of many factors that are reasonably independent, then the distribution of difficulty will be Gaussian according to the central limit theorem. There are longstanding findings that syntactic category (Bornstein et al., 2004), phonology (Swingley & Aslin, 2007), and frequency (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991), particularly in isolation (Brent & Siskind, 2003) all play a role in predicting when a word is learned. Given that these are largely independent, this seems a reasonable assumption.

While it is currently not possible to combine all of these factors to estimate difficulty of a single word, our analysis does not rest on the difficulty of any single word. Rather, we must be able to characterize the distribution, $g(a)$. In this regard, word frequency presents a particularly interesting case because its distribution is well defined across many languages by Zipf's (1935) law. Zipf's law simply states that the frequency of a word will be proportional to the inverse of its rank (e.g., the most frequent word has rank 1, the second most frequent has rank 2, etc.). This relationship can be modified with an exponent such that

$$\mathrm{Freq}(i) = \frac{f_1}{k_i^s} \tag{11}$$

Here, k_i represents the rank order of word i's frequency, s is a free parameter, and f_1 is the frequency of the most frequent word.

Zipf's law gives rise to a hyperbolic function, similar to Fig. 6. When converted to a statistical distribution (the number of words at any given frequency), it yields the function

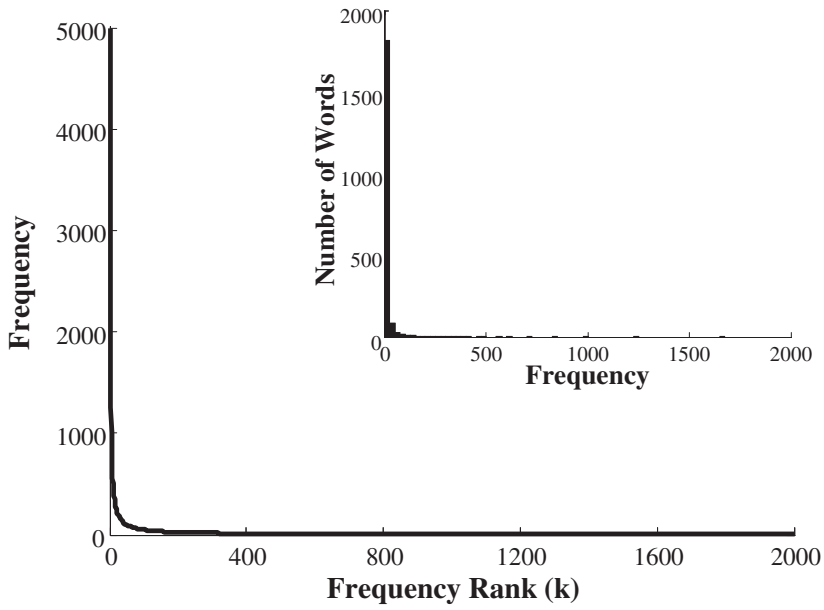

Fig. 6. An example of a Zipf relationship between frequency and rank (shown are the first 2,000 of 10,000 words). F_1 is 5,000 and $s = 1$. Inset: A partial histogram of the number of words at each frequency value.

shown in the inset. Here, most words have a very low frequency (very difficult) and there are a few high frequency (easy) words.

Theoretically, this distribution fits the criteria laid out by McMurray (2007). However, frequency is the inverse of difficulty—that is, the words with high frequencies should have low difficulties. This requires us to define the relationship between frequency and difficulty before it can be applied to the foregoing analyses. There are two ways to do this. The most obvious way is additive. Simply multiplying the Zipfian frequencies by a negative number to reverse the directionality, and adding a constant to make everything positive. In a sense, this switches the direction of the *x*-axis in Fig. 6 inset. However, recent analyses by Mitchell and McMurray (2008) suggest that frequency can be rescaled multiplicatively, where difficulty is proportional to $1/\text{frequency}$. This latter approach has the nice property that a word with twice the frequency will take half as long to learn.

Both have consequences for leveraged learning and will be discussed separately. For compatibility with the prior analyses we use the continuous version of Zipf's law, in which the word number, k, is continuous.

2.3.1. Difficulty $\propto$ frequency

In the first case, difficulty is scaled additively from frequency. Fig. 7A shows computations of $L(t)$ using Eq. 9 to integrate a Zipfian distribution of frequencies in which $s = 1$, $f_1 = 5{,}000$, and there were 10,000 words. The frequencies were then converted to difficulties by multiplying them by -1 and adding f_1 to make them positive.

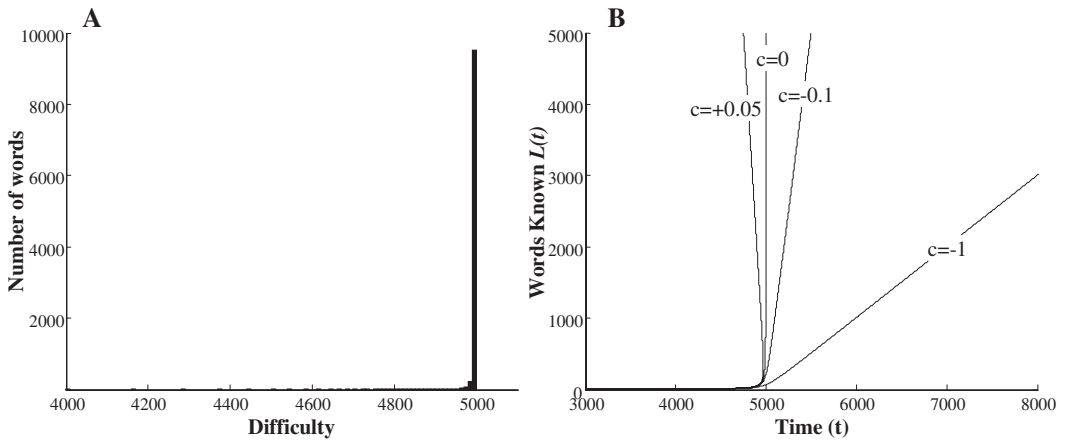

Fig. 7. Results from simulations with Zipfian distribution when frequency is converted to difficulty additively. (A) A histogram of difficulties. Most words take 5,000 time steps to learn and a handful can be learned earlier. (B) Growth functions for various amounts of leverage or interference. All curves show acceleration around 5,000, and even with no leverage learning is very steep. Moreover, at even small positive values of *c*, the function is poorly defined (has a negative slope), suggesting that leverage may not offer much benefit if difficulty is Zipfian.

$$\text{Difficulty}_i = f_1 - \frac{k_i^s}{f_1} \tag{13}$$

This predictably yields a positively increasing $g(a)$ (Fig. 7A). Integrating this to achieve the number of words known, $L(t)$, shows a very steep acceleration (Fig. 7B), and positive leveraging offers very little benefit. Moreover, in accord with our analysis, this same shape is observed with negative values of *c* (interference). This, if difficulty is distributed as a Zipfian distribution, the standard model holds and a very extreme acceleration will be seen.

2.3.2. Difficulty ∝ 1/frequency

The alternative approach to scaling Zipfian frequencies is to treat difficulty as proportional to the inverse of frequency:

$$\text{Difficulty}_i \propto \frac{k_i^s}{f_1} \tag{14}$$

This case gives rise to a couple of interesting properties.

First, if $s = 1$, then difficulty is proportional to each word's frequency rank. Ranks are equally spaced (that is the difference between rank 1 and rank 2 is the same as between rank 100 and rank 101), and there will always be one word with rank 1, one with rank 2 (and so on). As a result of these facts about ranks, the resulting distribution of difficulty, $g(a)$, will be uniform—*there are an equal number of words at each difficulty level*. Integrating this difficulty distribution leads to perfectly linear growth (no acceleration) and incorporating positive or negative leveraging can only change its slope (Fig. 8).

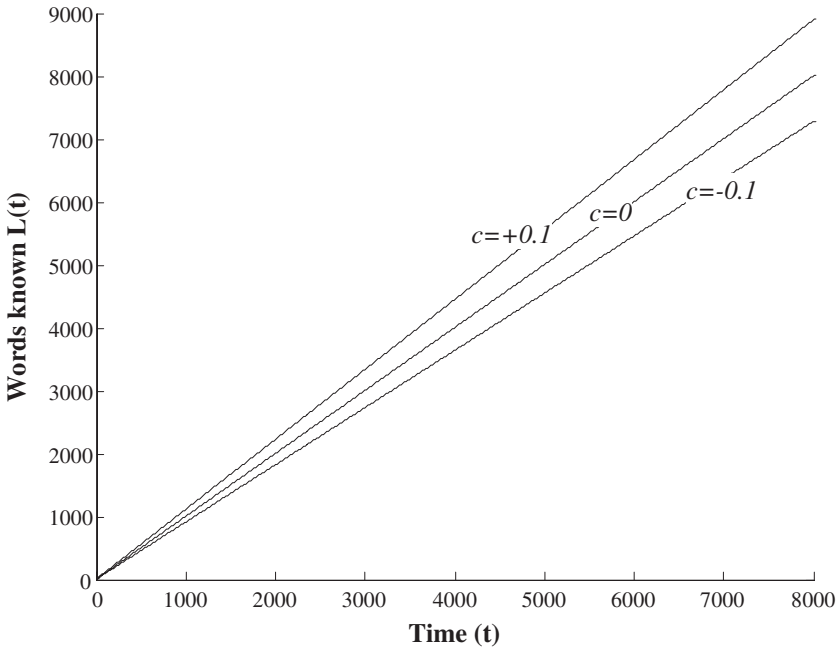

Fig. 8. Vocabulary growth as a function of time when difficulty is the reciprocal of Zipfian frequency. Here, with $s = 1$, learning is perfectly linear and leverage only affects the slope.

Interestingly, when $s > 1$, the distribution of difficulty, $g(a)$, shows slightly more easy words than hard words, and as a result the growth functions are decelerating (concave down).[1] Conversely, when $s < 1$, there are more hard words than easy words, and the growth functions show acceleration. However, in both cases, our leveraging analysis holds: positive values of c do not create acceleration when $s > 1$, and negative values for c do not eliminate it.

2.3.3. Implications

With difficulty as a concept still relatively undefined, it is hard to be certain of its precise distribution across words, $g(a)$, a factor that we have shown is important in vocabulary growth. It is attractive to gravitate toward one component of difficulty, word frequency, for which the distribution is well defined by Zipf's law. At the broadest level, this can be done quite simply in the context of the model, and it largely supports our findings: When there are fewer easy than hard words acceleration is guaranteed, and leveraged learning does not impact this.

However, in converting from frequency to difficulty, some subtleties arise. First, this can be done additively or multiplicatively. In the additive case, acceleration is observed because of the steeply increasing difficulty distribution. However, this may be too steep—virtually the entire lexicon is learned at once. In the multiplicative case there were some values of s for which acceleration will not be observed, and leveraging could not

create it. These limitations suggest that frequency alone may be insufficient to define difficulty. This, of course, must be true. According to Kucera and Francis' estimates, the 10 most frequent words in English (the, be, of, and, a, in, he, to, have, it) are all closed class words that are learned quite late. If difficulty were only a matter of frequency, this would not be true.

While the Zipfian distribution is an important starting point to determining $g(a)$, it will ultimately depend on how frequency is treated (additively or multiplicatively) and what the value of s is for a given language. However, beyond this particular model, it is important to remember that frequency is not a property of the word (like phonology or syntactic class); it is an estimate of how often it occurs (stochastically) in the child's environment. Thus, our model may be limited in its ability to handle frequency, and a stochastic model may be a better approach for dealing with it (e.g., Mitchell & McMurray, 2008).

3. Discussion

The foregoing analyses suggest a complex relationship between leveraged learning and acceleration. Leveraged learning cannot create acceleration where there is none—it is not causally related to the vocabulary explosion. It can affect the timing and shape of the acceleration, but only if the overall amount of leverage is small. Given the behavioral evidence that children do engage in many such processes, this would seem to suggest that leverage is a component of learning, but not fundamental to acceleration. Moreover, it cannot be large: either the amount of leverage provided by a word must be small, the scope of leverage (number of words affected) must be narrow, or there must be costs that counteract it. This does not minimize its importance: There are clearly learning situations in which the input is ambiguous enough, or the underlying structure complex enough, that leveraged learning is the only way to acquire the relevant information. However, leveraged learning does not explain acceleration.

Recent behavioral and computational work on mutual exclusivity suggests a framework for understanding the small benefit that leveraged learning must offer. Horst and Samuelson (2008) demonstrated that even though children can use mutual exclusivity to determine a novel name/referent mapping, they do not retain these mappings 5 min later when tested in a less supportive context. Mitchell and McMurray (2008) (see also Horst, McMurray, & Samuelson, 2006) modeled this in a dynamic neural network in which fast-mapping was simulated as an online competition process, while learning accumulated via slow associative mechanisms. This network was capable of fast-mapping and also failed to show retention when tested again. However, the model did acquire a small amount of knowledge on each fast-mapping trial, and over the course of thousands of such trials eventually acquired the mappings. This implies that leverage may exist on two timescales: the fast timescale of online referent selection (in which leverage could be large and useful) and a slower timescale of learning (the case discussed here), in which leverage is necessarily small.

One limitation of the model is the fact that difficulty, and hence the difficulty distribution, is still relatively undefined. Zipf's law presents an attractive possibility for modeling the difficulty distribution as a function of frequency. However, our analyses suggest that it is not sufficient (which may explain the relatively low percentage of variance that raw frequency accounts for in age of acquisition: Brent & Siskind, 2003). Factors like phonology, syntax, and the importance of the word in the child's own developmental ecology are more difficult to measure but ultimately may be more predictive.

A second limitation of the present work is our use of the deterministic model. In this model, easy words are always learned before hard words and there is little room for randomness. However, current work (Mitchell & McMurray, 2008) has generalized the McMurray (2007) model to a stochastic process and shows similar findings. In this model, on each time step, words obtain points with some probability. The frequency (probability of obtaining a point) and difficulty (number of points required) both contribute to the course of learning, and the time course of the model is not predetermined by the difficulty of the words. Crucially in this model, we have proved that acceleration is guaranteed as long as words require more than one point (exposure) for learning.

Analysis of this model is substantially more complex than the deterministic model; therefore, we have not yet implemented leveraged learning. However, our analysis of acceleration in the stochastic case suggests that the findings are unlikely to be different. As before, leverage offers a benefit, but, as acceleration is already guaranteed, leverage is not necessary to create it. Thus, it does not seem likely that our conclusions are restricted to the deterministic model.

Together with McMurray (2007), this suggests that acceleration is a general property that should be seen in virtually any parallel learning situation. However, this raises the question as to why so many laboratory tasks show deceleration. We offer three explanations. First, tasks that use reaction time and other continuous measures may be tapping something more akin to our variable a (the number of points), rather than $L(t)$ (the number of items). Second, they may also be partially measuring a general skill that applies across many items, rather than anything item specific (which would be captured in $L[t]$, the number of items learned). Finally, our analysis of the stochastic model (Mitchell & McMurray, 2008) suggests that increasing degrees of overall difficulty lead to increasingly ''spurty'' behavior. Language learning is clearly a task that spans many years, while most laboratory tasks span hours or days, and subjects come to the laboratory with many of the prerequisite skills (e.g., ability to flexibly use categories, push buttons, etc.).

To conclude, leveraged learning mechanisms are unnecessary to explain acceleration. The present analyses prove that it can be observed even when anti-leverage (interference) is implemented, and that leverage cannot create acceleration when the underlying function does not already display it. Moreover, while leveraged learning is empirically attested, our proof suggests that it must be small, limited in scope, or counteracted. Simply put, while they are important tools in the arsenal of the developing child, leveraged learning does not account for acceleration in vocabulary growth.

Note

1. The distribution of difficulty gives the number of new words, Δk, for some small change in difficulty, Δdifficulty. Thus, the distribution can be computed by inverting the difficulty function (e.g., Eq. 13 or 14) to give the word number, k, as a function of difficulty and then differentiating that function. Details are available from the first author.

Acknowledgments

The authors thank Charles Wang and the members of the University of Iowa Modeling Discussion group for helpful comments during the development of this work and the manuscript.

References

Anderson, J. R. (1982). Acquisition of a cognitive skill. *Psychological Review*, *89*(4), 369–406.

Bloom, L. (1973). *One word at a time: The use of single-word utterances before syntax*. The Hague, The Netherlands: Mouton.

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

Bloom, P. (2004). Myths of word learning. In G. Hall & S. Waxman (Eds.), *Weaving a lexicon* (pp. 205–256). Cambridge, MA: The MIT Press.

Bornstein, M. H., Cote, L. R., Painter, S. M. K., Park, S., Pascual, L., Pêcheux, M., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, *75*(4), 1115–1139.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, *16*, 298–304.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125.

Brent, M. R., & Siskind, J. M. (2003). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, *15*, 17–29.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech*, *51*, 61–75.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, *128*(2), 165–185.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: The MIT Press.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2/3), 245–264.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., & Thal, D. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59* (Serial no. 242).

Ganger, J., & Brent, M. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621–632.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, *98*, 3–53.

Gleitman, L. R., & Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, *1*, 31–35.

Golinkoff, R. M., & Hirsh-Pasek, K. (2006). Baby wordsmith: From associationist to social sophisticate. *Current Directions in Psychological Science*, *15*, 30–33.

Golinkoff, R. M., Mervis, C. V., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*, 125–155.

Gopnik, A., & 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.

Gopnik, A., & Meltzoff, A. N. (1992). Categorization and naming: Basic-level sorting in 18-month-olds and its relation to language. *Child Development*, *63*, 1091–1103.

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*, 310–344.

Horst, J. S., McMurray, B., & Samuelson, L. K. (2006). Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In R. Sun (Ed.), *The proceedings of the 28th meeting of the Cognitive Science Society* (pp. 339–344). Austin, TX: Cognitive Science Society.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention in 24-month-old infants. *Infancy*, *13*(2), 128–157.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248.

Logan, G. D. (1992). Shapes of reaction time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 883–914.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199–227.

Markman, E. M., Wasow, J. L., & 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., & Bertrand, J. (1994). Acquisition of the novel name/nameless category (N3C) principle. *Child Development*, *65*, 1646–1662.

Mervis, C. B., & Bertrand, J. (1995). Early lexical Acquisition and the vocabulary spurt—A response to Goldfield and Reznick. *Journal of Child Language*, *22*(2), 461–468.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Mintz, T. H. (2005). Linguistic and conceptual influences on adjective acquisition in 24- and 36-month-olds. *Developmental Psychology*, *41*, 17–29.

Mitchell, C., & McMurray, B. (2008). A stochastic model of the vocabulary explosion. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1919–1926). Austin, TX: Cognitive Science Society.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, *6*, 136–142.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Learning and Cognition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Peters, A. (1983). *The units of language acquisition*. London: Cambridge University Press.

Plunkett, K. (1993). Lexical segmentation and vocabulary growth in early language acquisition. *Journal of Child Language*, *20*, 43–60.

Quine, W. V. O. (1960). *Word and object: An inquiry into the linguistic mechanisms of objective reference*. Cambridge: MIT Press.

Rescorla, L., Mirak, J., & Singh, L. (2000). Vocabulary growth in late talkers: Lexical development from 2;0 to 3;0. *Journal of Child Language*, *27*(2), 293–311.

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

Reznick, J. S., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology*, *28*, 406–413.

Rost, G., & McMurray, B. (2009) Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Rumelhart, D., Hinton, G., & Williams, R. (1986) Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. (pp. 318–362). Cambridge, MA: The MIT Press.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382.

Stern, W. (1924). *Psychology of early childhood* (A. Barwell, Trans.). London: George Allen & Unwin (Original work published 1914).

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*, 99–132.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*, 553–566.

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456.

Zipf, G. (1935). *The Psychobiology of Language*. New York: Houghton-Mifflin.