Modeling the Development of Determiner Productivity in Children's Early Speech

Stephan Meylan
smeylan@berkeley.edu
Department of Psychology
University of California, Berkeley

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Roger Levy
rlevy@ucsd.edu
Department of Linguistics
University of California, San Diego

Abstract

The English definite and indefinite articles (also known as *determiners*) are a useful index of early morphosyntactic productivity in children's speech, and give evidence about children's representation of syntactic abstractions. Previous work (i.e. Pine & Lieven, 1997) used a measure of productivity that shows a strong sensitivity to sample size and does not account for the relationship between adult input and children's learning. In this paper, we develop a more robust metric by employing a hierarchical Bayesian model to characterize the degree of generalization implicit in observed determiner usage. By inferring parameters for a generative model over longitudinal corpora, we measure the trajectory of grammatical category abstraction. Our results are consistent with the hypothesis that child learners exhibit adult-like patterns of generalization quite early in the acquisition of determiners.

Keywords: grammatical productivity; development; syntax; morphosyntax; modeling

Introduction

How do children begin to use the rich combinatorial structure of language to express novel thoughts? Nativist accounts propose an innate specification of syntactic categories that allow the child learner to exploit regularities in language structure from birth (Valian, 1986). Constructivist theories, on the other hand, contend that abstract categorical knowledge is built up over time as the child learner generalizes from specific usages to form broader combinatorial rules (Tomasello, 2003). The indefinite determiner "a" and the definite determiner "the"—the shortest and most frequent words in the English language—are a locus of interest for both theoretical viewpoints. Because they are both frequent and obligatory, determiners are an early index of morphosyntactic[1] productivity that can be observed cross-linguistically.

A context-free grammar production rule (Figure 1) captures the intuition that a noun phrase can be created by choosing a determiner from the abstract DET category and a noun from the abstract N category. This noun phrase in turn combines with other phrase structures, like a verb phrase or preposition, to form higher-order structures. For most singular nouns, a grammatical NP can be formed using either determiner. Furthermore, hearing a novel word with one determiner suggests that use with the the other is also likely grammatical. Hearing someone introduce "a blickmoo" for the first time, you would not hesitate to request "the blickmoo" even if you had never heard that sequence of words before.

[1]In English, determiners and nouns are separate words by linguistic criteria (e.g. an adjective may come between a determiner and a noun). Many other languages use determiners that are morphologically integrated with the noun (see Kramsky, 1972 for an overview).

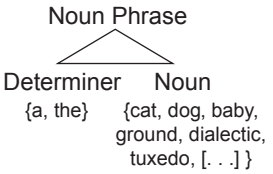

Figure 1: Many noun phrases can be created by combining a word from the abstract categories determiner and noun (NP→DET+N). It is an open question whether children's early representations are organized around these abstractions.

When do children share that same judgment? Valian (1986) showed that children between 2;0 and 2;6 demonstrate a variety of productive syntactic categories, including determiners. Using a distributional analysis of children's speech, she found that determiners were used in a fashion consistent with an adult-like grammar. Determiners were never used as the sole content of an utterance, never appeared in a sentence-final position, and were always sequenced correctly with respect to adjectives and nouns in noun phrases.

Pine & Lieven (1997) challenged Valian's assertion of adult-like grammatical productivity in children's speech by citing an apparent limit to productivity in determiner use. As a quantitative metric, Pine and Lieven presented the *overlap measure* for determiners: the number of nouns used with both determiners (in some sample), divided by the number of nouns used with either (in the same sample). For 11 children from 1;0 to 3;0 this proportion ranged from 0 to .23, which Pine and Lieven interpreted as being extremely low for a speaker with productive determiner syntax. Rather than making full use of the combinatorial productivity of nouns and determiners, on this metric children thus seemed to be very conservative in their productions and to show a strong tendency to use nouns with only a single determiner. Pine and Lieven interpreted this finding as supporting item-based theories of learning in which there is only gradual generalization from individual instances to abstractions like DET.

Valian et al. (2009) objected that Pine & Lieven (1997) failed to take noun frequency into account in considering determiner use. Because the overlap measure is necessarily 0 for all nouns that appear only once, Valian and colleagues argued that the overlap measure, especially when calculated over small datasets, under-represents productivity. Highly frequent nouns were much more likely to be used with both determiners: more than 80% of nouns used at least 6 times were used with both "a" and "the." Our own analyses of the Providence corpus (Demuth & McCullough, 2009) confirm

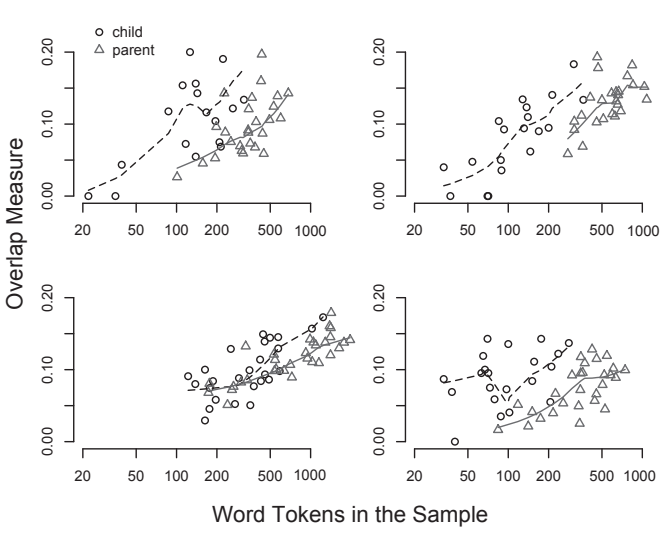

Figure 2: Determiner overlap (proportion of nouns used with both "a" and "the") increases as a function of the number of tokens in a speech sample, as seen from CHILDES files for 4 children from the Providence corpus. Dashed and solid lines show loess smoothers for the child and parent respectively.

this issue: overlap is deeply confounded with sample size. Sample size is the best predictor of both child and parent overlap, regardless of age (Figure 2).

Yang (2010) supplemented this argument by showing that—regardless of sample size—the overlap measure is necessarily low because of the Zipfian distribution of noun frequencies. The Zipfian frequency distribution of nouns results in a long tail of words seen only once, so if overlap is calculated as the proportion of nouns seen with both determiners, it will necessarily be low. Yang additionally observed that nouns vary in their determiner preference (e.g., "the bathroom" is more frequent than "a bathroom", but "a bath" is more frequent than "the bath"), unlike the simplest probabilistic instantiation of a productive context-free rule scheme as in Figure 1, where the probabilities of Determiner→"the" and Determiner→"a" would be independent of the noun's identity (Booth, 1969).

But while the overlap statistic is flawed, there is currently no replacement that directly measures the productivity of children's determiner use. Hence, in the current study, we develop a novel method for quantifying determiner productivity. We use a hierarchical Bayesian model to estimate adults' and children's determiner productivity (*metric model*) and then develop a variant that estimates the linkage between adult input and child generalization (*linking model*). In each model, one key parameter can be interpreted as a graded metric of productivity robust to variation in sample size and noun frequency distribution. Bayesian inference gives us the posterior distribution of this parameter given child and adult caregiver production data, allowing us to quantify determiner productivity and examine its developmental timecourse.

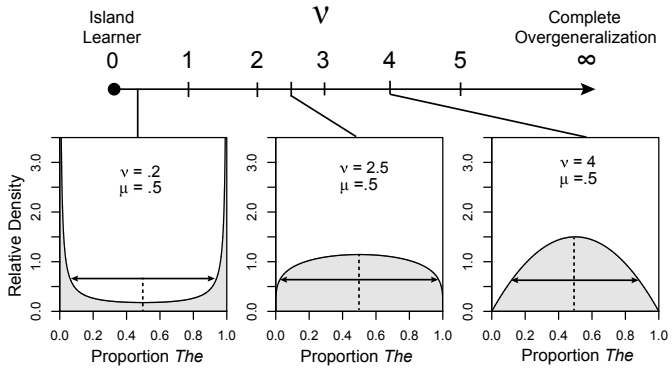

Figure 3: Interpretation of the ν parameter, a concise metric of grammatical productivity. At low values of ν, little or no information is shared between nouns. At higher ν values, nouns exhibit more consistent usage as a class, indicating the existence of a productive DET+N rule.

Metric Model

We model the use of each determiner with a noun as a draw from a binomial distribution (a single weighted coin flip). The use of "the" is heads, and the use of "a," tails. The idiosyncratic determiner preference for each noun can thus be thought of as a coin's weighting, ranging from zero (a noun used only with "a") to one (a noun used only with "the"). We model variability in noun-specific determiner preferences by assuming some distribution underlying these preferences; specifically, we assume that each noun's preference is drawn from a beta distribution with mean μ_0 (the underlying "average" preference across all nouns) and scale ν, giving us a *hierarchical beta-binomial* model (Gelman et al., 2004).[2]

The scale parameter ν in our model plays a central role in quantifying cross-noun variability and thus gives us a continuous space in which to quantify learner productivity (Figure 3). At one end of the range, when $\nu = 0$, we have an extreme "island" learner for whom every noun is produced with only one determiner or the other. At the other end of the spectrum, as ν approaches infinity, we have an extreme over-generalizer who has identical determiner preference for all nouns. The ν parameter thus establishes a continuum on which we can place constructivist and nativist hypotheses.

By estimating values of μ and ν for individual children over the course of their development, we can examine how these parameters change, potentially reflecting developmental changes in productivity. Here we use the metric model to compare mother and child productivity for the six children in the Providence corpus (Demuth & McCullough, 2009).

Model Details

A full graphical model representation of the linking model is shown on the left side of Figure 4. We assume that data d

[2]Many readers may be more familiar with the more common parameterization of the beta distribution in terms of shape parameters $\alpha = \mu\nu$ and $\beta = (1-\mu)\nu$.

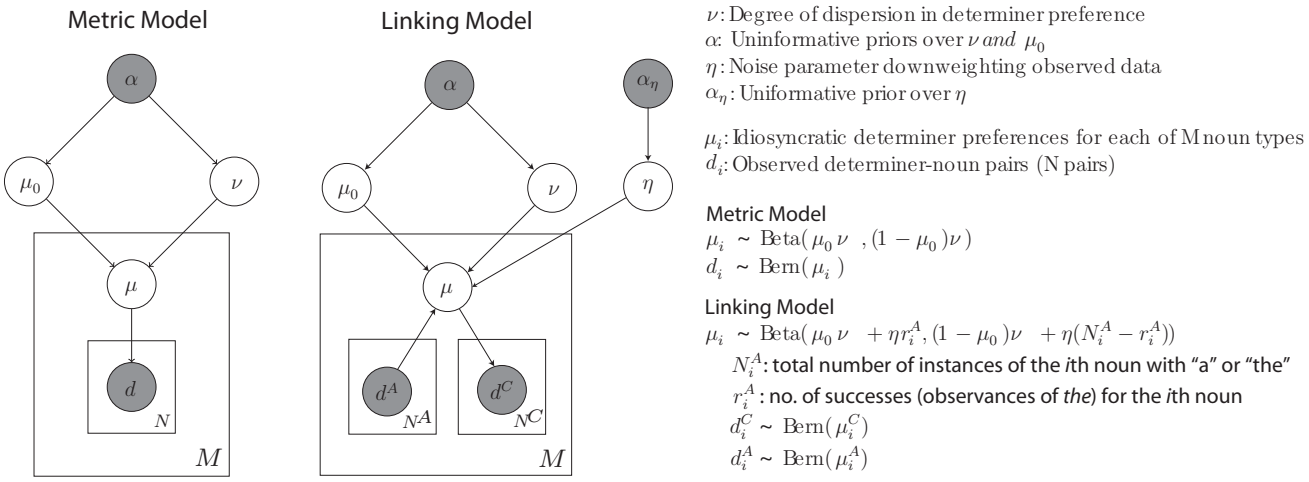

Figure 4: Graphical representations of the metric and linking model. Shaded nodes indicate observed data (determiner-noun productions) or uninformative priors set by the researcher.

(individual determiner observations) are generated as draws from a binomial with parameter μ_i for each of M noun types. These parameters are in turn drawn from a beta distribution with parameters μ_0 and ν. The μ_0 parameter describes the overall mean determiner preference, and the ν parameter—the central target of inference—describes the degree to which individual noun preferences vary around the overall average μ_0. We complete the model via an uninformative prior distribution over μ_0 and ν.

Given a sample of determiner-noun pairings, we can use Bayesian inference to produce full posteriors over μ and ν. In practice, we perform inference using Gibbs sampling via the JAGS package (Plummer, 2003); grid-sampling of posterior distributions and trace plots confirmed good convergence properties (see also the Appendix).

Corpus Selection and Extraction

The Providence corpus (Demuth et al., 2006) consists of longitudinal in-home recordings from six children from New England and contains a relatively high density sample from the onset of single words at about 1;3 to 3;0. Utterances from each child and their mother were extracted from CHILDES-formatted transcripts (MacWhinney, 2000) and augmented with a machine-generated syntax tier in CLAN (Sagae et al., 2010). Using these syntactic trees, we automatically extracted modifiers associated with each noun, as well as their part of speech. For the model input, noun uses were subset to those with a definite or indefinite determiner, yielding $5-15\times10^3$ age-referenced DET+N tokens for the mothers and $1.5-5\times10^3$ for the children.

For each mother and child, we performed a sliding-window analysis, examining successively older subsections of the corpus. On the basis of artificial corpus simulations (see Appendix), a window size of 1024 tokens was selected. On a linear sequence of tokens, each new window contained 10 new tokens from the full dataset and omitted the earliest 10. This method yielded on average 150 measures of determiner productivity for each speaker. Additionally, an overlap measure was calculated for each 1024 token window according to the procedure described in Pine & Lieven (1997).

Results and Discussion

An item-based learning theory predicts a developmental increase in children's generalization across nouns (as measured by ν) as individual item-based constructions give way to a general production rule. In contrast, a theory positing full morphosyntactic productivity predicts no major difference in generalization over development; instead, children and parents will show the same level of productivity from early on.

Our sliding window analysis reveals no clear developmental trend in children's productivity (Figure 5), consistent with the early productivity account. For both the adult and the child, individual conversational bouts show high variance, but ν values for the children are as higher or higher than those in the speech of their mothers, and children exhibit adult-like peaks of noun groupedness from the beginning of production. Nevertheless, for several of the children (e.g. William, Ethan, Violet), it is clear that the amount of data is not sufficient to allow the temporal granularity for a strong test.

Although it gives similar results to the overlap statistic, the ν parameter in our model is preferable. While the overlap measure is confounded by sample size (see above), additional data only improves our estimate of ν. Posterior inference gives an explicit representation of the model's uncertainty in a data set, making it readily apparent when the sample size is too small to estimate model parameters.

This property of the model allows us to note that the variability in the estimates of productivity for adults and children seem to be quite reliable. In both cases, there is substantial variability that is not explained by the child's age. We hypothesize that this variability is due to the changing conversational and discourse dynamics between recordings in the

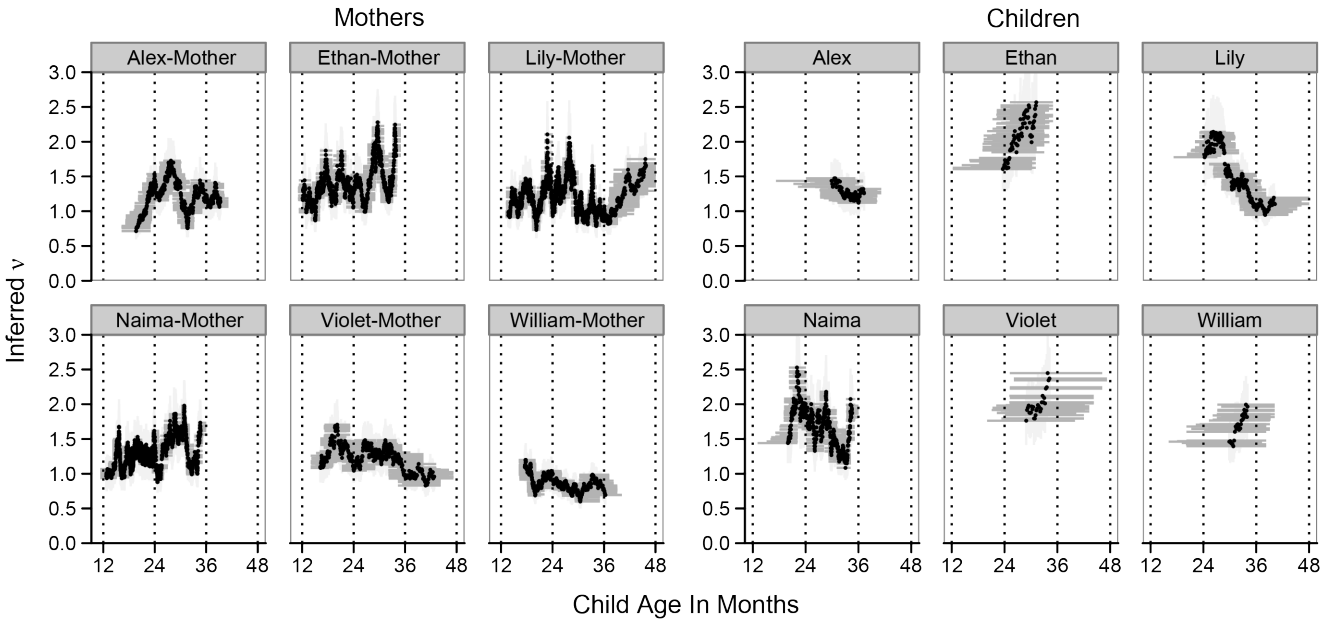

Figure 5: Sliding window analysis results. The metric model shows no clear developmental trend in children's productivity, nor a major difference in productivity between children and their mothers. On the left, black points and vertical gray bars represent the mean of the posterior and the 95% highest posterior density interval on ν; horizontal gray bars show the temporal extent of the window used in the model at each point.

corpus, leading to the introduction by chance of many nouns with similar or dissimilar determiner preferences in context. Denser data will be needed, however, to test this hypothesis more fully.

Linking Model

Although the metric model's results are suggestive of productivity from the earliest ages of children's determiner production, several aspects of the metric model limit the strength of the conclusions we can draw from it. First, the model fails to control for differences in the distribution of nouns for which determiners are produced by the speaker. For example, if children's determiner-noun productions disproportionately involve high-frequency nouns compared with adult productions, and if higher-frequency nouns tend to have more balanced determiner preference, it would inflate the metric model's estimate of children's productivity.

Additionally, an advocate of the island-learner position could justly point out that a child might produce relatively equal numbers of both determiners for a given noun Y—which disfavors low values of ν—not due to generalization but because the child has learned both "a Y" and "the Y" as islands from the input. Our linking model remedies these shortcomings by explicitly linking the determiner preference for child productions of a given noun to the experience the child has had with that noun in input from the caregiver. In the linking model, ν more directly represents the strength of a child's *generalization* across nouns: as ν approaches zero, we have a true island learner whose productions for a given noun reflect only experience with that noun from adult input; as ν approaches infinity, we have a true overgeneralizer for whom noun-specific variabilty in determiner frequencies in input are completely ignored. While it allows for a more nuanced picture of the relationship between a child's input and his or her productions, the linking model does not allow us to compare measures of adult and child productivity directly; in this sense it is complementary to the metric model.

Model Details

The generative structure for the linking model is given on the right side of Figure 4. As before, we assume a hierarchical beta-binomial model linking different noun-specific determiner preferences together into a general determiner preference with mean μ_0 and scale ν. Here, however, adult determiner productions d^A for a given noun in the child's input contribute explicitly to the child's determiner preference μ for that noun. We formalize the effect of the input on the child's determiner preference by assuming that the child acts as an ideal observer. Adult input for a given noun serves as binomial count observations, which the child combines with its beta-prior pseudocounts to yield Bayesian inference on the posterior distribution over the determiner preference for that noun.[3] We allow adult input to be downweighted by a "for-

[3]Note, however, that while the linking model contains an ideal-observer component, it is not an ideal-observer model in its totality. Most critically, μ_0 and ν are not learned by the child from adult data, but rather reflect the relationship between adult input and the child's productions. In principle, the child's productions can even be highly discrepant from the adult input, if ν is large and μ_0 does not match the overall distribution of adult determiner use. Conversely, if the posterior on μ_0 is a close match to adult determiner use, it suggests

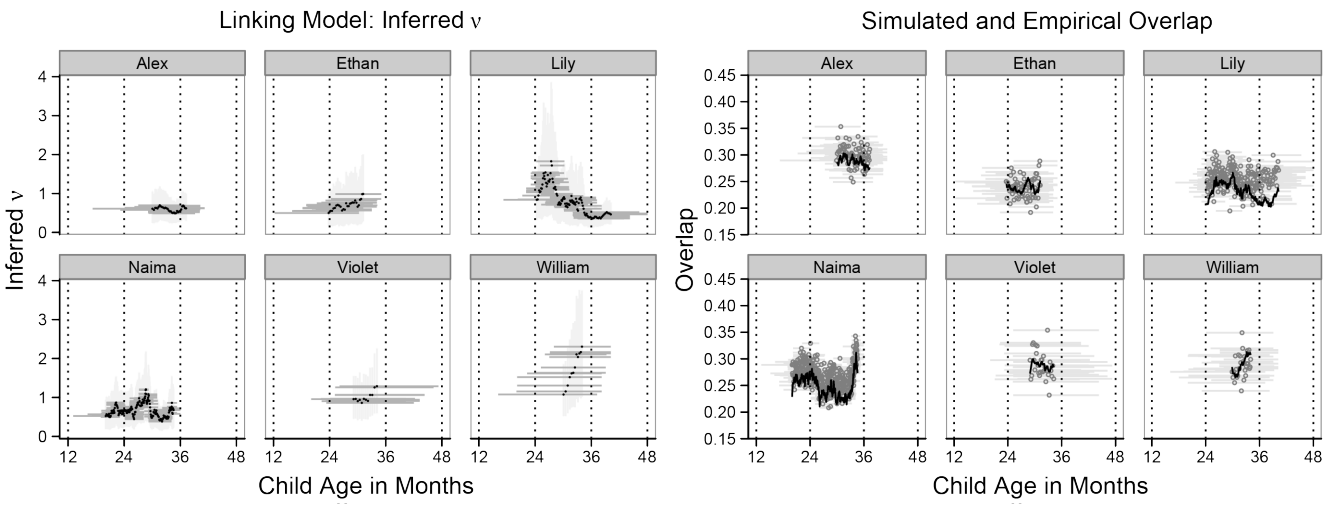

Figure 6: Linking model results for inference on ν (left) and simulated vs. empirical overlap measure (right). Black points and vertical gray bars represent the mean of the posterior and the 95% highest posterior density interval on ν, respectively. Horizontal gray bars express the temporal extent of the window used to fit the model at each point.

getting" or "noise" parameter η, motivated not only theoretically from the consideration that a child is unlikely to be able to store and learn with perfect fidelity from every determiner-noun production in its input, but also empirically: without it, it is hard for even an extreme island learner version of our model to reproduce a pattern sometimes seen in our dataset, where the determiner distribution for a given noun will be relatively balanced for the adult but highly skewed for the child.

Dataset, Results and Discussion

We used the same window size (1024 tokens) as for the metric model for a sliding window analysis using the linking model, but used all parent data up to and including the period of child usages for each window. Results of the linking model indicate that children generalize beyond the input that they receive (Figure 6, left), though there is some evidence of variation across children in generalization strength: the weakest generalizer, Alex, shows a ν around 0.6, and the strongest generalizer, William at the latest stage in our dataset, shows a ν around 2.3 Posterior means for η varied between 0.071 and 0.599, with substantial variation between children; posterior means for μ_0 varied between 0.145 and 0.717. As with the metric model, we observed no evidence for a developmental trend from lesser to greater generalization: while some children (Ethan, Violet, William) seem to show a trend toward increasing ν over time, other children (Naima, Alex) show no directional trend, and one child (Lily) has a decreasing trend.

Finally, although we have argued that the overlap measure is not useful for quantifying productivity *across* sample sizes, we can use it as a goodness-of-fit metric for our model *within* a sample. We do this by using the adult data and the joint posterior of the fitted model for each window to generated simulated determiner productions for the specific noun distribution in that window, and comparing the overlap measure for the simulated data with the empirical overlap measure in that window. For nearly all windows of all children, empirical overlap falls within the range of simulated overlaps, validating the model's overall fit to the data (Figure 6, right).

General Discussion

We constructed two models to quantify the productivity in children's early determiner usage and to compare this to that of their mothers. These models instantiated a statistical tradeoff between memorization of the observed data ("island learning") and extreme generalization. Results from both models suggested that the children in our sample were neither extreme generalizers nor extreme island learners. Contra the constructivist hypothesis, neither model provided clear evidence for developmental change in children's generalization behavior over time, and by the summary measure of productivity furnished by the metric model their speech was not quantitatively distinguished from that of their parents. Yet contra the full-productivity nativist hypothesis, there is clear evidence for item-specific combinatorial preferences between determiners and nouns (ν values are relatively low in the metric model; compare Figures 3 and 5) and that children are at least somewhat sensitive to the specifics of adult input (ν values are low in the linking model).

Nevertheless, while the current results are consistent with early productivity, our modeling work leaves unaddressed a number of issues that both preclude a conclusive judgment in this debate thus far and also point the way towards future work. As we alluded to when introducing the linking model, it is difficult to rule out the possibility that apparently "productive" determiner behavior for a given noun may reflect the child's having learned both determiners with that noun as is-

that the child is indeed generalizing from adult productions across nouns in his or her production behavior.

lands. This difficulty is compounded by the fact that though the Providence corpus is extensive, it still records only a small fraction of the total adult input each child in the corpus has received. In the future this difficulty may be addressed by more complete datasets; additionally, our model could be extended by allowing imputation of unrecorded adult data, which would allow our uncertainty regarding the content of this input to be incorporated into inferences about productivity in child behavior.

A second challenge is that an advocate of the full-generalization position could reasonably object that noun-specific determiner preferences in child productions that mirror adult input may be driven by other factors to which both adults and children are sensitive in determiner production, such as referential context (e.g., Maratsos, 1979; Karmiloff-Smith, 1981). Our model could be extended to account for these effects by conditioning determiner probabilities not only on noun identity but also on other contextual factors recoverable from corpus data; this move might allow a richer investigation of the developmental trajectory of how these aspects of the knowledge underlying fully proficient determiner deteminer use are learned and used in naturalistic production.

The gold standard for demonstrating the existence of productive knowledge of determiner syntax would, of course, be the combination of a novel noun with determiners that the child has not yet heard used with that noun. Regardless of the outcome of such a study, however, we believe that our probabilistic, data-driven approach would retain potential to advance our understanding of how linguistic knowledge develops. The modeling framework presented here provides an alternative to the extreme positions of all memorization or all generalization embodied by constructivist and nativist viewpoints. Although our model contained many simplifying assumptions, including not only those mentioned above but also the restriction to two determiners, it has given initial traction in measuring how experience from local episodes may lead to global generalizations. For the problem of determiner productivity, the simplifying assumptions can be relaxed one by one; and the general architecture can be applied to study a broad range of phenomena beyond the development of determiners, such as the emergence of plural markings and other morphological generalizations. We hope that exploring the space of models that combine the best features of both island- and generalizing-learner accounts may lead to new insights into the emergence of productive language.

Acknowledgments

Thanks to the members of the Language and Cognition Lab at Stanford and the Concepts and Cognition Lab at UC Berkeley for valuable discussion. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1106400.

References

Booth, T. L. (1969). Probabilistic representation of formal languages. In *IEEE conference record of the 1969 tenth annual symposium on switching and automata theory* (pp. 74–81).

Demuth, K., Culbertson, J., & Alter, J. (2006). *Word-minimality, epenthesis, and coda licensing in the acquisition of english* (Vol. 49).

Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of children's early english articles. *J. of Child Language*, *36*, 173–200.

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PLoS ONE*, *8*, e52500.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Karmiloff-Smith, A. (1981). *A functional approach to child language: A study of determiners and reference* (Vol. 24). Cambridge University Press.

Kramsky, J. (1972). *The article and the concept of definiteness in language*. Paris: Mouton.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Maratsos, M. (1979). Learning how and when to use pronouns and determiners. *Language acquisition. Cambridge*, 225–240.

Pine, J. M., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*, 123–138.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.*

Sagae, K., Davis, E., Lavie, A., Macwhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. *J. of Child Language*, *37*, 705–729.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *2*, 562–579.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? the case of children's determiners. *J. of Child Language*, *36*, 743–778.

Yang, C. (2010). *Who's afraid of George Kingsley Zipf?* (Unpublished manuscript)

Appendix: Model Validation with Artificial Corpora

To test the validity of our Gibbs sampling procedure procedure and establish the minimum number of DET+N samples necessary to parameterize the model, we tested the metric model on artificial noun and determiner counts generated according to known statistical properties. We varied μ was from .1 to .9 in increments of .1, and ν at .05, .1, .5, 1, 5, 10, and 50. We additionally varied the number of tokens from $2^0(1)$ to $2^{24}(1.6x10^7)$ (the upper limit corresponding to the order of magnitude of tokens heard by a child; Frank et al., 2013), with token distributions generated from both uniform and Zipfian word frequency distributions.

As in the main simulations, we estimated posteriors for the parameters μ and ν and compared with the known μ and ν used to generate the input data. MCMC chains here and in the main simulations consisted of 1000 samples after a burn-in of 1000 adaptive samples and 1000 updates, with no thinning. We employed Gelman diagnostics as well as manual inspection of traces to check for sufficient burn-in time and mixing. Grid sampling confirmed that likelihoods were sufficiently peaked to constrain parameter estimates and were consistent with posteriors produced inferred with MCMC. Measures of the reliability of inference (mean and standard deviation in the difference from the true value) helped establish a minimum window size for sliding window analyses to correspond with error less than some fixed value ε.