Representing and Learning a Large System of Number Concepts with Latent Predicate Networks

Joshua Rule, Eyal Dechter, Joshua B. Tenenbaum: {rule, edechter, jbt} @ mit.edu
MIT, 46-4053, 77 Massachussetts Avenue, Cambridge, MA 02139 USA

Abstract

Conventional models of exemplar or rule-based concept learning tend to focus on the acquisition of one concept at a time. They often underemphasize the fact that we learn many concepts as part of large systems rather than as isolated individuals. In such cases, the challenge of learning is not so much in providing stand-alone definitions, but in describing the richly structured relations between concepts. The natural numbers are one of the first such abstract conceptual systems children learn, serving as a serious case study in concept representation and acquisition (Carey, 2009; Fuson, 1988; Gallistel & Gelman, 2005). Even so, models of natural number learning focused on single-concept acquisition have largely ignored two challenges related to natural number's status as a *system* of concepts: 1) there is an unbounded set of exact number concepts, each with distinct semantic content; and 2) people can reason flexibly about any of these concepts (even fictitious ones like *eighteen-gazillion*). To succeed, models must instead learn the structure of the entire infinite set of number concepts, focusing on how relationships between numbers support reference and generalization. Here, we suggest that the latent predicate network (LPN) – a probabilistic context-sensitive grammar formalism – facilitates tractable learning and reasoning for natural number concepts (Dechter, Rule, & Tenenbaum, 2015). We show how to express several key numerical relationships in our framework, and how a Bayesian learning algorithm for LPNs can model key phenomena observed in children learning to count. These results suggest that LPNs might serve as a computational mechanism by which children learn abstract numerical knowledge from utterances about number.

Keywords: child development; concept learning; number; generalization; computational model; grammar induction

Introduction

Humans seldom learn concepts in isolation. We learn about *left* by comparing and contrasting it with *up*, *down*, and *right*, and about *red* by noting its similarities and differences with *green* and *blue*. The natural numbers (1, 2, 3, ...) are no exception: to understand a number such as *one*, we must not only ground it in terms of concepts and percepts we already know, but we must also relate it to other number concepts we are still in the process of acquiring. The natural numbers are particularly interesting in this respect. Because they are infinite, there is no way to learn all the individual concepts without learning a compositional structure for the system.

A great deal of empirical work has focused on the first part of this problem, on how initial number concepts are grounded in counting routines and the core systems of approximate magnitude and parallel object individuation (Carey, 2009; Dehaene, 2011; Feigenson, Dehaene, & Spelke, 2004). Recent studies have also proposed computational mechanisms to explain several key behavioral changes during early number learning (Piantadosi, Goodman, & Tenenbaum, 2012).

Far fewer studies have focused on the second half of the problem, on how numbers are learned as a system and partially defined with respect to each other. While the problems of how children link physical sets with the counting routine and develop their first number concepts are crucial, we direct our attention elsewhere in this paper. We focus on this second problem, on how children might acquire knowledge of an infinite number system, particularly for numbers they hear discussed but are unlikely to ever see counted out explicitly.

We ground our learning proposal in a new framework for representing number as a conceptual system, which on its own has presented a non-trivial challenge met in different ways by linguists and developmentalists. For example, Hurford (1975) proposed a single system differentiating primitive and compound number concepts, while Siegler and Robinson (1982) proposed a system with several stages of development, each containing minimal internal structure.

Our approach to representation and learning is in part inspired by, and shares much in common with, the recent family of Rational Rules models (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi et al., 2012; Ullman, Goodman, & Tenenbaum, 2012), exploring concept learning through Bayesian induction of compositional representations using sparse evidence. We agree that this framework is fundamental to understanding concept learning.

The major difference is in how our models represent concepts. In Rational Rules models, each concept is a single stand-alone rule supported by its own evidence. These rules are generated from a static grammar which defines the hypothesis space. Learning is determining which concepts (which rules) are supported by the evidence. In the model we present here, concepts are not stand-alone rules, but networks of possible relations generated according to a grammar. The hypothesis space is thus not over stand-alone rules generated by a prespecified grammar, but over millions of possible grammars, each defining a different network of relations. Learning is determining which grammar, which sets of relations, are supported by the evidence.

We begin by discussing how to represent the infinite conceptual system of natural number, and show how a particular formalism – the Probabilistic Range Concatenation Grammar (PRCG) – can represent number concepts this way (Boullier, 2005). We then show how a portion of this grammar can be learned using Bayesian inference in an LPN, a learning framework for PRCGs (Dechter et al., 2015).

A Grammar Representing Number Knowledge

Three challenges make learning systems of number concepts particularly difficult and interesting. First, very few number concepts are perceptually grounded. When, for example, did you last count exactly 254 objects? The problem only intensifies as we begin applying numbers to events, time periods, sets of objects, and eventually even other numbers. Second, the fact that there are infinitely many number concepts means that, much like in natural language sentences, the meanings of the numbers are compositional. The meaning of *four-hundred fifty-two*, for example, depends on but goes significantly beyond the meanings of *four* and *hundred*. Third, to understand numbers is also to understand the relations in which numbers participate. We are often interested in a number not for its cardinality but for some more complex property, such as whether it is more or less than another number or how it changes through addition or division. This diverse range of uses makes it impossible to fully describe *three* without referencing *two*, *four*, and eventually all other numbers.

How can we hope to represent systems of concepts which are: 1) learnable without direct perceptual grounding; 2) compositionally constructed; and 3) relationally defined? Happily, these properties are similar to those linguists face in studying natural language syntax. Grammars can be induced directly from a stream of utterances, are highly compositional, and define their constituents based on their relationships to each other rather than as discrete objects.

Motivated by this insight, we now present a grammar of number knowledge we have constructed to capture five key number relations learned during childhood and carried into adulthood: *Number*, capturing the distinction between valid and invalid number words; *Succ* and *Pred*, the successor and predecessor relations, respectively; and *More* and *Less*, the more-than and less-than relations, respectively. While seemingly basic tasks, children require years to master them (Fuson, Richards, & Briars, 1982). Whereas most work in natural language syntax uses context-free grammars, our focus on capturing structural relationships between concepts demands that we use a context-sensitive grammar. We specifically use PRCGs because they are expressive and context-sensitive while remaining relatively tractable (Boullier, 2005).

Capturing these relations with an RCG is not only possible but can be done quite compactly. Our grammar for the concepts of *Number*, *Succ*, *Pred*, *Less*, and *More* covers all numbers between zero and one-quadrillion, exclusive, and requires only 218 rules. Even considering just *Number*, *Succ*, and *Pred*, these 218 rules cover more than 10^{24} true relations. Figure 1 shows a schematic of the rules concerned with determining valid and invalid numbers, while the rest, due to space constraints, can be found online (`https://git.io/ruleEtAl2015CogSci`).

Three clarifications: first, our grammar never produces nor parses full English sentences. We model the structure of concepts, not the structure of language. When attempting to parse something like *Succ(ninety nine, one hundred)*, we assume another system more directly involved in language preprocesses utterances into predicates which are then checked against the knowledge encoded in our conceptual grammar. Second, this grammar has not been optimized for compactness or efficiency. We focus on providing a grammar that is correct, human-readable, and fits a prefix-base-suffix understanding of number, as discussed below. We do not claim this particular grammar is used by children or adults but rather that this framework, regardless of the specific grammar given, captures important aspects of concept learning, such as the rich and systematic relations between concepts, that are underemphasized in other models. Third, while the natural numbers form an infinite set, many numbers do not have convenient names. Our choice to examine what can be learned from conventional number names analyzed as words, rather than as morphemes or phonemes, means we examine only a finite subset of the natural numbers.

Intuitively, a number word like *six-hundred thirty-seven* is valid because we have six units of one hundred each and thirty-seven remaining units of one each. That is, we have some base unit (hundred) and we track both how many of them we have (six), and how many of the next smallest base unit (one) we have (thirty-seven). We denote the sum of these (six-hundred + thirty-seven) simply by concatenating the two terms from largest to smallest base (six-hundred thirty-seven). This structure is recursive. *Nine-thousand seven-hundred sixteen* is created by taking nine thousands units and tacking on a remainder, which is seven hundreds plus its remainder of sixteen ones: *nine* × *thousand* + (*seven* × *hundred* + (*sixteen* × *one*)). Note that there is no explicit mention of the base *one* in a valid number word - it is implied and marked by appending ∅, the empty string, instead of *one*.

Our grammar similarly uses a prefix-base-suffix system, and Figure 2 shows the concepts involved in deciding that *six-hundred thirty-seven* is a valid number word. As in our example above, we must show that *six* is a valid prefix for *hundred* and *thirty-seven* is a valid suffix or remainder:

$$\begin{aligned}&\mathit{Number}(\mathit{six\ hundred\ thirty\ seven}) \leftarrow \\ &\quad \mathit{Prefix}(\mathit{six}, \mathit{hundred}), \mathit{Suffix}(\mathit{hundred}, \mathit{thirty\ seven}).\end{aligned} \tag{1}$$

Six is a valid prefix for *hundred* because it is a number word representing a *ones* number, a number between one and nine. It would be incorrect for *hundred* to have no prefix, and it would also be incorrect to use a prefix larger than *nine*:

$$\mathit{Prefix}(\mathit{six}, \mathit{hundred}) \leftarrow \mathit{Ones}(\mathit{six}). \tag{2}$$

Thirty-seven is a valid suffix because it is a valid number for a previous base, in this case ∅, the ones base:

$$\begin{aligned}&\mathit{Suffix}(\mathit{hundred}, \mathit{thirty\ seven}) \leftarrow \\ &\quad \mathit{LargerBase}(\mathit{hundred}, \varnothing), \mathit{Number}(\mathit{thirty\ seven}).\end{aligned} \tag{3}$$

$$\mathit{LargerBase}(\mathit{hundred}, \varnothing) \leftarrow \mathit{PrevBase}(\mathit{hundred}, \varnothing). \tag{4}$$

Thirty-seven is one of these numbers because it is merely the concatenation of a *decade* word and a *ones* word:

$$\begin{aligned}&\mathit{Number}(\mathit{thirty\ seven}) \leftarrow \\ &\quad \mathit{Prefix}(\mathit{thirty\ seven}, \varnothing), \mathit{Suffix}(\varnothing, \varnothing).\end{aligned} \tag{5}$$

$$\begin{aligned}&\mathit{Prefix}(\mathit{thirty\ seven}, \varnothing) \leftarrow \\ &\quad \mathit{Decades}(\mathit{thirty}), \mathit{Ones}(\mathit{seven}).\end{aligned} \tag{6}$$

Ones(one).
. . .
Ones(nine).

Teens(ten).
. . .
Teens(nineteen).

Decades(twenty).
. . .
Decades(ninety).

(1, 5) Number(PBS) ← Prefix(P, B), Suffix(B, S).
Number(PB) ← Prefix(P, B).

LargerBase(X, Z) ← LargerBase(X,Y), LargerBase(Y,Z).
(4) LargerBase(X, Y) ← PrevBase(X, Y).

PrevBase(million, thousand).
PrevBase(thousand, hundred).
PrevBase(hundred, ∅).

Prefix(P, B) ← LargerBase(B, hundred), NormalPrefix(P).
(2) Prefix(P, hundred) ← Ones(P).
Prefix(X, ∅) ← Ones(X).
Prefix(X, ∅) ← Teens(X).
Prefix(X, ∅) ← Decades(X).
(6) Prefix(XY, ∅) ← Decades(X), Ones(Y).

NormalPrefix(S) ← Suffix(thousand, S).

(3) Suffix(B, PCS) ← LargerBase(B, C), Number(PCS).
Suffix(∅, ∅).

Figure 1: An RCG whose strings are valid number words. Numbered rules correspond to Figure 2.

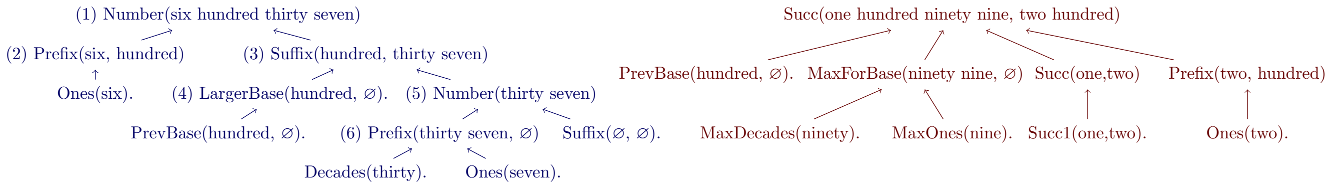

Figure 2: Example RCG parses for *Number* (Blue) and *Succ* (Red) relations.

The compositional use of simple predicates thus helps us analyze the structure of a complex phrase like *six-hundred thirty-seven* and show that while it is a valid number word, *hundred six seven thirty* is not. *Succ* can similarly be encoded (Figure 2) as can *More* (not shown), while *Pred* and *Less* can be encoded quite simply as $Less(X,Y) \leftarrow More(Y,X)$ and $Pred(X,Y) \leftarrow Succ(Y,X)$.

Learning Number Knowledge

How might children learn the number knowledge captured in the representation above? In this section, we present a computational model of learning PRCGs and take a first step toward evaluating this model against the learning trajectories and patterns of error reported in the literature on counting.

To match the literature's focus on counting, we restrict our experiments here to the successor relation, and, in particular, to learning to count from one to one-hundred. Several studies track children's learning trajectories and patterns of errors when acquiring the count sequence (Fuson et al., 1982; Miller & Stigler, 1987), making count sequence learning an interesting domain for evaluating our model of learning PRCGs against empirical data.

Latent Predicate Networks Latent Predicate Networks (LPNs) are PRCGs with three types of predicates connected in a layered fashion. *Observed* predicates are relations directly present in the data (e.g. *Succ* is observed if the data includes *Succ* relations). Observed predicates are defined in terms of layers of *latent* predicates. These relations are not directly observable in the data and their meanings are determined through learning. For example, the *Decade* predicate, which is true for "ten", "twenty", etc., might correspond to one of the latent predicates after the model is trained on pairs of successive number words. Each layer of latent predicates is defined in terms of the latent predicate layer beneath it, and the lowest layer of latent predicates is defined in terms of a collection of *lexicon* predicates, each of which is a unary predicate that is true of the atomic units (the words) of the system. The rules of the LPN consist of all definitions possible within the network architecture (for details see Dechter et al. (2015)). The parameters of the network are the probabilities of the rules.

Our model learns a distribution over the parameters of the LPN given the available data using hierarchical Bayesian inference: the model assumes that there is a prior distribution over the parameters of the LPN and, using Bayes' rule, infers a distribution over parameter values that balances the fit of the observations against the prior. We use a sparsity-inducing prior to formalize the intuition that latent predicates and rules should be shared in order to learn grammars that can generalize beyond the observed data.

Since exact inference in probabilistic grammars is computationally intractable, our model is simulated using the Variational Bayes EM approximate inference algorithm as implemented in the PRISM programming language (Sato, Kameya, & Kurihara, 2008).

All the simulations below were run on an LPN with three layers of five predicates each. The learning algorithm was run for a single iteration with a concentration parameter of $\alpha = 0.1$, and a convergence criterion of $\varepsilon = 1e-4$. We will refer to each separate simulation below as a *simulated child*.

Acquiring the count sequence Fuson et al. (1982) describes qualitative phenomena of count sequence acquisition and elaboration based on several surveys in which the authors asked American children between three and five years of age to count (either while counting a collection of objects or just reciting the count word sequence). The learning trajectories and error patterns they describe have inspired computational modeling efforts using connectionist networks; for example,

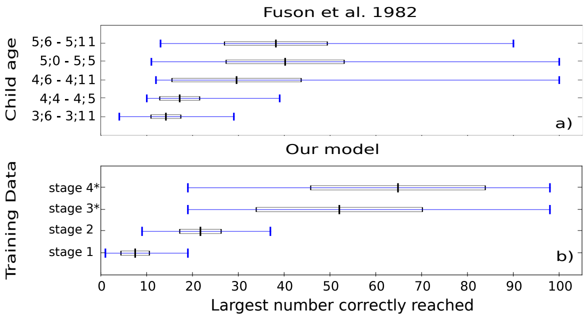

Figure 3: Our model compared with children's counting data a) Data from Fuson et al. (1982). The x-axis shows the highest number correctly reached when children were asked to count starting at "one." Boxes correspond to the standard deviation, central bands to the means, and whiskers to the range. b) Model performance, averaged over ten runs at four stages of increasing data quantity.

Ma and Hirai (1989) use an associative network to model the errors that young children typically make when learning the count sequence up to twenty or thirty. But, to our knowledge, such models have not been used to study how children acquire the count sequence beyond thirty.

Figure 3a shows the highest number correctly reached by children of various ages when tested by Fuson et al. The authors hypothesize that the large jump in range between the young three-year-olds and the four-year-olds and five-year-olds is due to the older children partially solving what they term "the decade problem" – i.e. recognizing both that there is a pattern that repeats across decades greater than twenty and that there is a particular sequence to the decade words.

We asked whether our model goes through a similar transition. To simulate learning, we generated data sets consisting of successor pairs between one and one-hundred, with the number of examples N of each pair $Succ(i,i+1)$ following the power law $N=\frac{K}{i}$, where K determines the overall size of the data set. To explore the effect of evidence quantity, and to simulate the effect that overall quantity of evidence has on a child's acquisition of the count list, we generated data sets for $K=10,100,1000,10000$, which we denote stages 1-4, respectively. The resulting histograms of data are shown in Figure 4a-d (the y-axes are logarithmically scaled).

For each of these data sets we ran our learning algorithm ten times, generating ten simulated children at each stage (the simulations differ due to different random parameter initializations). In Figure 4a-d, each line corresponds to one of the simulated children and shows the probability that the child will correctly count to the corresponding number on the x-axis. To generate this data, we asked the model for the distribution of successors for a given number and used a simple softmax decision procedure to determine the probability of the simulated child reporting each word. Specifically, if the simulated child believes x follows a with probability $p_a(x)$, then it says x after a with probability proportional to $p_a(x)^2$.

The stage 1 simulations are variable in performance, with some of simulated children unable to count further than the first few words and a few having a relatively high chance of reaching "twenty." The sharp drops in performance at "twenty," "thirty" and "forty" in stage 2, and the horizontal lines between them, indicate that here the simulation has learned the within-decade structure of the count list but is uncertain about the transitions between decades. In stages 3 and 4, nearly all simulated children master the numbers up to "twenty nine" but are unable to transition from "twenty nine" to "thirty." Only in stage 4 do we see any children making the transition from this state of knowledge to one in which they can reach "ninety nine."

The simulations in these first four stages suggest that even with large increases in the quantity of data, our model is unlikely to progress beyond "twenty nine." We hypothesized that this is due to a lack of evidence for the decade transitions. Mastering the decade transitions requires both learning that there is a special rule for the successor of numbers ending in "nine," and learning the order of the decade words. This adds considerable complexity to the grammar, and our simulations favor a more parsimonious explanation of the heavily weighted smaller numbers. Children, however, do not learn to count to a hundred by unsupervised exposure to naturally occurring count words; they are actively taught to do so. Although we know of no study of the pedagogical language used in teaching children to count, some kindergarten teaching blogs (e.g. `http://www.heidisongs.com/blog/2012/05/teaching-kids-to-count-to-100.html`) mention emphasizing decade transitions as useful in helping struggling students to learn the count sequence.

To confirm that increased emphasis on decade transitions can facilitate the transition to mastering counting up to a hundred, we created two additional data sets, stages 3^* and 4^*, that contain the same data as stages 3 and 4, respectively, but have an additional 10% of the data evenly distributed across the decade transitions (twenty nine, thirty; thirty nine, forty; ...; eighty nine, ninety). The simulated data for these stages is shown in Figure 4e-f. In both simulations, we observe a sharp increase in the number of simulated children who transition to counting to a hundred (from 0 to 4 children in stage 3^*, and 2 to 6 children in stage 4^*).

Figure 3b summarizes the simulation data for stages 1,2,3^* and 4^* for comparison against the Fuson et al. data in Figure 3a. For each stage and each simulated child, we computed the probability that the highest number reached by counting, starting from "one," would be x for $x=1,\dots,99$. We averaged these values across simulated children within a stage and used the resulting densities to calculate the means, standard deviations, and 10^{th} and 90^{th} percentiles for each stage (these percentiles were chosen to be comparable with the empirical ranges described by Fuson et al.).

In addition to examining the learning trajectories of our model, we also examined its mistakes. One interesting pattern of mistakes that young English-speaking children make

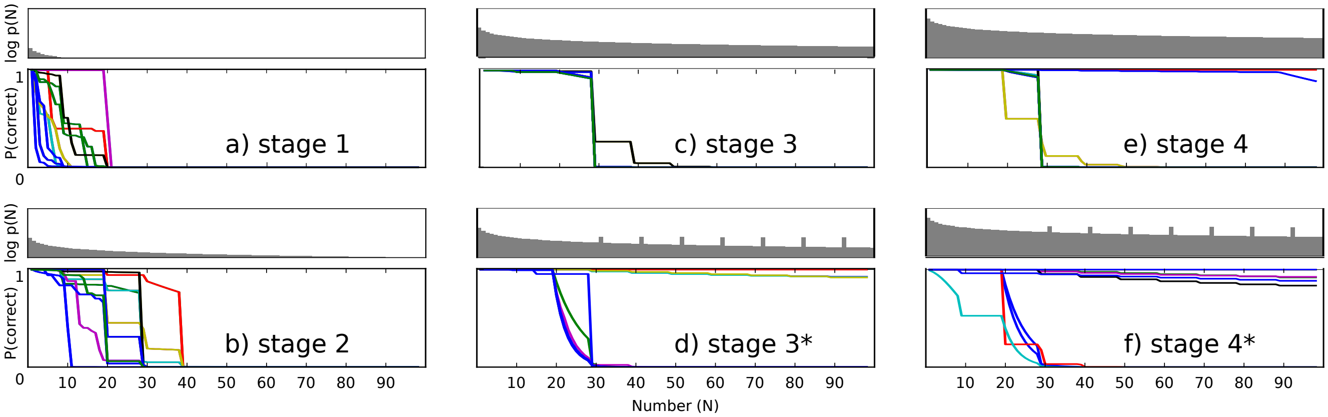

Figure 4: Our model's performance correctly reciting the count sequence. Each colored curve corresponds to a single run of the learning algorithm given the distribution of data in the histogram directly above it. For each number, N, along the x-axis, the y-axis corresponds to the probability that the model correctly counts from one up to N. The y-axes on the data histograms are shown on a logarithmic scale. The stages refer to the distributions of data available to the model (see text for details).

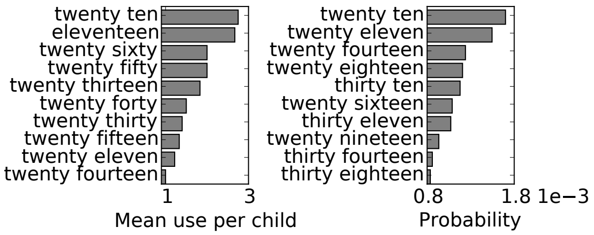

Figure 5: Top ten invented number words in children's counting. a) Data from Fuson et al. b) A simulated child at stage 2.

when reciting the count sequence is that they invent number words. Fuson et al. report that children invent such words both by combining morphological components of number words (such as "fiveteen" and "eleventy") and by combining decade words with incorrect digit-place words (such as "twenty-eleven" and "twenty-twenty"). In particular, they report that appending teen words to decade words is most common, creating sequences like "twenty-ten", "twenty-eleven", "twenty-twelve", etc. In Figure 5a we show the most common invented words that Fuson et al. report and the mean number of times a child used the word.

Since we do not model in this work how number word morphemes are composed to construct number words, our model cannot account for morphologically-based errors. We asked, however, to what extent it can model the other invented word errors that Fuson et al. report; the most common invented words are shown in Figure 5a. To compare these data to our model, we asked a stage 2 simulated child for the top ten non-number words that could appear as successor to a number word or non-number word (because this was a computationally expensive procedure, we restrict our analysis here to a single randomly selected simulation). The marginal probabilities of those non-number words is shown in Figure 5b.

Discussion

In this work, we have shown how exact number concepts and the relations among them can be represented using probabilistic context-sensitive grammars. We have also given a model for how children might learn such representations based on hierarchical Bayesian inference. Our simulations suggest this model captures several behavioral phenomena children exhibit while learning the count sequence – a critical and difficult prerequisite to adult-like numerical knowledge.

An interesting aspect of this process is the seemingly sudden transition from counting only through the first few decades to counting all the way to a hundred. Our model explains this transition as an inductive leap: for small amounts of data, learning is slow and incremental – adding a decade at a time – because the increased complexity of the conceptual knowledge is large compared to the gains in explanatory power. Eventually, however, enough evidence accumulates to warrant a more complex and more general grammar, resulting in a kind of phase transition between states of knowledge.

In many ways this phenomenon is analogous to the Cardinal Principle (CP) transition, in which younger children learning the relationship between small numbers and set sizes make slow and incremental progress when learning to count out sets matching the first three or four number words but then suddenly expand their ability to every other memorized number word. The theory that this rapid transition is due to what Carey (2009) refers to as *Quinian bootstrapping* has been formalized by Piantadosi et al. (2012) as probabilistic inference over a space of recursive programs defined by a grammar. As we do here, they explain the inductive leap of the CP transition as a result of the tension between program complexity and fit to the available data. Whereas Piantadosi et al. place a distribution over programs using a probabilistic context-free grammar, however, our model is learning a complete grammar, one that can accommodate many different concepts and relations and that can be seen as a probabilistic and declarative knowledge base.

Another difference is that our simulations require a pedagogical emphasis on critical evidence – the decade transitions – to master the count sequence robustly, suggesting that pedagogy may play an important role in facilitating these kinds of inductive leaps. Focusing on concept acquisition in slightly older children allows us to explore the relationship between computational level considerations driving inductive reasoning and the pedagogical factors enabling it in practice.

An important goal for future work is to apply our model to learning systems of number concepts in other languages besides English. In preliminary work we have applied our model to learning the Chinese number system and shown that it both learns the adult system with relative ease and explains why Chinese children generally make different patterns of mistakes than English-speaking children – in particular, why they are much less likely to invent number words like "twenty eleven" even though Chinese uses the same words to refer to both decade and ones values (e.g., "twenty one" is "two ten one") (Miller & Stigler, 1987).

Another important future step for this research will be to relate our model to those, like Piantadosi et al. (2012) for counting and Dehaene (2011) for the approximate magnitude system, that attempt to explain how abstract number knowledge becomes grounded in the perceptual and procedural primitives through which children learn about the world. The model we presented here does not attempt to explain how children come to understand that number words refer to cardinalities, though this is crucial to understanding number.

That said, we see no fundamental incompatibility between the model presented here and extensions to include approximate magnitude, object tracking, set manipulation, more complex morphology (e.g. the meaning of *-illion* or *-teen*), or different counting strategies (e.g. as used in Turkish, French, or Mandarin) as would be needed for a more comprehensive model of number learning. In fact, a key next step for us is to model the link between the relatively small set of named numbers (as modeled here) and the infinite set of numbers through more complex morphology and word invention (i.e. the -illion system, including *gazillion* or *bajillion*) or systems like Arabic or tally notation, where the infinite sequence is easier to express. We see our work here as a first demonstration of LPN's suitability for capturing a broad range of concepts in number and other semantic domains including space, kinship, and natural kinds. Whether these more general models are best approached by working strictly within the LPN formalism or by using it as one module within a more complex framework is an open question. Certainly, the human mind is more powerful than an RCG and is at least Turing-complete. RCGs provide a tractable way, however, to explore a restricted subclass of problems. The strategies and solutions we discover here are also available in Turing-complete systems, and are in fact implemented in one (PRISM Prolog), so our findings easily generalize to more expressive grammars.

More broadly, we see this paper as growing out of the hypothesis that much of human learning, including the explosion of knowledge during development, can be understood as inducing, from sparse and noisy data, a library of bits of conceptual knowledge, written in something like a programming language of thought. This vision of the *child-as-hacker* draws on and extends the notion of the *child-as-scientist* (Gopnik, 1996); not only are children forming theories about the world, but they are simultaneously developing the very conceptual language they use to formulate those theories.

Acknowledgments

The authors benefited significantly from conversations with Leon Bergen, Timothy O'Donnell and Steven Piantadosi. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, an NSF Graduate Research Fellowship, and the Eugene Stark Graduate Fellowship.

References

Boullier, P. (2005). Range concatenation grammars. In *New developments in parsing technology.* Springer.

Carey, S. (2009). *The origin of concepts.* Oxford University Press.

Dechter, E., Rule, J., & Tenenbaum, J. B. (2015). Latent predicate networks: Concept learning with probabilistic context-sensitive grammars. In *Papers from the 2015 AAAI Spring Symposium.*

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* Oxford University Press.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Fuson, K. C. (1988). *Children's counting and concepts of number.* Springer-Verlag.

Fuson, K. C., Richards, J., & Briars, D. J. (1982). The acquisition and elaboration of the number word sequence. In *Children's logical and mathematical cognition.* Springer.

Gallistel, C., & Gelman, R. (2005). Mathematical Cognition. In *The Cambridge handbook of thinking and reasoning.* Cambridge University Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*(4), 485–514.

Hurford, J. R. (1975). *The linguistic theory of numerals.* Cambridge University Press.

Ma, Q., & Hirai, Y. (1989). Modeling the acquisition of counting with an associative network. *Biological Cybernetics*, *61*(4), 271–278.

Miller, K. F., & Stigler, J. W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, *2*(3), 279–305.

Piantadosi, S. T., Goodman, N. D., & Tenenbaum, J. B. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199-217.

Sato, T., Kameya, Y., & Kurihara, K. (2008). Variational Bayes via propositionalized probability computation in PRISM. *Annals of Mathematics and Artificial Intelligence*, *54*(1-3), 135–158.

Siegler, R., & Robinson, M. (1982). The development of numerical understandings. *Advances in Child Development and Behavior*, *16*, 241–313.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.