Logical word learning: The case of kinship

Francis Mollica

Steven T. Piantadosi

University of Rochester

7 August 2019

Abstract

In this paper, we propose a framework for conceptual development through the lens of program induction. We implement this framework to model the acquisition of kinship term concepts, resulting in the first formal developmental model for kinship acquisition. We demonstrate that our model can learn several kinship systems of varying complexity using cross-linguistic data from English, Pukapuka, Turkish and Yanomamö. More importantly, the behavioral patterns observed in children learning kinship terms, under-extension and over-generalization, fall out naturally from our learning model. We conducted interviews to simulate realistic learning environments and demonstrate that the characteristic-to-defining shift is a consequence of our learning model in naturalistic contexts containing abstract and concrete features. We use model simulations to discuss the influence of simplicity and learning environment on the order of acquisition of kinship terms, positing novel predictions for the learning trajectories of kinship terms. We conclude the paper with a discussion of how this model framework generalizes beyond kinship terms and the limitations of our model.

Keywords: word-learning; conceptual development; Bayesian modeling

Logical word learning: The case of kinship

In order to acquire a language, learners have to map words to objects and situations in the world. From these mappings, they must then learn the underlying concept of the word that will generalize to new objects and situations. The mappings between words and concepts, acquired over a lifetime, will constitute the majority of information a language user stores about linguistic representations (Mollica & Piantadosi, 2019). While there is a vast literature on how children might solve the problem of mapping words to the world (e.g., Carey & Bartlett, 1978; Smith & Yu, 2008; Frank, Goodman, & Tenenbaum, 2009; Medina, Snedeker, Trueswell, & Gleitman, 2011; Siskind, 1996), we know less about children use these mappings to inform their concepts in order to generalize words to new contexts. Research on children's early word generalization has focused on uncovering biases in children's generalizations (e.g., taxonomic constraints, Markman, 1991) and explaining the mechanism and types of input children need to overcome these biases (e.g., Gentner & Namy, 1999; Graham, Namy, Gentner, & Meagher, 2010); however, research has yet to precisely predict children's behavior across the developmental trajectory. We propose a theoretical model from two first principles—simplicity and strong sampling, to scale up our understanding of how children's word meanings should change as they observe more data. In the process, we demonstrate that several seemingly unrelated patterns in children's early word use can be explained by the process of induction in naturalistic learning contexts.

Understanding how children's conceptual knowledge changes over development is a non-trivial task. It's no secret that children's early word usage does not reflect their underlying knowledge. In general, young children's definitions and, more importantly, their behavior suggest a partial knowledge of the underlying concept even though they can produce the word and appear to fully understand the word (Clark, 1973; P. Bloom, 2000). Interestingly, tasks assessing this partial knowledge have revealed systematic patterns of word use as children learn the true underlying meanings of words. Around their first birthday, children sometimes show a preference for words to label individual referents and, thus, under-extend a term to other correct referents (Clark, 1973; Kay & Anglin, 1982). For example, a young child may refer to their blanket as *blanky* and refuse to use *blanky* to refer to other blankets. Before their second birthday, children will often over-extend a term, using it to describe inappropriate but often similar referents (Clark, 1973; Rescorla, 1980). For example, children frequently over-extend *dog* to refer to any animal with four legs. In some complicated semantic domains (e.g., kinship, morality), young children continue to over-extend a term for several years. In these cases, children's over-extensions gradually shift from relying on characteristic features to more defining relations (Keil & Batterman, 1984; Keil, 1989).

While these behavioral patterns are consistently observed in children's early word use, it's unclear whether they reflect partial conceptual knowledge (Clark, 1973; Kay & Anglin, 1982), performance limitations–such as retrieving the correct word in the child's small but rapidly increasing vocabulary (Huttenlocher, 1974; Gershkoff-Stowe, 2001; Fremgen & Fay, 1980), or pragmatic reasoning (L. Bloom, 1973; Hoek, Ingram, & Gibson, 1986; Barrett, 1986). As a result, children's early patterns of word use are under-utilized as a source of data for conceptual development. A major obstacle to teasing apart these alternative hypotheses is the lack of a formalized account of conceptual development predicting children's word use over time. Specifically, what patterns of word use should we expect as children gather more data? How should these patterns hold cross-linguistically? How do these patterns change as children learn inter-connected conceptual systems (Murphy & Medin, 1985)?

Kinship is an ideal domain to demonstrate the universality of the learning mechanism and the importance of the data distribution. Kinship systems are present in almost every culture in the world; therefore, the task of learning kinship terms is present in almost every culture in the world. While the importance of kin relationships might vary across cultures, a prominent structure in the world supporting kinship terms, genealogy, is universal[1]. That being said, kinship systems show remarkable diversity across the languages and cultures of the world in terms of which relationships get expressed by words (e.g., Murdock, 1949). Analyses of the kin relationships that do get encoded in the languages of the world have shown that extant kinship systems are the optimal trade-off between communicative efficiency and simplicity (Kemp & Regier, 2012). Starting from the same underlying structures and ending with principled but diverse systems can be reconciled if we take the child's input to be the driving force in conceptual development.

The goals of this paper are i) to present a rational constructivist framework (Xu, 2007, 2016, in press) of conceptual development formalized as logical program induction, ii) to evaluate this framework against the literature on children's patterns of generalization over time—specifically under-extension, over-generalization and the characteristic-to-defining shift, and their order of acquisition. To that end, we implement a model based on this framework to learn kinship terms, providing the first formal developmental model for kinship term acquisition. The paper is organized as follows: First, we review the empirical literature on kinship term acquisition and computational models of kinship. We then flesh out our model framework and implementation. In presenting the results, we first demonstrate that the model is powerful enough to learn kinship systems of varied complexity based on its input data. We then provide simulations based on informant provided learning contexts to show that the general patterns of children's word use described above fall out naturally from framing conceptual development as program induction in

[1] Kinship as a construct potentially operates over multiple structures, including systems of address, sociological systems and social categories (Read, 2001, 2007). As a point of scope, we focus here on genealogical notions of kinship terms–i.e., kinship terms defined over a family tree.

naturalistic environments. In the process, we present evidence suggesting that children's early word use might be informative about conceptual development and derive a novel account of the characteristic-to-defining shift. To demonstrate how this model can be used to entertain important theoretical questions about how inductive biases and children's input drive children's behavior, we examine the roles of simplicity and environmental input in determining the order of kinship term acquisition. Lastly, we conclude with a discussion of novel predictions and limitations of our account.

Children's Acquisition of Kinship Terms

Interest in the acquisition of kinship terms began with Piaget (1928)'s study of logical relationships. Piaget (1928) conducted targeted interviews with 4-12 year old children to assess their knowledge of logical relations using the sibling concept as a case study. Piaget's task tested the reciprocity of sibling relationships by soliciting definitions and investigating if children could note the contradiction between the claims that "There are three brothers/sisters in your family" and "You have three brothers/sisters." Based on his interviews, Piaget proposed that children learning logical relations (like kinship) progress through three stages: egocentric, concrete relational (transitive), abstract relational (reciprocal). Piaget also noted significant increases in performance as age increased. Conceptual replications (Elkind, 1962; Danziger, 1957; Chambers & Tavuchis, 1976; Swartz & Hall, 1972) as well as more child-friendly elicitation (LeVine & Price-Williams, 1974; Price-Williams, Hammond, Edgerton, & Walker, 1977; Ragnarsdottir, 1999) and comprehension (Greenfield & Childs, 1977; Macaskill, 1981, 1982) tasks also find strong age effects in the acquisition of kinship terms; however, the explanation of age effects varies.

In terms of empirical support for Piaget's account, the literature provides sparse and conflicting evidence. Consistent with Piaget, children (5-8 years old) make less mistakes on egocentric concepts (*grandmother*) than other-centric concepts (*granddaughter*) (Macaskill, 1981, 1982). Children (4-10 years old) also perform better when questions are framed with respect to themselves (*What is the name of your sister?*) as opposed to another family member (*As for your sister Mary, what is the name of her aunt?;* Greenfield & Childs, 1977). However, equally young children succeed at taking other people's perspective when providing kin terms (Carter, 1984) and young adopted children (4-5 year olds) have more kinship knowledge than non-adopted children (Price-Williams et al., 1977). Moreover, it's unclear that children providing examples of family members when giving a definition reflects an egocentric understanding of kinship as opposed to the use of kinship terms as terms of address (for discussion see Hirschfeld, 1989). Given the limited and conflicting data on egocentric biases in kinship acquisition, we do not evaluate our model against the egocentric claims in the literature.

A second line of kinship research lies at the merger of componential analysis in anthropology (Goodenough, 1956) and the semantic feature hypothesis for word learning proposed by Clark (1973).

Componential analysis takes up the task of identifying the minimal set of features required to distinguish relevant distinctions in meaning. For example, gender is a required feature of the English kin system because gender is required to distinguish, for instance, MOTHER from FATHER. The semantic feature hypothesis posits that children acquire the semantics of a concept "component-by-component" (Clark, 1973). Thus, developmental studies of kinship acquisition could inform theoretical anthropological studies of componential analysis, especially when multiple sets of components are equally as expressive. As Greenfield and Childs (1977) points out, the pattern of children's mistakes in an elicitation task is informative about the actual features of meaning children have acquired. For example, 4-5 year old Zinacantan children's mistakes never violate the feature that siblings have common parentage; however, half of their mistakes violate gender. Whereas, 8-10 year olds never violate common parentage and gender, but violate relative age. Therefore, the componential analyses that includes features for common parentage and gender are more likely than componential analyses that do not. For our purposes, the developmental evaluation of componential analyses potentially highlights the dimensions on which children might generalize.

The semantic feature hypothesis has also been used to predict the order of acquisition of kinship terms. Haviland and Clark (1974) proposed and found evidence for simplicity to be a driving force in the order of acquisition of English kinship concepts. In their analysis, a relationship between two individuals was considered one feature. Relations that could be explain by appealing to one parent/child relationship (e.g., mother) were learned earlier than relations that required two parent/child relationships (e.g., brother). Similarly terms that required three relationships (e.g., aunt) were learned after those requiring two relationships. Surprisingly, terms that required both a parent and child relationship (e.g., brother) were learned before terms that required the same relationship twice (e.g., grandma). Further support for the semantic feature hypothesis has been found cross-linguistically in definition elicitation studies with German 5-10 years old children (Deutsch, 1979) and Vietnamese 4-16 years old children (Van Luong, 1986). A similar pattern was reported by Benson and Anglin (1987); however, they chose to explain their data not in simplicity but by differences in experience with relatives and input frequency of kinship terms. While experience seems to explain differences in adopted children, there was no effect of household size on kinship acquisition (Price-Williams et al., 1977). In general, the extent to which simplicity and experience contribute to the order of acquisition of kinship terms is still an open question. We resume discussion of these contributions in our analysis of order of acquisition effects from model simulations.

To summarise, studies on kinship term acquisition document a protracted developmental trajectory, providing modest evidence for patterns of over- and under- extension in children's use of kinship terms; although the exact patterns of extension vary across cultures. For example, Bavin (1991) and Greenfield and Childs (1977) find gender over-extensions in Walpiri and Zinacatan children's kin usage; whereas, Price-Williams et al. (1977)'s study of Hawiian and the studies on English kin acquisition report no incorrect gender extensions. Interestingly, the children in these studies are well older than the age range where the typical patterns of over- and under- extension are observed. While not all of these studies solicit definitions, the elicitation tasks used are still likely to be challenging for children who have limited verbal ability. Therefore, we should take these patterns with a grain of salt, as young children might not understand the task and older children might lack the verbal ability to articulate their knowledge. Given these limitations, it is unclear that these patterns should fall out of a model of conceptual development as opposed to a model of how children verify semantics or produce labels. This makes it all the more interesting if these patterns do emerge naturally from the inductive learning process, which would suggest that conceptual development may still be contributing to these patterns despite the limitations of the task.

To further ground the possibility of conceptual development giving rise to patterns of over- and under- extension, it is worth mentioning a related field of studies regarding the characteristic to defining shift observed in children's knowledge (Keil & Batterman, 1984; Keil, 1989; Landau, 1982). In Frank Keil's studies, children are presented with scenarios of a concept–take for example the concept, GRANDPA–that emphasize either characteristic features but not defining features (e.g., a nice old man who isn't related to you) or defining features but not characteristic features (e.g., your parent's evil father). Young children (mean 5;7) are more likely than older children (mean 9;9) to accept a scenario with characteristic features as being true than a scenario with defining features but not characteristic features. Older children are more likely than younger children to accept the scenarios with the defining features of the concept. Remarkably, even some of the older children were not at perfect performance, suggesting that there is significant conceptual development still taking place in kinship beyond the ages in which one typically observes patterns of over- and under- extension. Given this timescale, we argue that children's over-extensions and under-extensions might actually be due to conceptual development—in particular, rational construction of a logical theory—as opposed to performance-based or pragmatic-based alternative explanations.

In this paper, we implement an ideal learning model using the default assumptions from the rule-based concept learning literature. The model framework is designed to learn a kinship system consistent with the input; however, the model is *not* designed to match the patterns of behaviors children demonstrate when learning kinship. We evaluate the model against these patterns of behavior to show that a first principles learning mechanism provides an explanation for the patterns of over- and under- extension behavior we see in children even though there was no design pressure to do so. Further, we expand the model by adding assumptions about the learning context (via interviews) and the environmental distribution of data to show that when this first principles learning model operates under naturalistic contexts and distributions of data, it predicts both a characteristic-to-defining shift and the order of kinship term acquisition that we observe in children. Lastly, we identify how the model could be used to identify primitives functions, and test early pragmatics or retrieval issues in children's word use.

Computational Models of Kinship

From a formal modeling perspective, kinship is an ideal domain for studying how children's conceptual knowledge develops into the rich rule-like concepts and conceptual systems seen in adult definitions. Kinship easily lends itself to logical representation (e.g., Greenberg, 1949; Wallace & Atkins, 1960). It is relatively clear how to extensionally define the conceptually-aligned upon meanings of kinship terms. Kinship systems are relational concepts by nature, which allows us to look at the acquisition of concepts that are difficult to reduce to similarity. Further, kinship is a great test-bed for how inter-related conceptual systems are learned, as adult kinship knowledge suggests inter-related, not independent, concepts for kinship terms[2]. That being said, most of the previous computational models of kinship had other motivations.

The earliest computational models of kinship were primarily concerned with automating componential analysis. Given a large set of features about each kinship term in a language, what is the minimal set of features required to distinguish the terms (Goodenough, 1956; Lounsbury, 1956)? As Burling (1964) was quick to point out, the componential analysis of a kinship dataset has many possible solutions. Pericliev and Valdés-Pérez (1998) implemented a model to perform componential analysis that finds all possible solutions possessing both the smallest number of unique features and the shortest feature conjunctions required to define all terms. Proving Burling's point, Pericliev and Valdés-Pérez (1998)'s automated analysis of Bulgarian kinship systems found two equally complex feature inventories that use different features. To complement componential analyses, several behavioral studies utilized multidimensional scaling techniques to uncover the dimensionality of kinship components and arbitrate between different componential analyses (e.g., Wexler & Romney, 1972; Nakao & Romney, 1984). Recent work in the spirit of componential analysis has taken up the search for kinship universals using optimality theory (Jones, 2010) and Bayesian methods (Kemp & Regier, 2012).

Early connectionist models have used learning kinship as a test case for distributed models of relational concepts. Hinton et al. (1986)'s family tree task focused on learning an encoding for the family members on a given tree and the relationships between them. The connectionist model received input vectors reflecting an individual on the tree (e.g., *Simba*) and a kinship relationship (e.g., *father*) and output the individuals on the tree who completed the kin relation (e.g., *Mufasa*). The model learned interpretable

[2] We explore inter-related learning schemes in Appendix C.

embeddings for people on the tree, such that semantic features (e.g., gender) could be easily extracted. However, the relationship embeddings were not interpretable and the generalization performance of the model was poor. Using linear relational embedding, Paccanaro and Hinton (2001) greatly improved the generalization performance. Their model learned relationships as rotational transfers between point vectors in a space reflecting individuals. The model was successful at completing the implicit structure behind the training data especially when incorporating held out people into the system; however, the model did not fare as well when incorporating held out relations to the model. The model learns the family members and all of the relations on the tree without learning the actual tree structure. Therefore, it's unclear how well the relations learned will generalize to an entirely new family tree. Importantly, neither connectionist model makes any claims about children's behavior while learning. Though, Paccanaro and Hinton (2001) points out the most common generalization error was over-extension of sibling terms to include the speaker—i.e., the common failure of Piaget (1928)'s logic problem.

More recent computational models have approached the acquisition of kinship knowledge through a Bayesian relational-learning or theory-learning perspective. The Infinite Relational Model (IRM; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006) uses the presence or absence of relations between individuals and kinship term use to learn groupings of these individuals and properties shared by the groups, which are diagnostic of the relationship. For example, applying the IRM to data from a complex Australian kinship system results in groups of individuals that share "diagnostic" kinship relevant feature dimensions such as age and gender. Katz, Goodman, Kersting, Kemp, and Tenenbaum (2008) proposed a generative model similar to the IRM but with a richer representation system based in first order logic, Horn Clause Theories. Their model learns each individual's kinship relevant properties and the abstract rule governing how those properties give rise to the kinship relation. Katz et al. (2008)'s representation scheme has two advantages over the IRM. First, Horn Clause Theories are compressible probability models that license deductive inference, inductive inference and deductive inferences based on inductive inferences. Second, Horn Clause theories are context independent, which allows one's knowledge of kinship to easily generalize beyond the observed/training data. Similar first order logic representation schemes have been used to analyze the space of all possible kinship systems to identify the pressures that influence which kinship systems are extant in the world (Kemp & Regier, 2012). Surprisingly, extant kinship systems are found at the optimal trade-off between simplicity and communicative efficiency. Yet again, these computational models of kinship provide proof of learnability without making claims about children's behavior during learning.

Our model builds off the intuitions of the Bayesian models. Following Katz et al. (2008), we adopt the use of a context independent representation scheme. Our model also incorporates a pressure for simplicity, which is line with Kemp (2012) and other studies of kinship acquisition (e.g., Haviland & Clark, 1974). Our approach will depart from past models in two ways. First, our representation scheme is inspired by set theory instead of horn clauses[3], which provide poor fit to adult's induction and generalization behavior (Piantadosi, Tenenbaum, & Goodman, 2016). Operating over sets is a more functional representation scheme that emphasizes generating members of those sets, or possible word referents, as opposed to computing the truth of a logical expression. Second, we aim to provide not only a proof of learnability but an evaluation of the full developmental trajectory of concepts (illustrated here with kinship), including the the common behavioral patterns of mistakes children display. As a result, we use our model to formalize the contributions of data, the learning context and inductive biases towards explaining the patterns of children's behavior while learning these concepts; whereas, previous learning models with logic-like representation schemes have not been evaluated against behavior.

The approach: Concept induction as program induction

The basic premise of our approach is that conceptual knowledge can be likened to a computer program. One role of a concept is to point to entities in the context. For example, your concept of CHASE allows you to detect entities in the context that move in a particular relationship to each other as opposed to static entities or randomly moving entities. In this regard, a concept's ability to denote entities is like a program that takes as input a context of potential referents and returns a set (possibly empty) of referents consistent with that concept. We formalize this metaphor by defining concept induction as probabilistic program induction (e.g., Lake, Salakhutdinov, & Tenenbaum, 2015; Piantadosi & Jacobs, 2016; Goodman, Tenenbaum, & Gerstenberg, 2015; Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

This metaphor capitalizes on several similarities between programs and concepts. First, both programs and concepts are relational in nature. Concepts are defined in terms of both their extension and their relations between other concepts (e.g., DOG and WOLF share common origin). Whereas, programs can be mathematically defined in terms of input/output relations. Second, conceptual development and program induction both emphasize the dynamic nature of knowledge. When a young child originally pieces together a concept, it can be thought of as chaining inferences about what underlying features or relationships are good approximations to the concept's true meaning. Similarly in program induction, the model is chaining inferences about what underlying base functions or relationships between base functions are good approximations to the program's desired output. Lastly, concept and program induction can both result in many intensionally distinct representations that are extensionally equivalent. The principles that a programmer might use to choose between two equivalent representations (e.g., simplicity, minimal hidden structure and ease of deployment) are the same principles we see in children's explanations (e.g., Walker,

[3] Although see Mollica and Piantadosi (2015) for a first order logic implementation of our model.

Bonawitz, & Lombrozo, 2017; Johnston, Johnson, Koven, & Keil, 2016).

We flesh out our framework at the computational level of analysis (Marr, 1982) as an ideal learner model, which illustrates how a rational learner might solve the problem of program induction given properties of the environment and prior inductive biases (Tenenbaum, Griffiths, & Kemp, 2006). This approach is also a rational constructivist approach in that we are looking at how data drives the construction of a program (Xu, 2007, 2016, in press). In the past decade, research in this tradition has provided rich accounts of causal learning (e.g., Goodman, Ullman, & Tenenbaum, 2011), language learning (e.g., Chater & Vitányi, 2007), number learning (Piantadosi, Tenenbaum, & Goodman, 2012) and theory learning (Ullman, Goodman, & Tenenbaum, 2012). For our purposes, this approach comes with several advantages. First, the resulting family of models are explanatory in nature, meaning the behavioral predictions of the model can be attributed to underlying knowledge states (or competence) as opposed to performance concerns. Second, our model is sensitive to different data distributions, which provides a technique to address the effect of different data distributions on learning. Looking to the future, Bayesian data analyses linking this model to behavioral data can inform us about prior biases (Piantadosi et al., 2016; Hemmer, Tauber, & Steyvers, 2015). In this form, our model would no longer be an ideal learner but an arguably stronger descriptive Bayesian model (Tauber, Navarro, Perfors, & Steyvers, 2015).

The Model

For our ideal learner model, we must specify three components: a hypothesis space over concepts $\mathcal{H}$, a prior over hypothetical concepts $P(h)$ for $h \in \mathcal{H}$ and a likelihood function $P(d|h)$ to score the hypothesis according to the data d. The hypothesis space reflects the cognitive architecture supporting learning. For example, our hypothesis space consists of compositional functions over family trees. The prior reflects the inductive biases that we suspect children bring to a learning task. While there are many reputed biases in children's word learning (e.g., the whole object bias Markman, 1990), we only adopt a bias for simplicity in our model. The likelihood reflects how we think the data (i.e., instances of referential word use) are generated. For our model we adopt a noisy size principle likelihood.

For implementing our model, we must also specify how we simulate data for our learning analyses. Here, a data point d is a collection of four objects: a *speaker*, who uses a *word* to refer to a *referent* in a *context* (detailed further below). We model learning as the movement of probability mass across a hypothesis space as a function of observing data. Following Bayes rule, the posterior probability of a hypothesis h after observing a set of data points D is:

$$P(h|D) \propto P(h) \prod_{d \in D} P(d|h). \tag{1}$$

We will discuss each component in turn.

Hypothesis Space

Constructing the hypothesis space over possible programs involves specifying primitive[4] base functions for kinship that are available to the learner and the method by which these functions compose to form hypotheses. In our model we specify several types of base functions—tree-moving functions (parent, child, lateral), set theoretic functions (union, intersection, difference, complement), observable kinship relevant properties (generation, gender, co-residing adult[5]), and variables—the speaker (denoted X) and the individuals in the context. Tree-moving functions take as argument a reference node in a tree and return a set of nodes satisfying a specific relationship on the tree. As justification for including tree primitives, we note that affording these abilities to children is a common assumption in the literature (e.g., Haviland & Clark, 1974). Set functions allow for first-order quantification, which has been shown to be relevant for adults' concept acquisition (Piantadosi et al., 2016; Kemp, 2012). We assume that gender and generation can be approximated by children and that children can compute functions over speakers. Given the late timescale of children's acquisition of kinship concepts, we feel these assumptions are appropriate. That being said, that children learn these functions is a nontrivial assumption and what these conceptual primitives mean is an interesting and important part of learning the conceptual system for future study.

Unlike linguistic or componential analyses, we do not intend for these base functions to be a complete account of all of the functions required for learning kinship systems or all of the function children might bring to the task. For example, children would require primitives to compute relative age or patrilineage to learn some kinship systems (e.g., Japanese and Korean). On the flip side, children might approach the task with ultimately unnecessary primitive functions, which we will explore further in the section on the characteristic-to-defining shift. In choosing these primitives, we have attempted to focus on learning at a level where the base functions are effectively independent of each other. It is easy to see how one could decompose certain primitives into one level less of abstraction (e.g., generation might be represented in terms of primitives that check for perceptual features) or how one could choose to augment

[4] Our use of "primitive" reflects the atomic nature of the functions within the kinship domain and is not a claim about innateness.

[5] For ease of computational search, we modified the primitives used to capture the complex relations in Yanomamö. Specifically, we modified the generation functions to compute only related people at a given tree depth as opposed to all people at that tree depth. In the same vein, we added co-residing adult as a primitive only for Yanomamö. Both of these primitive could be constructed out of the other primitives but these modifications greatly decreased computational search time. That being said, related-generation and co-residing adults are easily noticed by children and would serve as a strong cue for relevant genealogical relationships in some complex kinship systems.

this set at a greater level of abstraction (e.g., adding a sibling primitive). For any model of learning, the granularity and span of a hypothesis space depends on the characterization of the learning problem. For our purpose, we are not interested in how children develop their function for gender or even the family tree itself. We are focused on how one learns relations over a structure and these primitives are an appropriate set to investigate this learning problem.

Our general findings will not strongly depend on any particular base function inventory; however, inventories can make different predictions about the precise pattern and timing of children's behavior over learning (see Piantadosi et al., 2016, for a method to evaluate different primitive inventories in a similar model framework). Currently there is insufficient empirical data and inconsistent qualitative reports to properly evaluate the precise predictions of different primitive inventories; however, we can still evaluate the coarse-grained predictions of the model: learnability, under-/over- extension and order of acquisition. For a more detailed discussion of hypothesis spaces see Perfors (2012).

SET $\xrightarrow{1}$ union(SET,SET)	SET $\xrightarrow{1}$ parent(SET)	SET $\xrightarrow{1}$ generation0(SET)	SET $\xrightarrow{1}$ male(SET)
SET $\xrightarrow{1}$ intersection(SET,SET)	SET $\xrightarrow{1}$ child(SET)	SET $\xrightarrow{1}$ generation1(SET)	SET $\xrightarrow{1}$ female(SET)
SET $\xrightarrow{1}$ difference(SET,SET)	SET $\xrightarrow{1}$ lateral(SET)	SET $\xrightarrow{1}$ generation2(SET)	SET $\xrightarrow{1}$ sameGender(SET)

SET $\xrightarrow{1}$ complement(SET)	SET $\xrightarrow{1}$ coreside(SET)	SET $\xrightarrow{\frac{1}{37}}$ concreteReferent	SET $\xrightarrow{1}$ all
			SET $\xrightarrow{10}$ X

Table 1

The Probabilistic Context Free Grammar (PCFG) specifying the base functions and the rewrite rules that govern their composition. Each hypothesis starts with a SET symbol and there are 37 *concrete referents in our learning context.*

We compose the base functions using a probabilistic context free grammar (PCFG; see Table 1) following Goodman et al. (2008); Piantadosi et al. (2012); Ullman et al. (2012). Briefly, a PCFG is a set of rewrite rules which describe how functions can compose while defining a potentially infinite space of possible compositions. For example, the composition leading to the concept of GRANDPA would require applying the male rule, parent rule, parent rule and speaker rule, resulting in the program: $male(parent(parent(X)))$. A program can then be evaluated in a context to produce a set of possible referents[6]. Here, we use a PCFG as a tool to generate a finite approximation to an infinite hypothesis space, not as a model of cognition. In addition to defining an infinite space, a PCFG also provides a probability distribution over that space. In this distribution, we weight each rule equally as likely with two exceptions. First to prevent infinite recursion when generating hypotheses, the speaker, X, is weighted 10 times as likely as the other rules. Second, we divide the weight for concrete referents equally among the

[6] We make the assumption that programs do not return the speaker as referent–i.e., a bias against interpreting kinship terms as self-referential. The reported results are robust if we relax this assumption.

individuals in our context (detailed below).

Simplicity Prior

One advantage of using a PCFG is that it builds in a natural prior towards simplicity. Hypotheses that compose more rules are less probable than hypotheses that compose less rules. We motivate this bias towards simplicity in several ways. First, adults have been shown to learn simpler concepts faster than complex concepts (Feldman, 2003, 2000). Second, children prefer simpler explanations over more complex explanations (Lombrozo, 2007; Bonawitz & Lombrozo, 2012)–although see (Walker et al., 2017). In language learning, simplicity has been suggested as a guiding principle (Chater & Vitányi, 2007). Further in kinship, simplicity has been proposed as the driving factor behind the order of acquisition of kinship terms (Haviland & Clark, 1974). In a global analysis of all possible kinship systems, simplicity is a good predictor of which kinship systems are actually observed in the languages of the world (Kemp & Regier, 2012). Therefore, we believe simplicity is an important inductive bias to be incorporated in our model. The prior probability of a hypothesis, h, according to our PCFG is:

$$P(h) = \prod_{r \in h} P(r), \tag{2}$$

where r reflects a single use of a base function following the rules in the PCFG (Table 1).

Size Principle Likelihood

The last component of the model to be specified is the method of scoring each hypothesis according to the data. Based on past research with adults (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001), children (Xu & Tenenbaum, 2007a, 2007b; Lewis & Frank, 2018) and infants (Gweon, Tenenbaum, & Schulz, 2010), we use a size-principle, or strong sampling, likelihood for our model of concept induction. This choice of likelihood comes from the notion that the data we observe is generated from a structure in the world (i.e., strong sampling) as opposed to randomly generated (i.e., weak sampling). In strong sampling, the learner weighs positive evidence with respect to their hypothesis about how the data were generated; whereas, in weak sampling each data point of positive evidence is weighed equally regardless of how likely it was to be generated. As a result, positive evidence for a hypothesis only distinguishes between hypotheses under a size principle likelihood. For example, consider a learner trying to decide if apples are small, red fruit or if apples are just fruit. Under a strong sampling likelihood, observing a small red apple would provide more evidence for the hypothesis that apples are small red fruit than for the hypothesis that apples are fruit because the data better matches the predictions of that hypothesis. Under a weak sampling likelihood, the same data point would be equally likely under both hypotheses. Strong sampling is a powerful likelihood function that can lead to convergence on the true generative process of the data from positive evidence alone (Tenenbaum, 1999) and even in the presence of significant noise (Navarro, Dry, & Lee, 2012).

For our model we use a noisy size principle likelihood, which marginalizes over two possible ways a learner might think the data was generated. First, the data might be generated according to the learner's current hypothesis. For a given context, there is a finite set of data points that a learner expects to receive. Following a size principle likelihood, data points are sampled randomly from these expected data points: $\frac{1}{|h|}$, where $|h|$ is the number of unique data points (i.e., speaker-word-referent combinations) that a learner expects to see in a given context. Second, a learner might think that a data point was generated by noise—i.e., randomly mapping a speaker, word and referent. In this case, the probability of a data point is given by $\frac{1}{|\mathcal{D}|}$, where $|\mathcal{D}|$ reflects the number of all possible speaker-word-referent pairs in a given context. Our noisy size principle likelihood mixes these two generative processes together by adding a new parameter α reflecting the reliability of the data. At high values of α, the learner thinks that most of the data is being generated by their conceptual hypothesis; whereas at low values of α, the learner thinks the data they see is randomly generated. Combining both of these processes, our likelihood function is given by:

$$P(d|h) = \delta_{d \in \{h\}} \frac{\alpha}{|h|} + \frac{1-\alpha}{|\mathcal{D}|}. \tag{3}$$

Having a noisy process allows us to account for any issues the learner has mapping words to referents, or resolving the mapping for genitive (e.g., *your daddy*) or allocentric (e.g., a mother saying *daddy is coming*) uses of kinship terms. If the learner cannot successful map words and referents, they should act as if their data is being generated randomly, which would be implemented in the model as having low values of α.

It is also worth mentioning that the latent scope bias observed in adults (Khemlani, Sussman, & Oppenheimer, 2011) and children's explanations (Johnston et al., 2016) makes similar predictions as a size principle likelihood. According to a latent scope bias, adults and children prefer explanations that both match all of the observed data and do not predict data that is not observed. Therefore, we think that the size principle likelihood is an appropriate choice as it captures both intuitions about the data distribution and explanatory preferences.

Environmental Assumptions for Simulating Data

Our model acts as a linking hypothesis between data, inductive biases and word use/generalization. Ideally, we should be using this model on empirical measures of comprehension and production to predict children's future word use and to infer the inductive biases and conceptual architectures supporting conceptual development. Unfortunately, there are no existing data sets that either quantitatively measure children's kinship term use or span the nine years of a single child's experience with kin and kinship terms with the required detail to fully specify this model. As a result, we adopt a simulation approach, which generates predictions about children's word use from first principle assumptions about data distributions and inductive biases. We can then qualitatively compare our predictions to the trends in children's behavior reported in the literature.

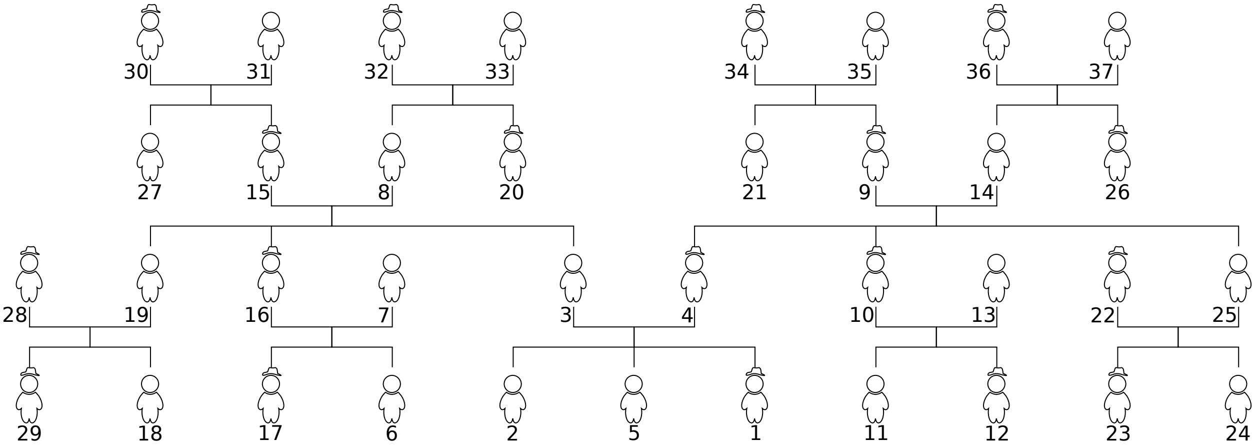

Figure 1. Family tree context for our simulations. Connections above figures reflect parent/child relationships. Connections under figures reflect lateral/spousal relationships. Male denoted with hats. Numbers reflect the rank order of the amount of interaction a learner (i.e, 1) has with the other individuals on the tree.

For our model, a data point has four components, the speaker, the word, the referent and the context. The context is a family tree, which contains each member of the family, their parent, child and lateral connections and their gender (see Figure 1). To simulate the data for learning, we first generate all true possible data points given the target word and the context. We then sample data points from the true set with probability α or construct a random data point with probability $1 - \alpha$. For all analyses reported in the paper, α was set at 0.90.[7] In simulating the data this way, we make two simplifying assumptions. First, we are only sampling the data from one family tree and it is likely that children are exposed to multiple family tress. This limitation is mitigated to some extent by our choice to vary the speaker, which changes the anchor on the tree across data points. Second, allowing the speaker to vary does not capture the use of genitives or perspective taking—i.e, we assume that the referent is always with respect to the speaker.

Results

We divide the results into three sections: Model Outcomes, the Characteristics-to-Defining Shift and Order of Acquisition. In Model Outcomes, we first check that the model successfully learns the conventionally agreed upon extension for each kinship term in finite amounts of data. We conduct this analysis using four different kinship systems: Pukapukan, English, Turkish and Yanomamö. We then take

[7] In Supplementary Figure A1, we emulate the simulations conducted by Navarro et al. (2012) to demonstrate that our main findings are robust under realistic values of α.

a closer look at how the model behaves locally at the outset of learning to demonstrate how children's early preference for concrete reference–i.e., under-extension, naturally follows from the process of induction with few data points. Afterwards, we look at how broad patterns of over-generalization fall naturally out of the process of induction when trading off simplicity and fit to the data.

In Characteristic-to-Defining Shift, we augment the model's hypothesis space, allowing rules based on characteristic features (e.g., UNCLE : $union(big, strong)$). We first replicate our previous analyses using simulations based on naturalistic learning contexts–i.e., informant provided family trees. For each word learned by each informant, we demonstrate the characteristic-to-defining shift. We discuss how the characteristic-to-defining shift arises from properties of the learning context and under what circumstances we would predict to see a characteristic-to-defining shift.

In Order of Acquisition, we return to an open question in the kinship acquisition literature: is the order of acquisition driven by experience or the conceptual complexity of the kinship relations? We evaluate the model predicted order of English kinship acquisition against the empirically observed order of concept acquisition in children. We illustrate that while the simplicity of the minimal description length correct kinship concepts aligns with the observed order of acquisition in children, the model does not predict acquisition in that order. Inspired by accounts of children's experience with kin relations (Benson & Anglin, 1987), we simulate several plausible data distributions based on kin experience and find that the order of acquisition is more likely driven by naturalistic data distributions than by conceptual simplicity.

Model Outcomes

The model learns typologically diverse systems as input varies. By framing concept induction as program induction, we can look at how the same inductive mechanism and primitive functions can give rise to very different programs depending on the data provided for learning. A breadth of ability is logically required for explaining how children learn a range of kinship systems across typologically diverse languages and cultures. We first simulated data for four kinship systems that vary in their complexity and are common in the languages of the world: Pukapukan, English, Turkish and Yanomamö. In the tradition of Morgan (1871), Pukapukan, English, Turkish and Yanomamö are from the Hawaiian, Eskimo, Sudanese and Iroquois family of kinship systems respectively. Extensions for the kinship terms of these languages are provided in the insets of Figure 2 and Table 2. The Pukapukan kinship system is relatively simple, with six kinship terms that are fully described by generation and gender. The English kinship is slightly more complex, with nine terms that require representing parent/child relations. The Turkish system is even more complex with high specificity in the first generation. In addition to requiring tree moving functions, the fourteen kinship terms reflect increased specificity in referents, separating paternal and maternal brothers and sisters and their spousal relationships. The Yanomaö system has interesting structures

	Word	**Extension**	**MAP Hypothesis**
Pukapuka	*kainga*	Z, PGD, PED	difference(generation0(X), sameGender(X))
	matua-tane	PB	male(child(parent(parent(X))))
	matua-wawine	PZ	female(child(parent(parent(X))))
	taina		B, PGS, PES intersection(generation0(X), sameGender(X))
	tupuna-tane	PF	male(child(parent(parent(parent(X)))))
	tupuna-wawine	PM	female(child(parent(parent(parent(X))))
English	*aunt*	PS, FGW	female(difference(generation1(X), parent(X)))
	brother	B	male(child(parent(X)))
	cousin	PGC, PGEC	difference(generation0(X), child(parent(X)))
	father	F	male(parent(X))
	grandma	PM	female(parent(parent(X)))
	grandpa	PF	male(parent(parent(X)))
	mother	M	female(parent(X))
	sister	Z	female(child(parent(X)))
	uncle	PB, PGH	male(difference(generation1(X), parent(X)))
Turkish	*abi*	B	male(child(parent(X)))
	abla	Z	female(child(parent(X)))
	amca	FB	intersection(sameGender(*fabio*), difference(child(parent(male(parent(X)))), parent(X)))
	anne	M	female(parent(X))
	anneanne	MM	female(parent(female(parent(X))))
	baba	F	male(parent(X))
	babaanne	FM	female(parent(male(parent(X))))
	dayi	MB	male(child(parent(female(parent(X)))))
	dede	PF	male(parent(parent(X)))
	eniste	PGW	intersection(lateral(child(parent(parent(X)))), male(complement(parent(X))))
	hala	FZ	female(child(parent(male(parent(X)))))
	kuzen	PGC, PGEC	difference(generation0(X), child(parent(X)))
	teyze	MZ	difference(difference(female(generation0(female(parent(X)))),X),parent(X))
	yenge	PGH	difference(female(generation1(X)),union(child(parent(parent(X))),parent(X)))
Yanomamö	*amiwa*	Z, FBD, MZD	female(child(close(X)))
	eiwa	B, FBS, MZS	male(child(close(X)))
	haya	F, FB	male(close(X))
	naya	M, MZ	female(close(X))
	soaya	MB	male(difference(generation1s(X), close(X)))
	soriwa	MBS, FZS	difference(male(generation0(X)), child(close(X)))
	suaboya	MBD, FZD	female(difference(generation0(X), child(close(X))))
	yesiya	FZ	difference(female(generation1s(X)), close(X))

Table 2

The maximum-a-posterior (MAP) hypotheses after learning. F*:father,* M*:mother,* P*:parent,* S*:son,* D*:daughter,* C*:child,* B*:brother,* Z*:sister,* G*:sibling,* H*:husband,* W*:wife,* E*:spouse*

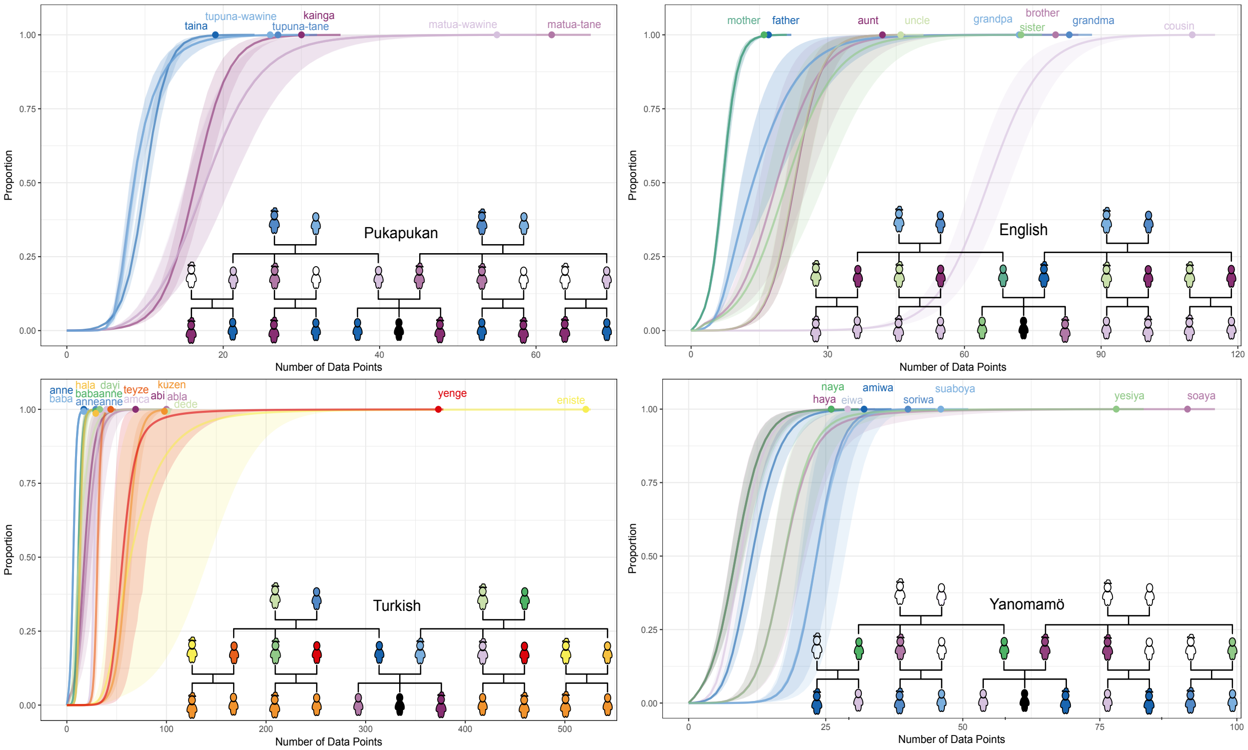

Figure 2. Average lexicon posterior-weighted accuracy for each word as a function of data points of that word. Shaded region denotes 95% bootstrapped confidence intervals. Insets show the color-coded extension of the terms.

representing cross- and parallel- cousins. Without coresidence base functions, Yanomamö is the most complex, requiring both tree and set functions to specify cross- and parallel- cousins.

Figure 2 shows the predicted learning curves for each kinship term in Pukapuka, English, Turkish and Yanomamö. The x-axis shows the number of data points for each word observed by the child. Note the differences in scale across languages. The y-axis is the probability that a learner has acquired the conventionally-aligned upon meaning of that term–i.e., extends the term appropriately. The shaded region represents the 95% bootstrapped confidence interval. The line for each word is color coded to match the word's extension in the inset. Table 2 provides the maximum-a-posteriori hypotheses learned for each kinship term.

Despite variegated reliance on base functions and differential complexity, the model successfully learns the conventional kinship systems for each of these languages based solely on differences in data input. Further, the model learns these kinship systems with fairly few data points, on average between $30 - 50$ data points per word learned. We discuss the differences between this model's predicted acquisition order and children's empirical order for English in the Order of Acquisition section. Unfortunately, we could not find empirical data for the order of acquisition of Pukapukan, Turkish and Yanomamö kinship terms.

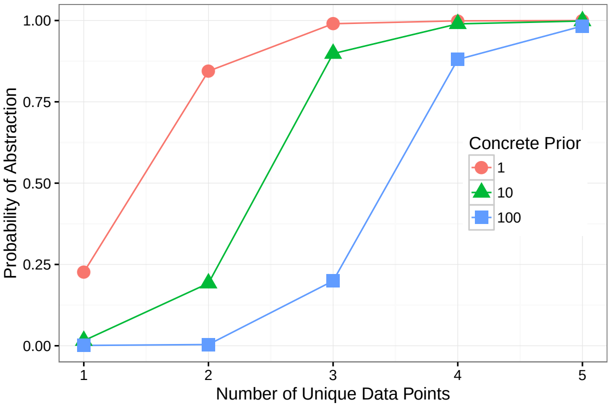

Figure 3. Probability of using abstraction as a function of unique data points at several different prior strengths for concrete reference. At higher prior values of concrete reference, the rise in the probability of abstraction is shifted to require more unique data points.

The model shows an early preference for concrete reference. Young children typically restrict their word usage to refer to particular individuals, or concrete referents, rather than draw abstractions over individuals (Clark, 1973; Kay & Anglin, 1982). This pattern naturally falls out of our model's push to explain the data when there are few unique data points, suggesting that the preference for using concrete reference is driven by the data observed rather than by inductive biases of the model. To look at the model's preference for concrete reference, we highlight a single concept, UNCLE, and focus on the first five unique data points that the model observes (see Figure 3). The x-axis in Figure 3 reflects the number of unique data points (i.e., distinct referents) for a word. The y-axis represents the probability the model uses abstraction to move away from concrete reference. With no inductive bias favoring concrete reference (red circles), the model initially favors concrete referents approximately 75% of the time. As more unique data points are observed, the model quickly switches to abstracting away from concretes referents.

This behavior is observed because at low data amounts, the best hypothesis that explains the data is a concrete referent. For example, if you only ever encounter the word *uncle* to refer to Joey the best hypothesis is to think that UNCLE just denotes Joey—regardless of how full the house is. As the model observes more data, it becomes too complicated to store all the possible referents and so the model adopts simpler rules that abstract away from the data. This movement away from concrete reference after seeing two unique referents might seem fast, given that children are often willing to provide multiple example referents before their definitions use abstraction. One possibility is that children are using kinship terms as a form of address. Therefore, their provision of referents is not a reflection of their kinship concept but of their terms of address for specific people, which extends beyond kin (e.g., *teacher*). Another possibility is that children have an inductive bias favoring concrete referents. In Figure 3, we plot the probability of abstraction when the model has a 10 : 1 (green triangles) and 100 :1 (blue squares) bias for using concrete reference as opposed to abstraction. As the bias for concrete referents increases, more unique data points need to be observed before the model favors using abstraction.

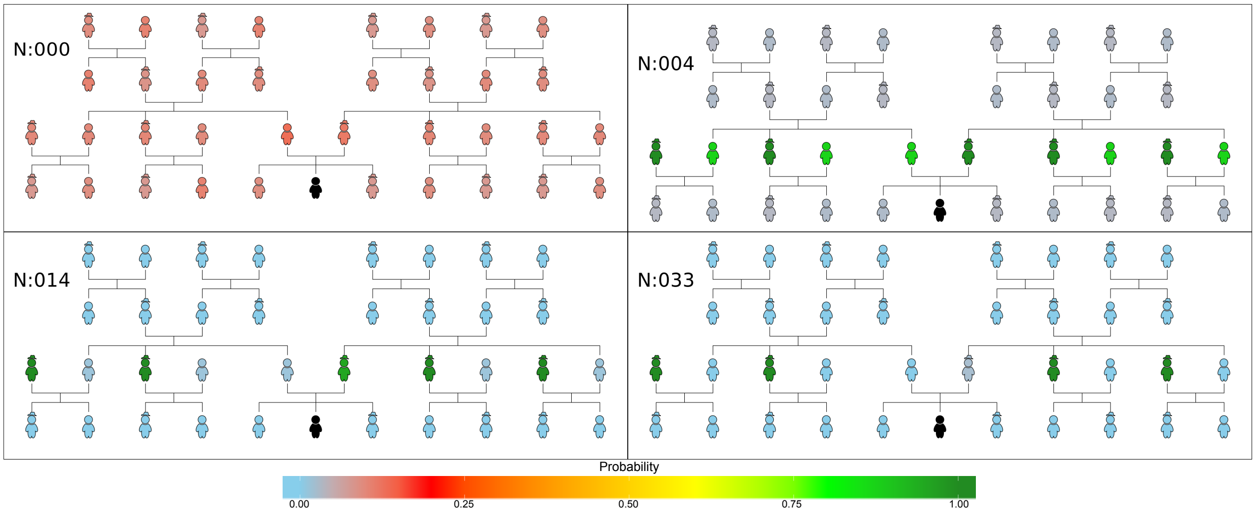

Figure 4. The posterior probability that each person on the tree is an uncle of the learner (in black) at various data amounts. Red indicates high probability and blue indicated low probability.

The model predicts over-extension as seen in children. Older children embrace abstraction; however, the rules they learn often over-extend a word to include incorrect referents (Clark, 1973; Rescorla, 1980). For example, all women might be referred to as *aunts*. Unlike under-extension, which is driven by the local data distribution at the onset of learning, over-extension is a global behavior of our model. What is interesting is that the model not only predicts over-extension but predicts specific patterns of over-extension as a function of the data it has observed and the base functions supporting the hypothesis space. For example, Figure 4 shows the model's predicted pattern of use for the term *uncle* conditioned on a learner, represented in black. At low amounts of data, everyone in the context is equally unlikely to be denoted by UNCLE. Within the first 5 data points, the model extends the term to all members of the learner's parent's generation (which is a base function). By 14 data points, the model has narrowed that down to only the males of that generation (which is the composition of two base functions). Near 33 data points, the model's extension looks very human-like; however, it is important to note that the model still needs to tease apart several different hypotheses that might make unwarranted predictions if the context was to vary. In fact, the model does not come to learn the context-invariant concept of UNCLE until around 45 data points.

Over-extension in the model falls out of the interaction between the size-principle likelihood and the base functions supporting the hypothesis space. The size principle likelihood posits that it is better to predict both observed and unseen data than to fail to predict observed data. Therefore, once the model has exhausted simple concrete hypotheses, it begins to abstract but it prefers to abstract using base functions that cast wide nets over referents–i.e., predicting many referents. The model will shift from these simple wide-reaching hypotheses to narrower hypotheses as it observes more data that can be explained better by a more complicated hypothesis. As a result, the patterns of over-extension should be predicted by base functions and compositions of base functions that increasingly approximate the true concept. We provide model predictions of the over-extension pattern for each kin term in supplemental material as an illustration. The specific patterns of over-generalization depend heavily on the base functions and more empirical data is needed to distinguish between base function inventories.

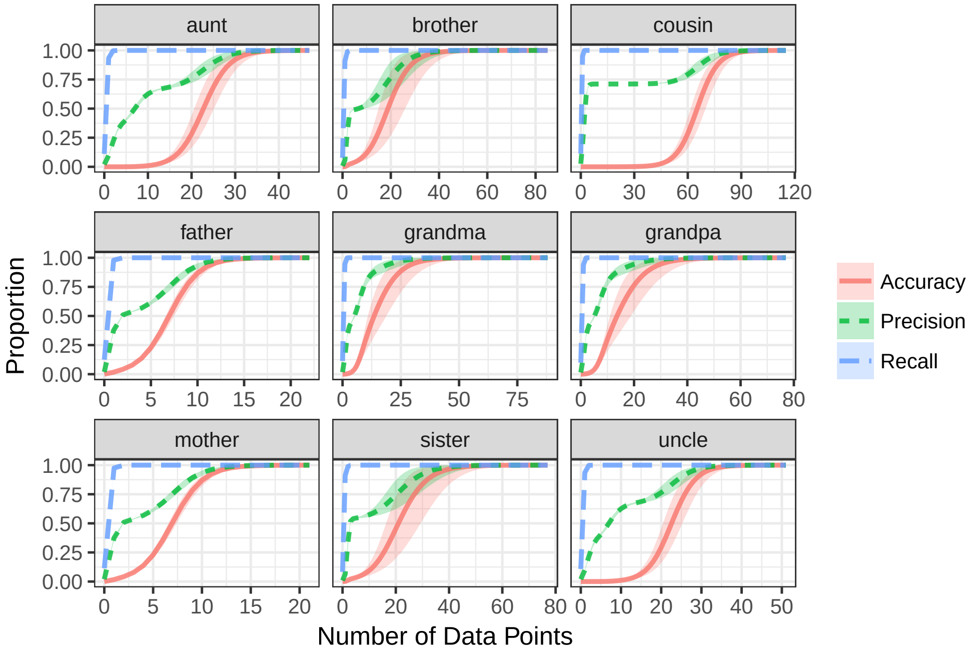

Figure 5. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Recall greater than precision is a hallmark of overgeneralization. Shaded regions represent 95% bootstrapped confidence intervals.

For a bird's eye view of over-extension in the model, we can compare the model's posterior weighted recall and precision. Recall is the probability of comprehending a word when it is used correctly. With a wide enough hypothesis, a learner will accept all of the correct uses of a word—although they will often accept incorrect uses of a word as well. Precision is the probability of producing a correct referent given the learner's current hypothesis. For example, if the learner had the correct definition of *uncle*, she would produce only and all the correct uncles and so precision would be 1.0. If the learner had a current hypothesis that over-generalized, she would produce correct uncles only a fraction of the time, even if her current hypothesis contained all of the real uncles. As a result, precision would be less than one. To visualize the presence of over-generalization, we use an F_1 score plot to compare posterior weighted precision to posterior weighted recall. Greater recall than precision is a hallmark of over-extension. Figure 5 illustrates this signature pattern of over-extension for each word in English[8].

The Characteristic-to-Defining Shift

As introduced earlier, the characteristic-to-defining shift is a prevalent pattern of children's over-extension. Young children are more likely to over-extend using characteristic features (e.g., robbers are *mean*) as opposed to defining features (e.g., robbers *take things*). While the characteristic-to-defining shift is commonly observed in concept acquisition, the process by which this occurs is unclear. One possibility is that the characteristic-to-defining shift is a stage-like transition that occurs in the representational system (Werner, 1948; Bruner, Olver, & Greenfield, 1966). For example, the shift could be explained by a transition from representing concepts wholistically—i.e., using all the features of objects, to representing concepts analytically—i.e., narrowing in specific relevant features of objects (Kemler, 1983). Neural network models of conceptual classification inherently capitalize on this idea when demonstrating a shift (e.g., Shultz, Thivierge, & Laurin, 2008). Another possibility is that there is a change in the mechanism by which one learns concepts. For example, concept learning might change from storing exemplars to constructing prototype or rule-based representations. These hypothetical changes in representation or processing might be maturational in nature, such as the development of abstraction (Piaget & Inhelder, 1969). Alternately, they may be driven by inductive inference mechanisms operating over observed data, a la rational constructivism (Xu, 2007, 2016, in press).

From the outset we can narrow down this space of theoretical hypotheses. The conceptual to defining shift is most likely a function of data, not maturation (Keil, 1983). One prediction of a maturational-shift is that at a single time-point, children should represent all words using characteristic features or defining features, whereas a data-driven shift predicts that both adults and children should have more exemplar-based representations in unfamiliar domains, and more rule-based representations in familiar domains. The former does not explain children's behavior—children seem to possess characteristic representations and defining representations of different words at a single time point. The prediction of the latter—that individuals have more exemplar-based representations in unfamiliar domains and more rule-based representations in familiar domains, is observed in children (Chi, 1985) and in adults (Chi,

[8] Appendix B contains F_1 score plots for every language and context simulated in this paper.

Feltovich, & Glaser, 1981).

All of the aforementioned explanations for the characteristic-to-defining shift require a discrete shift in representation or process. However, it is unclear whether a representational or mechanistic shift is entirely warranted. To date, no model has tested whether a characteristic-to-defining shift could be a natural by-product of the continuous data-driven construction of concepts. Here, we illustrate that the characteristic-to-defining shift could manifest even without discrete changes in representation, processing or abstraction ability. Under our model, the characteristic-to-defining shift is a consequence of incremental learning within certain learning contexts, similar to conceptual garden-pathing (Thaker, Tenenbaum, & Gershman, 2017).

We expect our model to demonstrate a characteristic-to-defining shift only if the characteristic features of the people in the context are informative but imperfect in their ability to capture the underlying concept (by denoting the proper referents). If the characteristic features accurately capture a concept, the model should never shift from favoring characteristic hypotheses to defining hypotheses. On the contrary, if the characteristic features are uninformative, and thus poor at capturing a concept, our model should favor defining hypotheses, predicting either no shift or an implausibly rapid shift from characteristic to defining hypotheses. The extent to which individual features apply beyond the learner's family tree context will also influence it's utility in explaining the shift. As a result, the feature landscape across contexts could influence the timing of both the shift and acquisition of the term. For example, if characteristic features explain the learner's data equally as well as defining features, a learner would require more disambiguating data points before learning the term than if they hadn't considered any characteristic hypotheses. Similarly, if the characteristic features that best explain family relations on a learner's own tree apply broadly to individuals outside the family context, the features are not informative enough and the characteristic-to-defining shift should occur sooner. Therefore, it is crucial that we collect data about the characteristic and logical relationships of real people to test if natural data will contain features within the range of informativity that will show a characteristic-to-defining shift.

START $\xrightarrow{1}$ SET	FSET $\xrightarrow{1}$ union(FSET,FSET)	FSET $\xrightarrow{1}$ intersection(FSET,FSET)	FSET $\xrightarrow{1}$ feature(VALUE)
START $\xrightarrow{1}$ FSET	FSET $\xrightarrow{1}$ complement(FSET)	FSET $\xrightarrow{1}$ difference(FSET,FSET)	VALUE $\xrightarrow{1}$ {Yes|No}

Table 3

Additional rules for the PCFG in Table 1. Now, each hypothesis starts with a START symbol.

We asked informants to provide us with information about their family trees. Four informants, who were blind to the experiment, drew their family tree, ranked each family member in terms of how frequently they interacted with them as a child (see Figure 6), and provided ten one-word adjectives for

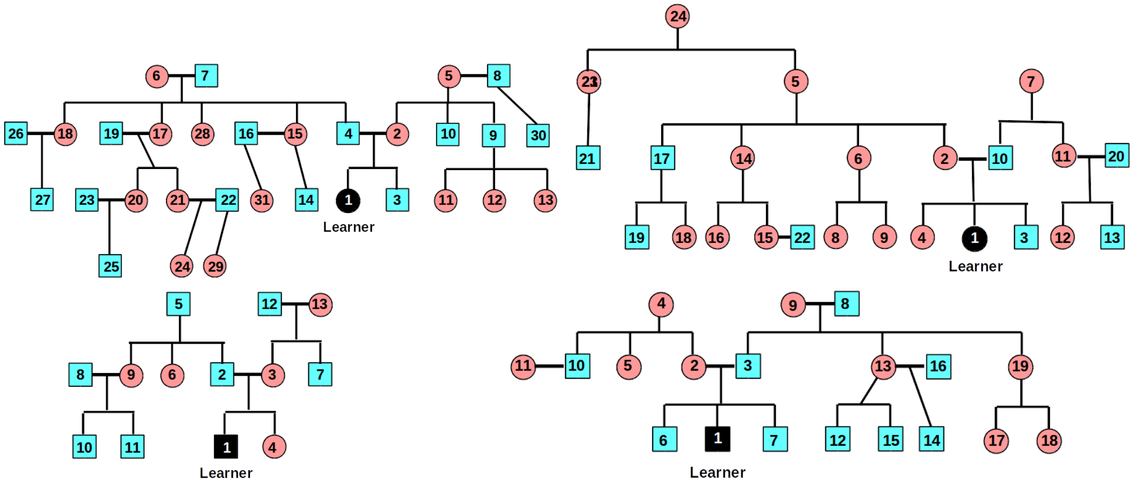

Figure 6. Distance-ranked family trees from informants. Circles represent females; squares males. Bold lateral lines denote spousal relationships. Informant 1 (top left) provided 107 unique features; Informant 2 (top right) 88; Informant 3 (bottom left) 92; and Informant 4, 59.

each family member. For each informant, the unique adjectives were used to construct a binary adjective by family member feature matrix (e.g., Figure 7). Each informant was presented with the feature matrix and asked to indicate if each feature applied to each family member. Informants made a response to every cell of the matrix: zero if the feature did not apply; one if the feature did apply. The informants provided between 59–107 ($M = 86.5$) unique features including both experiential features (e.g., *strict*) and perceptually observable features (e.g., *blonde*)[9]. We used these features to augment the hypothesis space with the rules in Table 3. One limitation of our design is that across feature matrices there was no requirement for shared features. In our matrices, there is little overlap in the solicited features, which prevents us from simulating data for a learner from other contexts. The main consequence for our analysis is that we can only predict the upper limit for the number of data points required to observe a shift as features applying more broadly or incorrectly across contexts would hasten the shift. Features applying less broadly or correctly across contexts would not introduce a bias.

The informant provided contexts are smaller/sparser than the context used in our previous analyses (Figure 1). Consequently, the types of data points the model is given in our informant analyses are restricted to a subspace of all possible types of data points, which could impede learning. The model could accommodate for this limitation by sampling across multiple contexts; however, this is computationally expensive to do for each of our informants. For computational efficiency, we only sample data for each informant within their context, which does not influence our ability to observe a characteristic-to-defining

[9] All family trees, feature matrices and code can be found at `https://github.com/MollicaF/LogicalWordLearning`

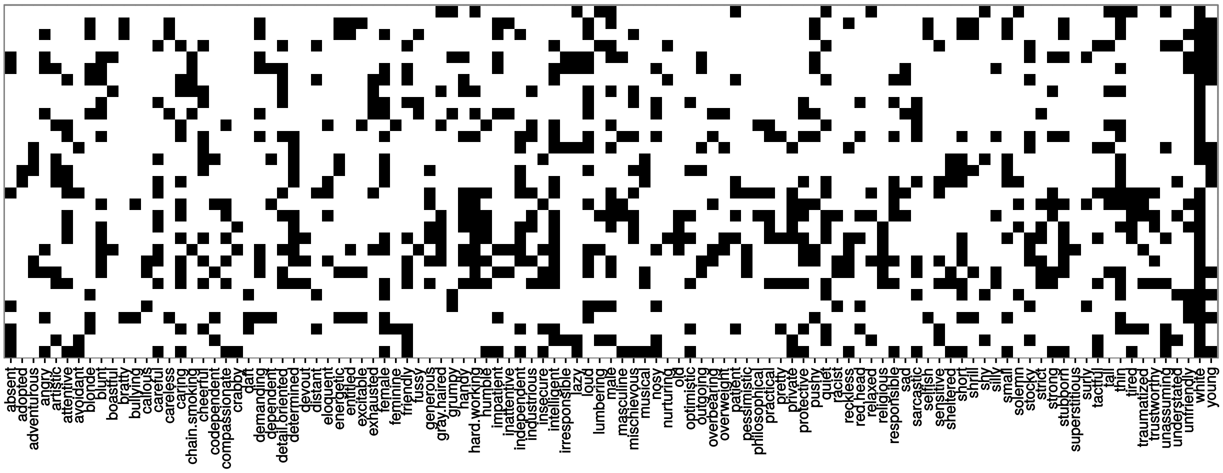

Figure 7. Feature matrix (adjective by family member) supplied by Informant One.

shift. That being said, the impoverished data/context sometimes prohibits the model from learn the conventionally-aligned upon extension of a kinship term. Nonetheless, the model does always learn a program that selects the individuals consistent with the observed data. In Appendix B, we provide F_1 plots for all informants and English kinship terms, and discuss the situations in which the model does not learn the “correct” concept for a kin term. Our failure to learn all terms from these simulations suggest that egocentric kinship data is not always sufficient for learning kinship terms.

To visualize the characteristic-to-defining shift (Figure 8), we plot the posterior probability of entertaining either a characteristic or defining hypothesis (y-axis) as a function of the amount of data observed (x-axis). For all of the words[10], we observe the characteristic-to-defining shift–i.e., the probability of entertaining a characteristic hypothesis is initially greater than the probability of entertaining a defining hypothesis. This means that a simple conceptual learning model shows a characteristic-to-defining shift purely due to the learning context–i..e, realistic data about logical relations and characteristic features. As these graphs average over the exact data points a learner observes, they hide the early preference for concrete referents; however, when plotted in terms of unique data points the early preference for concrete referents holds.

To further illustrate the model's explanation of the characteristic-to-defining shift, we have replicated the table from Mollica, Wade, and Piantadosi (2017) as Table 4, which contains the three most likely hypotheses at different data amounts for Informant One's simulated learning of GRANDMA. Recall from the Model Outcomes that before seeing data, the model prefers simpler hypotheses that tend to over-extend. As the model sees more data points, the broad over-extensions narrows to better approximate

[10] Informant 2 has no grandpa relations in their family tree context.

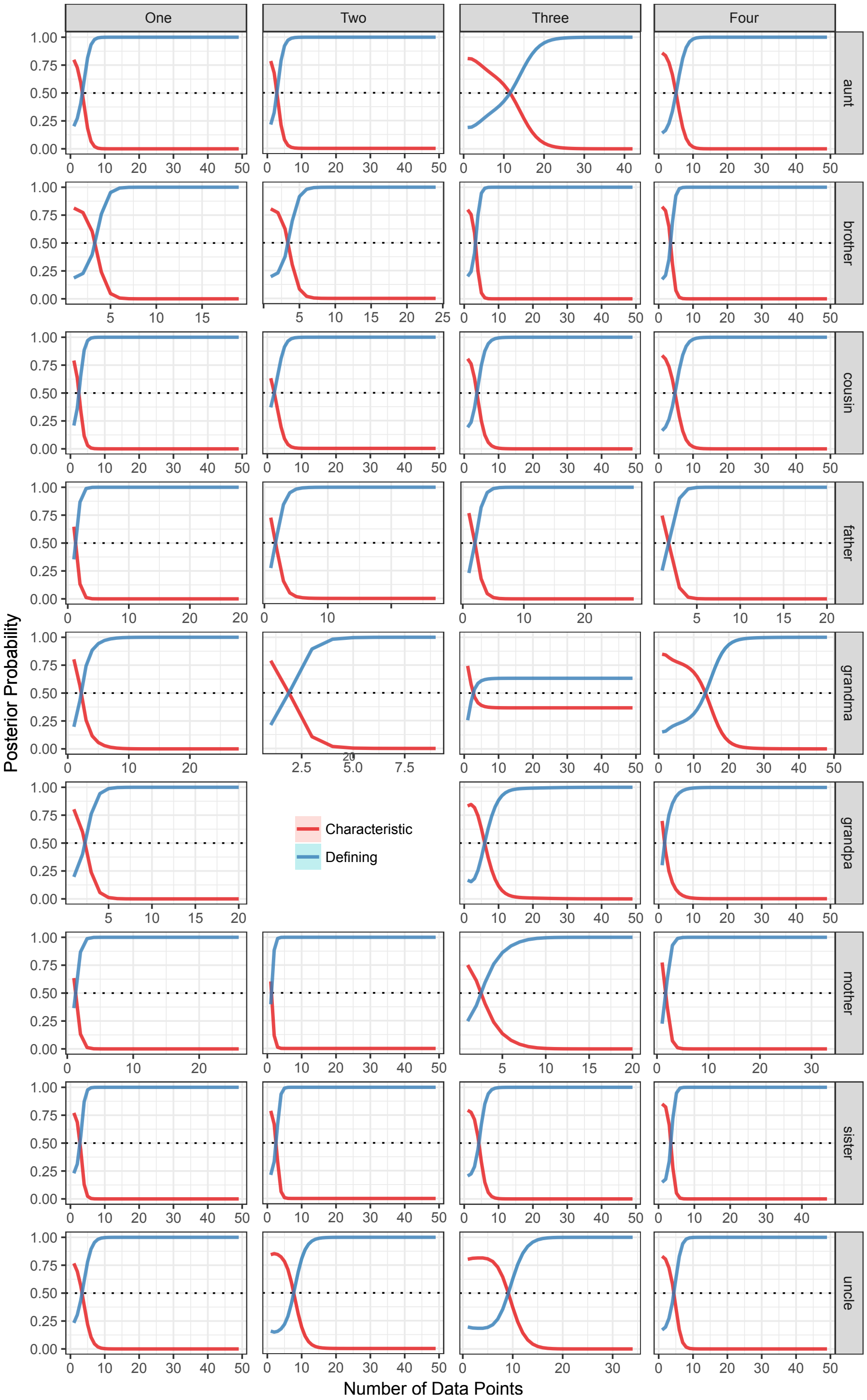

Figure 8. Average posterior probability of using a characteristic or a defining hypothesis (y-axis) as a function of the amount of data observed (x-axis) for words (rows) and informants (columns). Shaded regions reflect 95% bootstrapped confidence intervals. For all words, there is a characteristic-to-defining shift.

	Hypothesis	**Posterior Probability**
Before seeing data	X (i.e., the speaker)	0.354
	male(X)	0.006
	complement(X)	0.006
After seeing 3 data points	outgoing	0.283
	nosy	0.283
	small	0.084
One data point after shift	parents(parents(X))	0.289
	female(parents(parents(X)))	0.268
	outgoing	0.219

Table 4

Best hypotheses for Informant One learning GRANDMA *at three different time points.*

the data. This is present in Table 4 as after seeing 3 data points, the extensions narrow from, for example, all females in the context to the outgoing individuals in the context, which include both of our informant's grandmas as well as an aunt and a cousin. Importantly, the hypotheses that are favored after three data points are characteristic in nature yet imperfect in representing the concept. At one data point after the shift (i.e., the 13th data point), the most likely hypothesis still over-extends (in Table 4 by including grandpas) and is defining in nature; however, there still is mass on characteristic hypotheses. At the model observes more data, the expected extensions will continue to narrow until the correct concept for GRANDMA is the most probable.

It's important to note that our model does not have a discrete change in processing or representation as appealed to by previous research (e.g., Kemler, 1983). Additionally, our model had access to abstraction from the outset of learning. Recall from Model Outcomes that without a bias promoting concrete referents, the model without characteristic features had a 25% chance of using abstraction after only observing a single data point (Figure 3). Therefore, Piaget and Inhelder (1969)'s explanation, that the characteristic-to-defining shift reflects the development of abstraction, is not supported. With a precise, formal model of conceptual development like ours, it is possible to demonstrate that a rational learner would still undergo a characteristic-to-defining shift even if they had perfect access to the data and the ability to abstract from the outset of learning.

Compared to previous accounts of the characteristic-to-defining shift, our model proposes a new explanation: characteristic features are useful because they are simple and explain children's initial data well. As children observe more data, children can justify more complex defining hypotheses if and when characteristic features fail to explain the data. If the characteristic features perfectly explain the data, children should never switch to defining hypotheses. Perhaps this is why the characteristic-to-defining shift is only observed in some conceptual domains and absent in others. For example, even adults are hard pressed to describe concepts like ART in defining features.

Order of Acquisition: Simplicity and Data Distributions

The extent to which simplicity, as opposed to experience, drives the order of acquisition of kinship terms is an open theoretical question. Previous research has found that American children tend to acquire kinship terms in a specific order: mother/father, brother/sister, grandpa/grandma, aunt/uncle and cousin. Haviland and Clark (1974) first explained this in terms of simplicity, measured as the number of predicates in first order logic required to define the kinship term. They later revised their account to additionally penalize reusing the same relational predicate (e.g., [X PARENT A][A PARENT Y] is more complicated than [X PARENT A][A CHILD Y]). Other researchers have argued that data and the environment drive the order of kinship term acquisition. Benson and Anglin (1987) had parents rank order how frequently children spend time with, hear about or talk about twelve different kinship terms. They found that children's experience with different kinship relations correlated with their observed order of acquisition. In our model, we can directly pit experience against simplicity and evaluate these theoretical hypotheses to determine if simplicity or experience drive the order of acquisition.

To evaluate the role of simplicity and experience in determining the order of acquisition, we implement the model with different inductive biases and in different environments. For each model implementation, we simulate 1000 data sets from the tree in Figure 1 and run the learning model with only the base primitives to measure the probability that kinship terms are acquired in a specific order[11]. Figure 9 illustrates four patterns that we might see with these simulations: an accurate and reliable order of acquisition (top left panel), an inaccurate, reliable order (top right), an accurate, unreliable order (bottom left) and an inaccurate, unreliable order (bottom left). In each panel, the x-axis reflects the ordinal position in which words were learned. The fill reflects the probability that a word was acquired at that time. If the order of acquisition is reliable, there should be only one probable word acquired at each ordinal position (top panels of Figure 9). Whereas, if the order of acquisition is unreliable, there should be several probable words at each ordinal position (bottom panels of Figure 9).

Starting with *simplicity*, our initial prior distribution over hypotheses (i.e., the PCFG in Table 1)

[11] We return to the simulated tree for practical convenience and because the sparseness of the solicited trees lead to incomplete learning of kinship terms (see Appendix B). While the presence of characteristic features has the potential to influence the order of acquisition, our analyses in Figure 8 suggests the shift would occur before any of the terms are learned and, thus, have little to no effect on the order of acquisition.

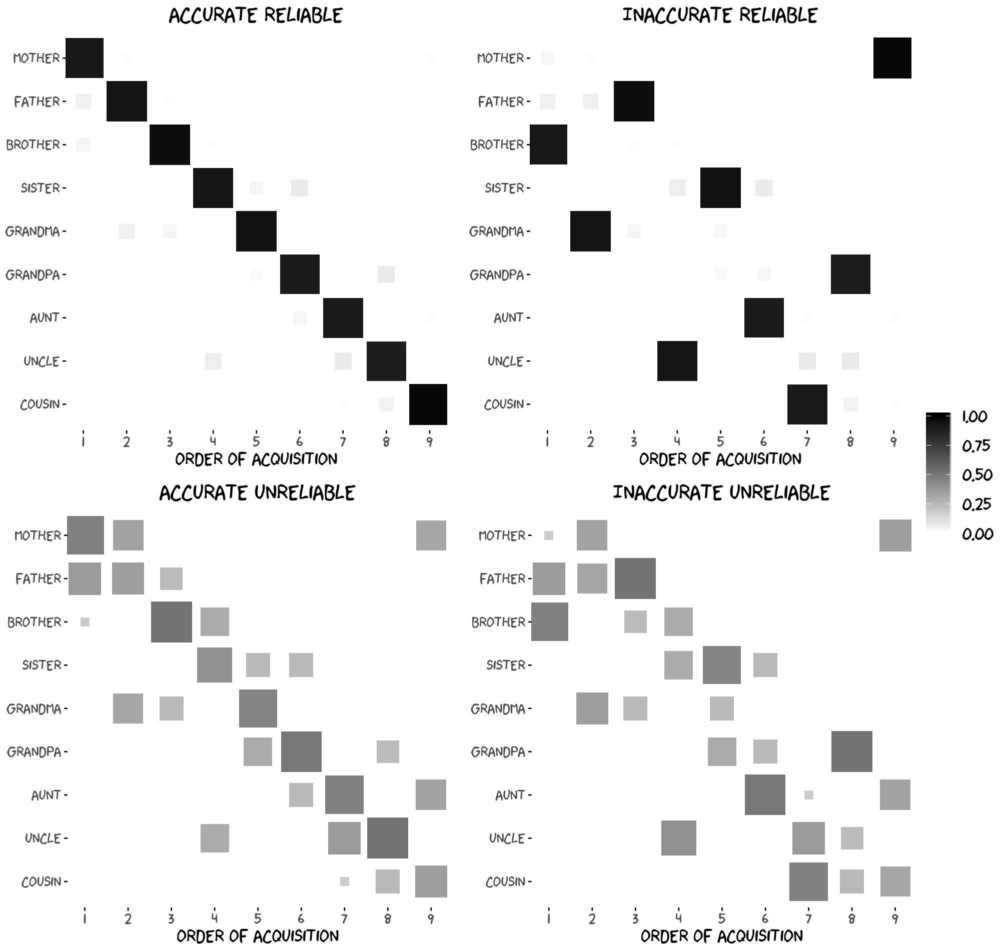

Figure 9. Possible patterns of order of acquisition. The x-axis reflects the ordinal position of acquisition. The y-axis represents each word. The tiles are filled according to the probability of acquisition. Words that have zero probability at a given ordinal position are omitted.

mostly aligns with Haviland and Clark (1974)'s original formulation of simplicity, as seen in Table 5. If data comes at a uniform rate for each word and the likelihood of a data point across words was equal, we would expect to recover this order of acquisition. However, CHILDES frequencies suggest that the frequency distribution for kinship terms is not uniform (MacWhinney, 2000) and under the size-principle, the likelihood of a data point is not equal across words in this context. The top left panel of Figure 10 shows the order of acquisition for the model given 1000 different data sets from the environmental distribution based on CHILDES frequencies and our simplicity prior. As expected, the model does not predict the empirical order of acquisition. Instead, the model is mainly uncertain about the order.

One possibility for this pattern is that CHILDES frequency estimates are not representative of children's actual input. CHILDES frequency estimates differ from the surveys of Benson and Anglin (1987)

Empirical Order	Word	Original H&C Order & Formalization	Log Prior	CHILDES Freq.
1	*mother*	Level I: [X parent Y][female]	-9.457	6812
1	*father*	Level I: [X parent Y][male]	-9.457	3605
2	*brother*	Level III: [X child A][A parent Y][male]	-13.146	41
2	*sister*	Level III: [X child A][A parent Y][female]	-13.146	89
3	*grandma*	Level II: [X parent A][A parent Y][female]	-13.146	526
3	*grandpa*	Level II: [X parent A][A parent Y][male]	-13.146	199
4	*aunt*	Level IV: [X sib A][A parent Y][female]	-19.320	97
4	*uncle*	Level IV: [X sib A][A parent Y][male]	-19.320	68
4	*cousin*	Level IV: [X child A][A sib B][B parent Y]	-18.627	14

Table 5

Complexity in terms of Haviland and Clark (1974) aligns with the prior probability of our model. Contrary to Benson and Anglin (1987)'s survey, CHILDES frequencies do not align with order of acquisition.

and a larger corpus analysis of kinship term use across Indo-European languages (Racz & Jordan, 2017), which finds that frequency decreases as genealogical distance increases[12]. As a larger point, children do not utilize every instance of a word in their environment as an effective learning instance and frequency is not strongly correlated to data usage for early word learning (Mollica & Piantadosi, 2017). Further, there is evidence to suggest that children filter their input (Perkins, Feldman, & Lidz, 2017; Kidd, Piantadosi, & Aslin, 2012).

To account for the discrepancy between environmental input and the latent distribution of effective learning instances utilized by a learner, we operationally define *experience* according to Benson and Anglin (1987)'s surveys: children are more likely to be spoken to by people closer to them; and children are more likely to hear about people who are closer to them. There are two ways in which these intuitions can be implemented in the model: through assumptions about the learner's inductive biases (i.e., in the likelihood), and through assumptions about the environment (i.e., the data distribution). We can add these assumptions to the learner's inductive biases by adopting a weighted size principle likelihood, or Zipfian likelihood. Similarly, we can add these assumptions to the environment by sampling data from two Zifpian distributions (see methods for details).

For both implementations of these assumptions, the strength of the bias is modulated by the Zipfian exponent s. When $s = 0$, the data are randomly generated–i.e., no bias, and the likelihood is equivalent to a size principle likelihood. When $s \sim 1$, the environment is biased to an extent consistent with the distribution of words in English, and the learner expects to see data points reflecting this bias. When $s > 1$, the environment is heavily biased with some "black sheep" family members almost never spoken

[12] Although, Racz and Jordan (2017) did not include grandparents in their analysis.

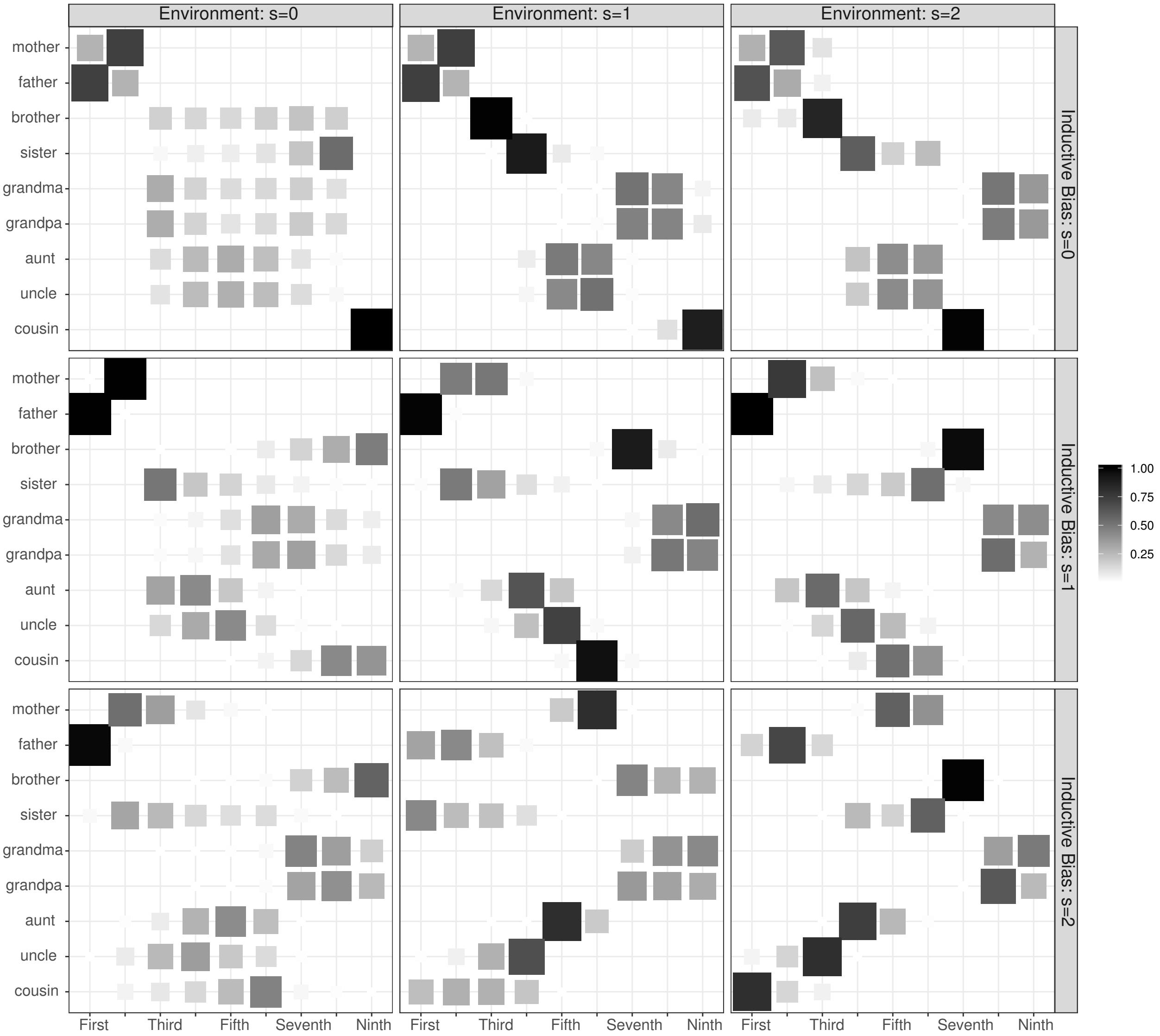

Figure 10. Simulations of the order of acquisition of kinship terms as a function of changes in environmental data distributions and the inductive biases of the learner. The strength of these biases are reflected in the s parameter with $s = 0$ reflecting no Zipfian bias. A tiny amount of random noise was added to probabilities in each simulation to settle ties.

about. Similarly, the learner does not expect to see these “black sheep” family members and discounts data including them. For simulation purposes, we assigned distances to family members loosely based on Euclidean distance to the learner in the tree context (see Figure 1).

In Figure 10, we systematically vary the environment, via the Zipfian exponent of the data distribution, and the inductive biases of the learner, via the Zipfian exponent of the likelihood function. In an unbiased environment (left column), the order of acquisition is relatively inconsistent, suggesting that order highly varies with learning data. The order of acquisition is most consistent when there is a biased environment and the bias does not greatly diverge from the learner's inductive biases. The order most closely matches the empirical order of acquisition when the environment is more biased than the learner's inductive bias (i.e., Inductive Bias $s = 0$ and Environment $s = 1$), reflecting naturalistic environments where the Zipfian exponent is ~ 1 and an unweighted size principle likelihood. The discrepancies between empirical order of acquisition and our predictions can be explained by our a-priori assignment of distances. If aunt/uncles were further from the learner than grandparents, we would expect grandparents to be acquired earlier. Differences between concepts of the same complexity (e.g., GRANDMA and GRANDPA) are slightly influenced by ties such that the alphabetical order appears dominant in Figure 10 where there is likely no bias. Importantly, under this Zipfian environmental distribution the model still shows under-extension, over-generalization and the characteristic-to-defining shift (Mollica et al., 2017). To be clear, both a simplicity bias and a Zipfian environmental distribution are required to successfully predict children's order of acquisition of kinship terms.

Our simulation analyses suggest that a latent Zipfian environmental distribution of learning data is more important that an inductive bias to expect to see certain relatives infrequently or an inductive bias for simplicity *alone*. That being said, our analysis of CHILDES word frequencies is inconsistent with this latent Zipfian distribution. How do children decide which input is useful for learning? There are multiple factors that potentially influence this filter, including the rate of metaphorical use of kinship terms, the child's ability to resolve the deixis involved in an instance of kinship term use (e.g., kinship terms are used with genitives–*your daddy is coming home*, and altercentrically–*daddy is coming home*, which involves selecting a perspective with which to represent the relation) and the utility of genealogical kinship relations over the lifespan (e.g., to young children kinship might just be an address system; whereas, genealogical relations are of more use to older children in the context of expanding their family). As mentioned earlier, our model would treat these data points as noise and still learns kinship terms under significant noise (see Appendix A). Further research is needed on how exactly children filter their linguistic input.

Discussion

By framing concept induction as logical program induction, we have demonstrated how two first principles—i.e, simplicity and the size principle, predict several of the empirical phenomena seen in kinship term acquisition. Specifically, an ideal learner model incorporating these principles learns the kinship system consistent with its input like children, offering a cross-linguistic proof of learnability. The trade-off between simplicity and the size principle drives the model to predict both an early preference for concrete reference and patterns of over-generalization consistent with children's behavior, including the characteristic-to-defining shift. Additionally, our model provides a novel explanation for the

Empirical Behavior	**Model Explanation**	**Behavioral Predictions**
Cross-linguistic learnability	Inductive learning	The number of data points before acquisition
Under-extension	Local data distribution	The number of data points before abstraction.
Over-generalization	Trade-off between prior and likelihood	The pattern of generalization at each data amount
Characteristic-to-defining shift	Learning context	The presence of and the number of data points before the shift
Order of Acquisition	Environmental experience	The order of acquisition and number of data points before each term is acquired

Table 6

Summary of the empirical behavior, how the model explains this behavior and the behavioral predictions to be generated by the model.

characteristic-to-defining shift seen in children's early understanding of words, highlighting the role of the learning context instead of proposing discrete changes in representation and processing. Lastly, the model has addressed open theoretical questions about the forces driving the order of acquisition of kinship terms in English.

Table 6 outlines each behavioral phenomena we attempted to explain, the components of the model that explain that phenomena and the behavioral predictions from the model. There are two ways in which the behavioral predictions of our computational model can be used. First, experiments can be designed to directly assess components of the model, and the learning environment. For example, the children's patterns of generalization could be used in the tradition of componential analysis to empirically ground the primitive functions used by children. Similarly, assumptions about how children use data (i.e., the likelihood function) and the inductive biases they bring to the learning task make different predictions for patterns of generalization and the timing of those behaviors. The model also makes predictions for if and when a learning context should result in a characteristic to defining shift. Second, this model can be used as a baseline or normative model for comparison against other theories of conceptual learning and for the development of theories of related processes. For example, this model shows how a learner should behave if their goal was learning the structure in the world; however, it's possible that learners are not trying to learn the structure in the world, but instead the conventions of lexical production through linguistic structure. Comparing the predictions of our model with those of formal models built to learn from linguistic structure would give us leverage to tell when and to what extent children are learning from world structure or through linguistic structure. Additionally, the model makes predictions of how children's competence should change as a function of data, which has the potential to aid the construction of theoretical models of pragmatic and retrieval processing in children's early word use, theoretical models of children's exploration and information extraction, and theoretical models of the other affordances of children's concepts (e.g., property induction/generalization).

It is important to highlight several links between this model approach and past approaches, which may be connected more formally in future research. First, the model framework is compatible with similarity based approaches to early concept acquisition. For example, a program could capture first and higher order correlations between features. While an individual program is currently implemented as deterministic in terms of referents, the posterior weighting of hypotheses allows for probabilistic interpretation. It would also be possible to extend the individual hypotheses to themselves be probabilistic in nature (see Church program; Goodman et al., 2015).

Second, the model framework is amenable to theory-theory approaches (e.g., Gopnik & Wellman, 2012) in several ways. For example, this framework is compatible with the idea of constructing *overhypotheses* from the data, which is a form of non-parametric structure learning in which higher level consistencies with the data are then given independent explanatory power (Kemp, Perfors, & Tenenbaum, 2007; Perfors, Navarro, & Tenenbaum, submitted). Learning higher level constraints on which hypotheses are more likely has the ability to fundamentally change the predicted pattern of behavior and influence future learning problems. Similarly, structures can be learned simultaneously from the same data and then be incorporated into the model.

Third, the model framework could incorporate several types of reuse and recursion (for one possibility see O'Donnell, 2015), providing a formal link to analogical transfer. For example, you can learn a specific function composition that is useful across many different hypotheses and many different learning problems. Alternatively, once you successfully learn a program you can use that program as a function in another program. Preliminary evidence suggests humans do both (Cheyette & Piantadosi, 2017). That being said, our reported model does not permit recursion and still matches several important developmental phenomena in kinship term acquisition. Future research will be needed to assess when and how recursion might be used in conceptual development.

Lastly and with regard to kinship, future models should shift focus from learning generalizable relational functions over trees grounded in genealogical structure, to learning both the component and parallel structures supporting kinship—i.e., how kinship terms (sometimes simultaneously) map to address, sociological and attitudinal structures. For example, it's easy to imagine a child construing *uncle* in *Uncle Ben* as a term of address like *doctor* in *Doctor Octavius*. Similarly, kin terms can be used to express an attitude toward an individual. For example, calling an individual a *grandpa* because they go to sleep and wake up early. Future implementations of models in our could map kinship terms to different structure or simultaneously learning multiple mappings. These models have the potential to tease apart the interactions between these structures and behavior and uncover the underlying representations for kinship by generating hypotheses for children's behavior and highlighting the empirical tests that will be most informative for distinguishing between alternatives. The understanding we derive from our current model learning the genealogical structure will be important for investigating how the alternative structures and the relationships over them are learned simultaneously.

Our work differs from past work in several ways. First our model is the first rational constructivist model (Xu, 2007, 2016, in press) that captures the behavioral phenomena observed in kinship learning. Beyond kinship, our model derives strong predictions for how conceptual development should unfold over time from first principles—i.e., simplicity and strong sampling. Previous research has highlighted the limitations of using children's early word use as evidence for their comprehension, arguing that performance limitations and pragmatic language use heavily influences early productions (Fremgen & Fay, 1980; L. Bloom, 1973). Having independent predictions for how conceptual knowledge unfolds over time provides leverage to further investigate these performance limitations and this type of early pragmatic reasoning. As a result, we may be able to gain insight from records of children's early word use, which is currently an under-utilized source of data.

Second, our account is a continuous account of conceptual development. There are no fundamental changes in the mechanism of learning or the representation of the hypothesis space. One noteworthy difference between previous accounts is that no change results in incommensurable theories (Carey, 2009); however, the conceptual system that the model ends on may be non-apparent given the likely initial hypotheses and the infinite space of hypotheses. From a child's perspective, their later theories may be incommensurable with the past theory because it is highly unlikely to move back to that area of the hypothesis space. As an argument against the implausibility of such a large hypothesis space[13], we provide evidence that the number of hypotheses actually worth consideration (i.e., within the top 95% posterior probability) at any given amount of data is manageable (median: 9, range:$5 - 30$)[14]. Although at this time, we do not provide a mechanism for how children might generate the hypothesis space, we do not mean to suggest that children will be considering the entire hypothesis space. Our goal in presenting this model is not to account for all conceptual change, but rather to provide both convergent evidence for accounts of conceptual development and a tool for predicting children's behavior as they come to wield adult like

[13] Although, our hypothesis space is no larger than the hypothesis space of any other learning model—including neural network approaches.

[14] The upper end of our range comes from Pukapuka, where the concepts often have multiple, simple, extensionally equivalent hypotheses

concepts.

Conclusion

We hope to impart two lessons learned from our model. First, programs are a powerful representational scheme to formalize concepts. Programs have the ability to capture both logical and graded/stochastic aspects of conceptual structure. When combined with data-driven learning techniques, programs not only capture the end state representation of concepts but provide rich behavioral predictions across the entire developmental trajectory, capturing phenomena like the characteristic-to-defining shift in a single model. A critical component of our program representation scheme is that our programs are functions of contexts, similar to Katz et al. (2008). Concept deployment and language use are heavily context-sensitive. To generalize across contexts, we need something like a program, that can operate over a given context. Additionally, generative programs have the potential to bridge the gap between the denotation, simulation and reasoning affordances of concepts.

Second, a precise formal model of conceptual development, like ours, allows us to rigorously test theories and questions developmental science has put forward without committing to often necessary data analysis assumptions. For example, fundamental questions in developmental science include: What biases or abilities (e.g., simplicity, compositionality, abstraction, recursion) must be in place for children to learn X? How much data do children need to learn X? Which types of data do children find most useful for learning X? What resource limitations must be in place to explain the developmental trajectory of X? Are cross-cultural differences or differences across populations learning X caused by different biases/abilities, different data availability or different consumption/usage of data? These questions can all be addressed within our model framework through Bayesian data analyses and model comparisons. As a result, formal models of conceptual development provide important and substantial convergent evidence and insight about developmental theories, which might not be possible or, more realistically, feasible to gather from behavioral experiments/observation alone.

Methods

Generating the Hypothesis Space

To construct a finite lexicon space appropriate for our analyses, we utilized a variety of Markov Chain Monte-Carlo methods to draw samples from the posterior distribution over lexicons at different data amounts. Our model is implemented using the Language of Thought Library for python (Piantadosi, 2014). As this is a computational level analysis, our goal is not to provide an account of the algorithms and processes behind hypothesis generation. Our goal is to describe learning as the movement of probability mass over a hypothesis space. Therefore, it is important to ensure that the finite approximation of the space that we use contains as many lexicons that are developmentally plausible as possible. Here a lexicon is a collection of hypotheses, one per kinship term. Our method of constructing a finite lexicon space had two phases. First, we searched the space of all possible lexicons, resulting in many partially correct lexicons. Across all of these lexicons, every word was learned and therefore, the learning trajectory for each word was present in the space. Nonetheless, few if any lexicons contained the correct hypothesis for all of the words. In our second phase, we mixed the hypotheses generated in the first phase to construct lexicons that contained the developmental trajectories of multiple words. A small percentage of these lexicons contained correct hypotheses for all of the words. Phase one and two combined generated too many lexicons to tractably analyse further. Therefore, we truncated the space by normalizing the lexicons and selecting the top 1000 hypotheses at various data amounts. For our main analyses, we collapse across lexicons and analyse developmental trajectories for each word independently to avoid any complications with not having a complete lexicon space. In Appendix C, we show that all results reported in the main text hold when analyses are conducted over lexicons.

To generate an initial set of hypotheses, we used the Metropolis-Hastings algorithm using tree-regeneration proposals following (Goodman et al., 2008; Piantadosi et al., 2012). For each language, we ran 16 chains at each of 25 equally spaced data amounts between 10 and 250. Due to memory limitations, we only saved the top 100 best lexicons from each chain. For English, Pukapukan and Yanomaman lexicons, each chain was run for one million steps. For Turkish, we first ran 5 chains for three million steps on a smaller lexicon—i.e., the search did not include the three words for grandparents or the word for cousin. We then ran 5 chains for three million steps on the full lexicon. Few if any lexicons resulting from this search contained the correct hypothesis for all words; however, across all lexicons the correct hypothesis for every word was learned.

In our second phase, we used Gibbs sampling to mix the hypotheses generated in the first phase, constructing lexicons that contained the developmental trajectories of multiple words. A small percentage of these lexicons contained correct hypotheses for all of the words. Phase one and two combined generated too many lexicons to tractably analyse further (around $200,000$ nine-word lexicons for English). Therefore, we truncated the space by normalizing the likelihoods and selecting the top 1000 lexicons at various data amounts favoring lower amounts (8 equally spaced intervals between 1 and 25, and 6 equal intervals between 25 and 250 data points). For the analyses presented in the main text, we marginalize over lexicons to analyse hypotheses for different kinship terms independently. As hypotheses are included in the space based on their performance at varying data amounts, we normalize the likelihood by simulating 1000 data points, computing the likelihood of each hypothesis and taking the average likelihood for each hypothesis.

Learnability, F_1 and Over-extension Analyses

To evaluate if a hypothesis $\hat{h}$ was correct, we compared the hypothesis's extension to the hand-constructed, ground truth hypothesis h for each kinship term system. We obtain the trajectories for posterior weighted accuracy, precision and recall by marginalizing over hypotheses at each data amount. For example, the posterior weighted accuracy is given by:

$$P(\hat{h} = h|d) = \sum^{\mathcal{H}} \delta_{\hat{h}h} P(h|d). \tag{4}$$

We adopt this same approach to estimate the extension probability for each referent x in a context as a function of data:

$$P(x|d) = \sum^{\mathcal{H}} P(x \in |h|)P(h|d), \tag{5}$$

where $P(x \in |h|)$ is given by:

$$P(x \in |h|) = \begin{cases} 1 \text{ if } x \in |h| \\ 0 \text{ else} \end{cases}. \tag{6}$$

Concrete Reference Analysis

As concrete reference is heavily influenced by local data distributions, we constructed a fixed data set of five unique data points for UNCLE and ran one MCMC chain $100,000$ steps for each amount of data. We collected the top 100 hypotheses from each chain to use for analysis. We operationalize abstraction as the probability the hypothesis is a function of the speaker:

$$P(r_{SET \to p} \in h) = \begin{cases} 1 \text{ if } r_{SET \to p} \in h \\ 0 \text{ else} \end{cases}. \tag{7}$$

The posterior probability of using abstraction at a given data amount is therefore:

$$P(r_{SET \to p}|d) = \sum^{\mathcal{H}} P(r_{SET \to p} \in h)P(h|d). \tag{8}$$

We manipulate the prior bias for concrete reference by changing the PCFG production probabilities given in Table 1, which influences the prior probability following Equation 2.

Characteristic-to-Defining Shift

We build the hypothesis space for characteristic and defining features separately for each informant. To gather defining hypotheses, we ran 7 chains at each of 25 equally spaced data amounts between 10 and 250 using the PCFG in Table 1 for $500,000$ steps. To gather characteristic hypotheses, we ran 7 chains at each of 25 equally spaced data amounts between 10 and 250 using the PCFG in Table 3 for $500,000$ steps. Due to memory limitations, we only saved the top 100 best lexicons from each chain. For each informant, the defining and characteristic hypotheses were concatenated to form a single finite hypothesis space. As our analyses collapsed over lexicons, we did not perform Gibbs sampling as above.

We replicate the learnability and F_1 analyses (described in Appendix B) using the same methods described above. Our analysis of the characteristic-to-defining shift is similar to our analysis of concrete referents. The posterior probability of using a characteristic hypothesis at a given data amount is

$$P(r_{FSET \rightarrow \text{feature}}|d) = \sum^{\mathcal{H}} P(r_{FSET \rightarrow \text{feature}} \in h) P(h|d), \tag{9}$$

where $P(r_{FSET \rightarrow \text{feature}} \in h)$ is:

$$P(r_{FSET \rightarrow \text{feature}} \in h) = \begin{cases} 1 \text{ if } r_{FSET \rightarrow \text{feature}} \in h \\ 0 \text{ else} \end{cases} . \tag{10}$$

Order of Acquisition Analysis

For the unweighted order of acquisition analysis, we sampled 1000 different datasets each containing 1000 data points as follows. A kinship term w is sampled from a multinomial distribution with θ values reflecting CHILDES frequencies. Given that term, a speaker-referent pair (x, p) is sampled uniformly from all possible speaker-referent pairs.

$$w \sim \text{Multinomial}(\theta) \tag{11}$$

$$(x, p) \sim \text{Uniform}(|(x.p)|) \tag{12}$$

To simulate *experience* according to Benson and Anglin (1987), we modified the likelihood 3 and the data generating process. We replaced our size principle likelihood 3 with a Zipfian likelihood:

$$P(x|h, p) = \delta_{d \in \{h\}} \alpha \frac{d_x^{-s}}{\sum_{x \in h(p)} d_x^{-s}} + (1 - \alpha) \frac{1}{|X|}, \tag{13}$$

where x is the referent, d_x is the rank distance of x from the learner, p is the speaker, X is the set of all possible referents, and s is the Zipfian exponent. This can be understood as a child expecting kinship terms to refer to people they frequently interact with as opposed to people they rarely hear about or see.

We added these assumptions to the data provided to the learner by sampling data from two Zipfian distribution. For each data point, speakers ranked closer in distance to the learner are more likely to be sampled than data from speakers ranked distant to the learner. Conditioned on a speaker and a word, valid referents ranked closer to the learner are more likely to be sampled than referents ranked distant to the learner. We implement both of these models with the same noise model used in Equation 3.

$$P(p|w) \sim \alpha \ \text{Zipf}(p|w, s) + \frac{(1 - \alpha)}{|X|} \tag{14}$$

$$P(x|p, w) \sim \alpha \ \text{Zipf}(x|w, p, s) + \frac{(1 - \alpha)}{|X|} \tag{15}$$

We assigned distances to the tree context in Figure 1 by fixing the learner as the central female in the youngest generation that had both a brother and a sister, and assigning relatives closer in Euclidean distance smaller distance values. As a result, aunts and uncles are assigned smaller distance values than grandparents, which results in learning aunt/uncle before grandparents (against the canonical order). The assignment of distance in our informant provided data suggests this relationship has great individual variability, so we refrain from making strong predictions about the order of acquisition for individual terms. Data is then sampled from Zipfian distributions as outlined in Equations 14 and 15.

For both schemes, we calculate the posterior accuracy of each hypothesis as a function of data following Equation 4 after each data point is sampled. If the posterior weighted accuracy is greater than or equal to 0.99, we mark the word as learned and record it's ordinal position. Ties were resolved alphabetically. As a result, we do not make strong predictions about order of acquisition for equally complex concepts (e.g., the ordering of MOTHER and FATHER), which often pattern alphabetically in our simulations.

References

Barrett, M. D. (1986). Early semantic representations and early word-usage. In *The development of word meaning* (pp. 39–67). Springer.

Bavin, E. L. (1991). The acquisition of warlpiri kin terms. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *1*(3), 319–344.

Benson, N. J., & Anglin, J. M. (1987). The child's knowledge of english kin terms. *First Language*, *7*(19), 41–66.

Bloom, L. (1973). *One word at a time.* Mouton The Hague.

Bloom, P. (2000). *How children learn the meanings of words* (No. Sirsi) i9780262523295). MIT press Cambridge, MA.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, *48*(4), 1156.

Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1966). *Studies in cognitive growth.* Wiley.

Burling, R. (1964). Cognition and componential analysis: God's truth or hocus-pocus? 1. *American anthropologist*, *66*(1), 20–28.

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

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Carter, A. T. (1984). The acquisition of social deixis: children's usages of 'kin'terms in maharashtra, india. *Journal of child language*, *11*(01), 179–201.

Chambers, J. C., & Tavuchis, N. (1976). Kids and kin: Children's understanding of american kin terms. *Journal of Child Language*, *3*(1), 63–80.

Chater, N., & Vitányi, P. (2007). 'ideal learning'of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

Cheyette, S. J., & Piantadosi, S. T. (2017). Knowledge transfer in a probabilistic language of thought. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 222–227).

Chi, M. T. (1985). Interactive roles of knowledge and strategies in the development of organized sorting and recall. *Thinking and learning skills*, *2*, 457–483.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121–152.

Clark, E. V. (1973). *What's in a word? on the child's acquisition of semantics in his first language.* Academic Press.

Danziger, K. (1957). The child's understanding of kinship terms: A study in the development of relational concepts. *The Journal of genetic psychology*, *91*(2), 213–232.

Deutsch, W. (1979). The conceptual impact of linguistic input*: A comparison of german family-children's and orphans' acquisition of kinship terms. *Journal of child language*, *6*(2), 313–327.

Elkind, D. (1962). Children's conceptions of brother and sister: Piaget replication study v. *The Journal of genetic psychology*, *100*(1), 129–136.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585.

Fremgen, A., & Fay, D. (1980). Overextensions in production and comprehension: A methodological clarification. *Journal of Child Language*, *7*(01), 205–211.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive development*, *14*(4), 487–513.

Gershkoff-Stowe, L. (2001). The course of children's naming errors in early word learning. *Journal of Cognition and Development*, *2*(2), 131–155.

Goodenough, W. H. (1956). Componential analysis and the study of meaning. *Language*, *32*(1), 195–216.

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.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *Concepts: New directions.* Cambridge, MA: MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, *118*(1), 110.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Graham, S. A., Namy, L. L., Gentner, D., & Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. *Journal of Experimental Child Psychology*, *107*(3), 280–290.

Greenberg, J. H. (1949). The logical analysis of kinship. *Philosophy of science*, *16*(1), 58–64.

Greenfield, P. M., & Childs, C. P. (1977). Understanding sibling concepts: A developmental study of kin terms in zinacantan. *Piagetian psychology: Cross-cultural contributions*, 335–358.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Haviland, S. E., & Clark, E. V. (1974). 'this man's father is my father's son': A study of the acquisition of english kin terms. *Journal of Child Language*, *1*(01), 23–47.

Hemmer, P., Tauber, S., & Steyvers, M. (2015). Moving beyond qualitative evaluations of bayesian models of cognition. *Psychonomic bulletin & review*, *22*(3), 614–628.

Hinton, G. E., et al. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).

Hirschfeld, L. A. (1989). Rethinking the acquisition of kinship terms. *International Journal of Behavioral Development*, *12*(4), 541–568.

Hoek, D., Ingram, D., & Gibson, D. (1986). Some possible causes of children's early word overextensions. *Journal of child language*, *13*(03), 477–494.

Huttenlocher, J. (1974). *The origins of language comprehension.* Lawrence Erlbaum.

Johnston, A. M., Johnson, S. G., Koven, M. L., & Keil, F. C. (2016). Little bayesians or little einsteins? probability and explanatory virtue in children's inferences. *Developmental Science*.

Jones, D. (2010). Human kinship, from conceptual structure to grammar. *Behavioral and Brain Sciences*, *33*(5), 367–381.

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

Kay, D. A., & Anglin, J. M. (1982). Overextension and underextension in the child's expressive and receptive speech. *Journal of Child Language*, *9*(01), 83–98.

Keil, F. C. (1983). On the emergence of semantic and conceptual distinctions. *Journal of Experimental Psychology: General*, *112*(3), 357.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development.* Cambridge, MA: MIT Press.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of verbal learning and verbal behavior*, *23*(2), 221–236.

Kemler, D. G. (1983). Exploring and reexploring issues of integrality, perceptual sensitivity, and dimensional salience. *Journal of Experimental Child Psychology*, *36*(3), 365–379.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological review*, *119*(4), 685.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Aaai* (Vol. 3, p. 5).

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527–535.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Landau, B. (1982). Will the real grandmother please stand up? the psychological reality of dual meaning representations. *Journal of Psycholinguistic Research*, *11*(1), 47–62.

LeVine, R. A., & Price-Williams, D. R. (1974). Children's kinship concepts: Cognitive development and early experience among the hausa. *Ethnology*, *13*(1), 25–44.

Lewis, M. L., & Frank, M. C. (2018). Still suspicious: the suspicious-coincidence effect revisited. *Psychological science*, 0956797618794931.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.

Lounsbury, F. G. (1956). A semantic analysis of the pawnee kinship usage. *Language*, *32*(1), 158–194.

Macaskill, A. (1981). Language acquisition and cognitive development in the acquisition of kinship terms. *British Journal of Educational Psychology*, *51*(3), 283–290.

Macaskill, A. (1982). Egocentricity in the child and its effect on the child's comprehension of kin terms. *British Journal of Psychology*, *73*(2), 305–311.

MacWhinney, B. (2000). *The childes project: The database* (Vol. 2). Psychology Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive science*, *14*(1), 57–77.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction.* Mit Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.*

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Mollica, F. (2019). *The human learning machine: Rational constructivist models of conceptual development* (Unpublished doctoral dissertation). University of Rochester.

Mollica, F., & Piantadosi, S. T. (2015). Towards semantically rich and recursive word learning models. In

Proceedings of the 37th annual meeting of the cognitive science society (pp. 1607–1612).

Mollica, F., & Piantadosi, S. T. (2017). How data drive early word learning: A cross-linguistic waiting time analysis. *Open Mind*.

Mollica, F., & Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society open science*, *6*(3), 181393.

Mollica, F., Wade, S., & Piantadosi, S. T. (2017). A rational constructivist account of the characteristic to defining shift. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Morgan, L. H. (1871). *Systems of consanguinity and affinity of the human family* (Vol. 218). Smithsonian institution.

Murdock, G. P. (1949). *Social structure.* Macmillan.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Nakao, K., & Romney, A. K. (1984). A method for testing alternative theories: An example from english kinship. *American Anthropologist*, *86*(3), 668–673.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage.* MIT Press.

Paccanaro, A., & Hinton, G. E. (2001). Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, *13*(2), 232–244.

Perfors, A. (2012). Bayesian models of cognition: what's built in after all? *Philosophy Compass*, *7*(2), 127–138.

Perfors, A., Navarro, D. J., & Tenenbaum, J. B. (submitted). Simultaneous learning of categories and classes of categories: acquiring multiple overhypotheses. *Manuscript submitted for publication.*

Pericliev, V., & Valdés-Pérez, R. E. (1998). Automatic componential analysis of kinship semantics with a proposed structural solution to the problem of multiple models. *Anthropological Linguistics*, 272–317.

Perkins, L., Feldman, N., & Lidz, J. (2017). Learning an input filter for argument structure acquisition. In *Proceedings of the 7th workshop on cognitive modeling and computational linguistics (cmcl 2017)* (pp. 11–19).

Piaget, J. (1928). *Judgment and reasoning in the child.*

Piaget, J., & Inhelder, B. (1969). *The psychology of the child.* Basic Books.

Piantadosi, S. T. (2014). *LOTlib: Learning and Inference in the Language of Thought.* available from https://github.com/piantado/LOTlib.

Piantadosi, S. T., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, *25*(1), 54–59.

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

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, *123*(4), 392.

Price-Williams, D., Hammond, O., Edgerton, C., & Walker, M. (1977). Kinship concepts among rural hawaiian children. *Piagetian psychology: Crosscultural contributions*, 296–334.

Racz, P., & Jordan, F. (2017). What explains the frequency of use in kinship terms across indo-european languages? In *Annual meeting of the european human behaviour and evolution association.*

Ragnarsdottir, H. (1999). The acquisition of kinship concepts. *Language and Thought in Development: Cross-linguistic Studies*, *26*, 73.

Read, D. W. (2001). What is kinship? In R. Feinberg & M. Ottenheimer (Eds.), *The cultural analysis of kinship: The legacy of david schneider and its implications for anthropological relativism* (p. 78-117). Urbana: University of Illinois Press.

Read, D. W. (2007). Kinship theory: A paradigm shift. *Ethnology*, 329–364.

Rescorla, L. A. (1980). Overextension in early language development. *Journal of child language*, *7*(02), 321–335.

Shultz, T. R., Thivierge, J.-P., & Laurin, K. (2008). Acquisition of concepts with characteristic and defining features. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 531–536.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Swartz, K., & Hall, A. E. (1972). Development of relational concepts and word definition in children five through eleven. *Child Development*, 239–244.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2015). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Manuscript submitted for publication.*

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). Citeseer.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20.

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.

Van Luong, H. (1986). Language, cognition, and ontogenetic development: A reexamination of piaget's premises. *Ethos*, *14*(1), 7–46.

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic bulletin & review*, 1–10.

Wallace, A. F., & Atkins, J. (1960). The meaning of kinship terms. *American Anthropologist*, *62*(1), 58–80.

Werner, H. (1948). *Comparative psychology of mental development.* Follett Pub. Co.

Wexler, K. N., & Romney, A. K. (1972). Individual variations in cognitive structures. *Multidimensional scaling: Theory and applications in the behavioral sciences*, *2*, 73–92.

Xu, F. (2007). Rational statistical inference and cognitive development. *The innate mind: Foundations and the future*, *3*, 199–215.

Xu, F. (2016). Preliminary thoughts on a rational constructivist approach to cognitive development. In *Core knowledge and conceptual change* (p. 11). Oxford University Press.

Xu, F. (in press). Towards a rational constructivist theory of cognitive development. *Psychological Review*.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental science*, *10*(3), 288–297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Supplementary Materials

Supplementary Materials can be found at `mollicaf.github.io/kinship.html`.

Appendix A

Alpha Analysis

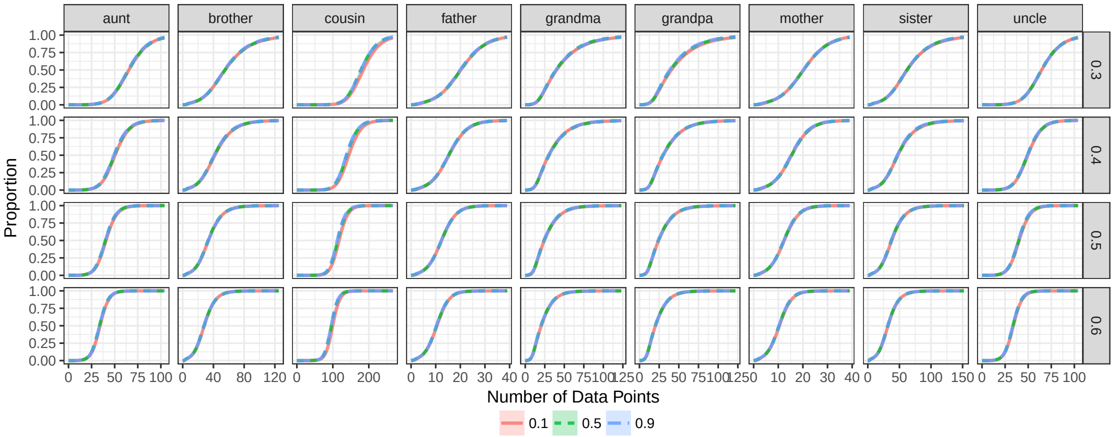

Figure A1. Posterior weighted accuracy (y-axis) as a function of data (x-axis) for models with different sampling assumptions (linetype and color) for different words (columns) and environmental reliability values (rows). The virtually invisible shaded regions reflect 3 standard errors of the mean.

Navarro et al. (2012) investigated how the reliability parameter α, which mixes between strong and weak sampling influences an inductive generalization task. They simulated environments where the data was generated to be reliable $30 - 60\%$ of the time, and checked how distinguishable a model with different sampling assumptions would be from pure strong sampling ($\alpha = 1$). They found that in the limit of data, models with reliability parameters as low as 0.1 converge to the predictions of strong sampling. We parametrically vary the reliability of the environment by simulating data with $30 - 60\%$ reliability and set our model's sampling assumptions to either 0.1, 0.5 and 0.9 to gauge whether learning in our simulations will be robust to unreliable environments and variable sampling assumptions. As can be seen in Figure A1, we find no significant differences in learning across sampling assumptions and environments.

Appendix B

F_1 Score Plots

As described in the main text, F_1 score plots are a visualization of learnability and over-generalization. Each figure in this appendix plots the posterior weighted accuracy, precision and recall (y-axis) as a function of data (x-axis). Accuracy reflects the the probability that the model has acquired the adult-like concept for that kinship term. Recall corresponds to the probability that the model will recognize a correct referent, and is given by:

$$\frac{\sum_{x \in \hat{h}} [x \in h]}{|h|}, \tag{16}$$

where x is a referent, $\hat{h}$ is the proposed hypothesis, h is the ground truth hypothesis. Precision corresponds to the probability that the model will propose a correct referent, and is given by:

$$\frac{\sum_{x \in \hat{h}} [x \in h]}{|\hat{h}|}. \tag{17}$$

When recall is greater than precision, the model is over-extending the term.

Figure B1 displays the F_1 plots for Pukapuka, Turkish and Yanomamö. As shown in the main text, the model learns the correct extension for every word. As expected, the posterior weighted recall is greater than the posterior weighted precision for every word, suggesting that the model over-extends the meaning of kinship terms. Predictions for the pattern of over-extension for each word is provided in supplemental material.

The Characteristic-to-Defining Shift

Figure B2 displays the F_1 plots for each of our informants. For all words, posterior weighted recall is greater than posterior weighted precision, consistent with over-extension of kinship words. As discussed in the main text, the model fails to learn the correct hypothesis for some words due to the impoverished input/context. That being said, the model always learns a hypothesis that is consistent with it's input. If we had provided evidence from multiple family tree contexts, we expect the model to learn the adult-like extension for all of the concepts. This suggests that having evidence from multiple families is likely an important property of the kinship data that childern use to learn their kinship terms.

In the majority of cases where the model does not acquire the correct extension, the conventional hypothesis was blocked by a hypothesis that overfit the context. For example, Informant 3 overfits for GRANDMA and Informant 4 overfits for GRANDPA because there is only one of those relations in their family tree. Hence, it is sufficient to just point to that person. Informant 2 does not learn AUNT, Informant 3 does not learn SISTER and Informant 4 does not learn COUSIN for similar reasons. In these cases, the conventional hypotheses do have some posterior probability (as evidenced in Figure B2 by non-zero Accuracy) but do not come to dominate the posterior distribution of possible hypotheses. The conventional hypotheses are blocked by hypotheses that are less complex, explain the observed data, but would not generalize properly across contexts.

Instead of overfitting, Informant 1 and 4 do not learn the conventional hypotheses for AUNT and UNCLE because there are children out of wedlock, which complicates how we have defined the conventional hypotheses. Importantly, the maximum-a-posteriori, or best, hypothesis recovered by the model actually generalizes correctly over trees without out of wedlock children. Informant 2 does not have any grandfathers in their family tree context and, therefore, the model never receives data to learn GRANDPA.

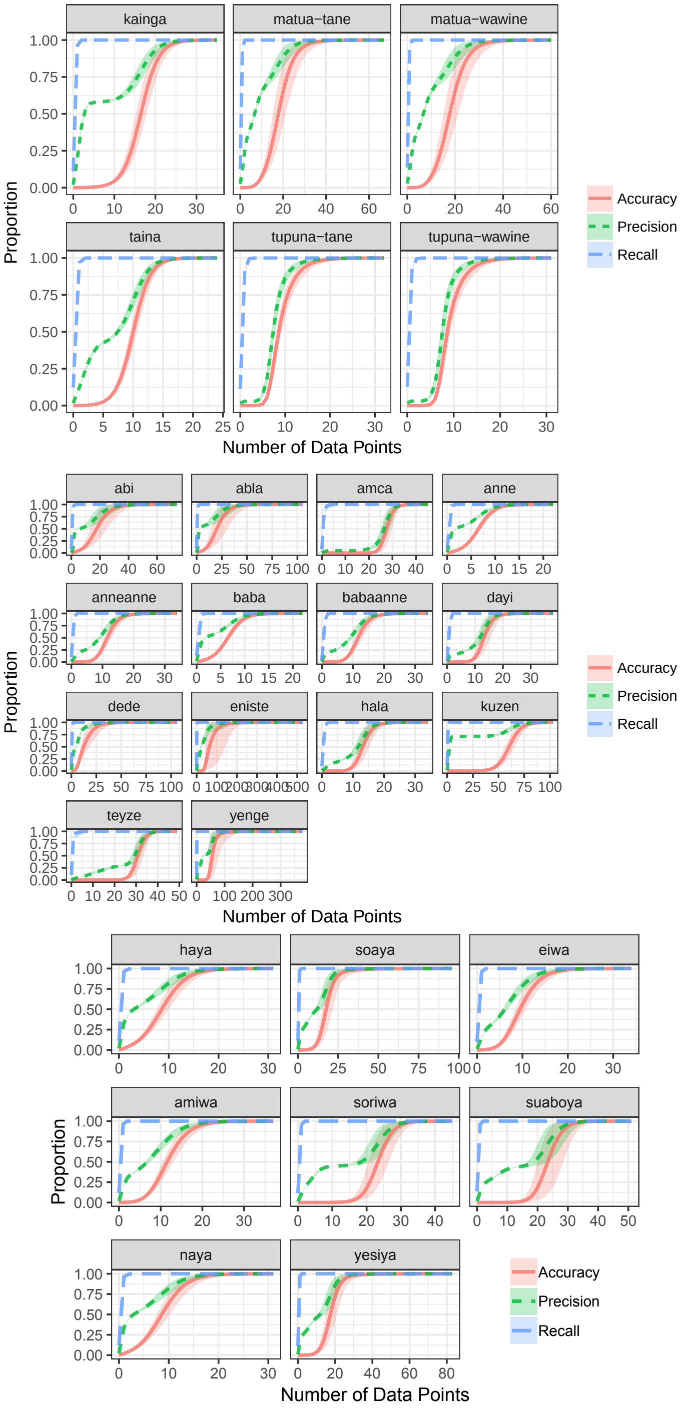

Figure B1. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

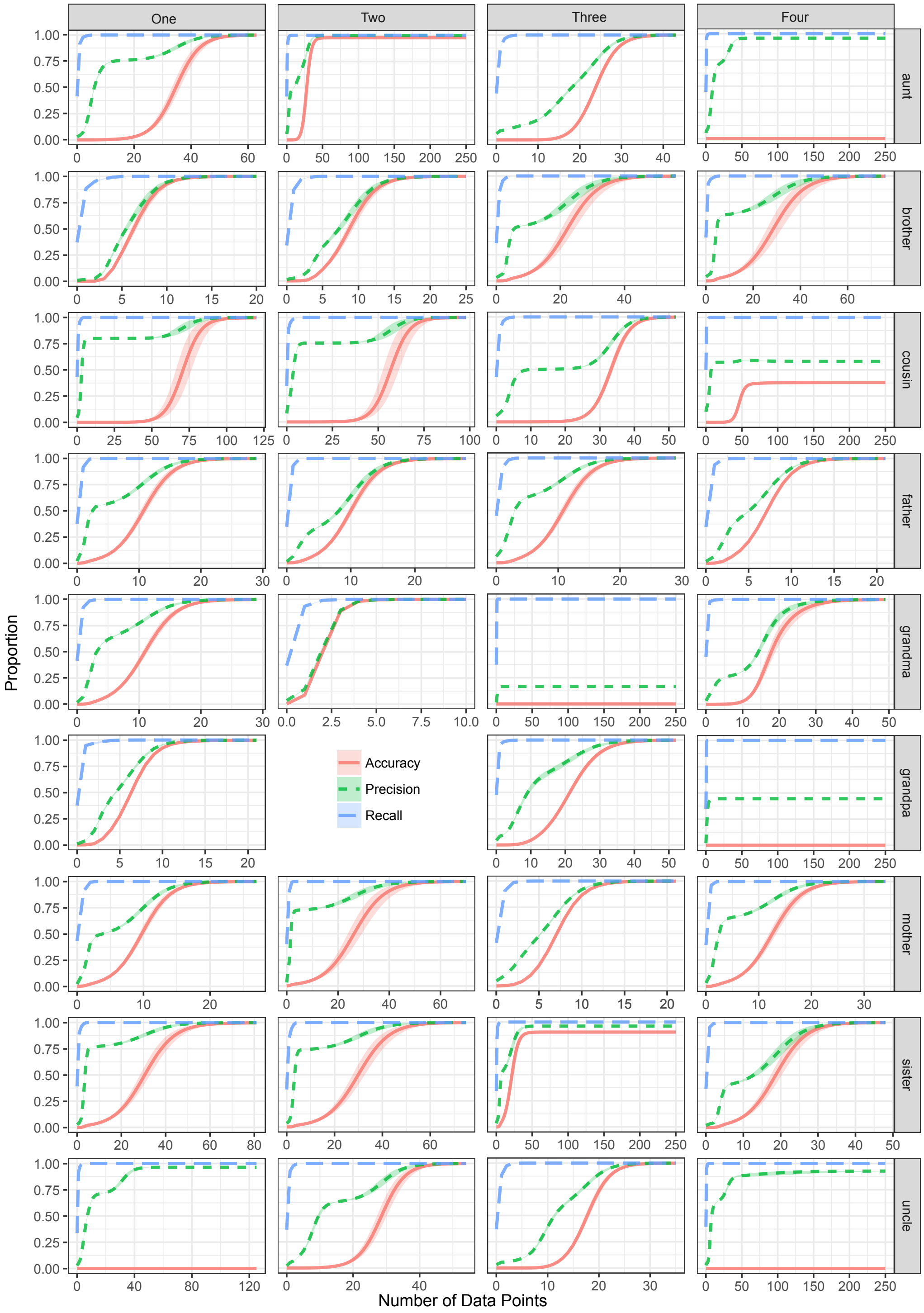

Figure B2. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

Appendix C

Learning an inter-related system

Throughout the paper, we have described a model that learns kinship terms independently of each other. One trivial way to implement learning an inter-related system would be to change the likelihood function to operate over the lexicon instead of individual words (e.g., Mollica, 2019). However, the more natural way to think of learning an inter-related system like kinship would be to allow for recursive calls. For example, a learner might use their current concept for BROTHER in their concept for UNCLE. We have implemented recursive calls in the model; however, despite multiple attempts, we were unable to construct an acceptable lexicon space to evaluate the model against developmental behavior. Without a proper finite approximation to the space of probable lexicons, there are no guarantees that any "conclusions" drawn will be robust.

One common issue with the search was finding lexicons that only learned a subset of the words after a lengthy search process. In the main text, we could easily mix lexicons using Gibbs sampling to help ensure the relevant lexicons—i.e., lexicons that contain all high probability combinations of hypotheses across the developmental trajectory, were in our finite approximation of the space. Unfortunately, recursive calls introduces dependencies between words in a lexicon, which prohibits techniques like Gibbs that rely on independence.

FATHER	male(parent(X))
MOTHER	female(parent(X))
BROTHER	child(parent(X))
SISTER	female(BROTHER(X))
UNCLE	male(BROTHER(parent(X)))
AUNT	female(BROTHER(parent(X)))
COUSIN	difference(generation0(X), BROTHER(X))

Table C1

An example local max lexicon when permitting recursive calls in the lexicon space.

Another common issue was the presence of local maxima in search. Often the model would construct a useful primitive instead of the definition of a word, which blocks that word from being acquired. Consider the example lexicon in Table C1. There are no simple proposals to BROTHER that do not negatively alter the ability of several other words in the lexicon to explain data.

Due to the search issues, we adopted a different tactic to explore the effect of recursion on kinship learning. Hypotheses with recursive calls have extensionally-equivalent hypotheses defined in terms of the base primitives. For example from Table C1, SISTER could be expressed as female(child(parent(X))). Being extensionally-equivalent, the two hypotheses have the same likelihood. The only difference on the posterior probability is in the prior. Recursive hypotheses should be simpler and thus more probable. Therefore, we can change the prior distribution over our existing hypothesis space to behave equivalently as if it was recursive. We capture the same intuitions as recursion using the Lempel-Ziv compression of the lexicon in terms of the grammar as a prior over lexicons. This prior distribution favors the reuse of specific combinations of primitives in the lexicon similar to the recursive calls in Table C1.

We found that when using a compression prior, the model predicts an inductive leap from most of the kinship terms not being properly acquired to all of the kinship terms being learned. We see this leap because the correct lexicon under the compression prior is significantly less complex than the lexicons required in search space to get you there (Figure C1). As a result, order of acquisition behavior are not predicted To remove this inductive leap, we could add a parameter that penalizes recursion (as in Piantadosi et al., 2012); however, we think that the better explanation would be through the development and integration of a more cognitively grounded notion of hypothesis generation—i.e., an algorithmic level explanation.

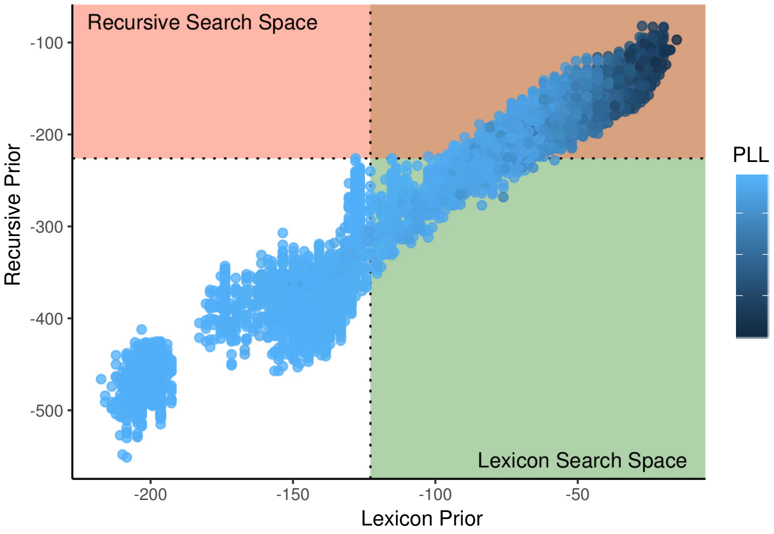

Figure C1. The English lexicons are plotted as a function of the recursive (compression) and lexicon prior. The color of each point represents the point log likelihood (PLL) of the lexicon. If the learner searched the space starting from the simplest to the most complex lexicon and terminated at the first correct lexicon, they would have to search a smaller space under a compression prior (red shade) than under a lexicon prior (green shade). Importantly, the developmental trajectory is not predicted under the recursive prior without additional assumptions about the complexity/development of recursion.