Logical word learning: The case of kinship

Francis Mollica

Steven T. Piantadosi

University of Rochester

5 March 2018

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 learning model for kinship acquisition. We demonstrate that our model can learn any kinship system consistent with its learning data using cross-linguistic data from English, Pukapuka and Turkish. 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 an epiphenomenon of our learning model in naturalistic contexts. 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 under different conceptual architectures for learning inter-related systems. 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

When you ask an adult "What is an uncle?", they often have no problem responding "An uncle is your parent's brother." This exchange conveys the responder's knowledge of UNCLE, an inherently relational concept. For relational concepts, adult definitions tend to be rule-like and they capitalize on the structure of the underlying relations. For example, the above definition for *uncle* could easily be expressed as a rule in logic: $\exists.x[\text{uncle}(x, you) \Leftarrow \exists.z[\text{parent}(z, you) \wedge \text{brother}(x, z)]]$, which capitalizes on knowledge of the relation between uncles and parents. Additionally, the above rule recursively uses a kinship term, *brother*, rather than defining the rule solely with logical operations. For example, *brother* could be replaced with $\exists.x\exists.y\exists.z\text{parent}(z, x) \wedge \text{parent}(z, y) \wedge \text{male}(x) \wedge x \neq y$. Given the complexity of the rule-like and inter-related conceptual systems of adults, learning inter-related conceptual systems is not a trivial task for a child.

When you look at how children define terms that denote relations, they often do not provide short and simple rules. 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. At young ages, children 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, Benson and Anglin (1987) reports an interview with a three year old child, who when asked "What is an uncle?" responded "Uncle Anthony." When further prompted, "Tell me everything you know about an uncle," the child offered up another concrete referent, "Uncle Henry." Slightly older children will also over-extend a term, using it to describe inappropriate but often similar referents (Clark, 1973; Rescorla, 1980). For example, a child might walk up to any man their father's age and address him as *daddy*.

This behavior is consistent with the definition Benson and Anglin (1987) reports from an interview with a five year old: "He's a man."

While these behavioral patterns are consistently observed in children's early word use[1], 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)?

In this paper, we describe a rational constructivist framework (Xu, 2007, 2016) of conceptual development formalized as logical program induction. For demonstrative purposes, we implement a model based on this framework to explain the behavioral patterns of acquisition we see in the rich, cross-linguistically varied, inter-related conceptual domain of kinship. Consequently, we provide the first computational learning model for the conceptual development supporting kinship term acquisition. We demonstrate that the model is powerful enough to learn any kinship system consistent with its input data. Crucially, we use simulations based on informant provided learning contexts to show that the patterns of children's word use, specifically over-generalization, under-extension and the characteristic-to-defining shift (Keil & Batterman, 1984), fall out naturally from framing conceptual development as program induction in naturalistic

[1]While the provided examples are all kinship related, these patterns are attested across conceptual domains.

environments, suggesting that children's early word use might be informative about conceptual development. Additionally, we examine the roles of simplicity and environmental input in determining the order of kinship term acquisition, making novel predictions for how different conceptual architectures for learning inter-related systems (e.g., lexicons) might influence learning times.

Children's Acquisition of Kinship Terms

Piaget (1928) was first interested in the logical relationships involved in kinship terms. His work was followed by a litany of interviews with children attempting to characterize children's knowledge of kinship terms (e.g., Danziger, 1957; Carter, 1984; Chambers & Tavuchis, 1976). The main conclusion from these investigations is that children, some even at nine years of age, do not demonstrate mastery of the kinship relations in their language. More theoretically motivated studies of kinship relations have attempted to explain 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. 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). 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. The extent to which simplicity and experience contribute to the order of acquisition of kinship terms is still an open question.

As alluded to earlier, interviews on kinship terms have provided some evidence for patterns of over- and under- extension in children's use of kinship terms (e.g., Benson & Anglin, 1987). While these examples provide evidence for these two patterns, there are two

limitations worth noting. First, these children are not in the age range where the typical patterns of over- and under- extension are observed; they are older. Second, the task of defining terms is a challenging task for children and we know that verbal ability increases with age. Therefore, we should take these patterns with a grain of salt as young children just 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. This makes it all the more interesting if these patterns do emerge naturally, suggesting 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.

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.

Previous computational models have approached the acquisition of kinship knowledge through a 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 analyse 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.

Our model builds off the intuitions from these past 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[2], 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 account for 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

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

probabilistic program induction (e.g., Lake, Salakhutdinov, & Tenenbaum, 2015; Piantadosi & Jacobs, 2016; Goodman, Tenenbaum, & Gerstenberg, 2015).

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; whereas, programs are defined in terms of base functions, the compositions of these functions and the relations between variables and functions. Second, placeholder structures are important in both program induction (e.g., creation of new variables or sub-routines) and conceptual development (e.g., the count list in number learning) (Carey, 2009). Third, 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, 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 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, a prior over hypothetical concepts $P(h)$ and a likelihood function $P(d|h)$ to score the hypothesis according to the data. The hypothesis space reflects the cognitive architecture supporting learning. If we imagine that the child is a scientist (Gopnik, Meltzoff, & Kuhl, 1999), what do we think their hypotheses look like? For example, hypotheses might look like first order logic, an associative network or compositional functions. The prior reflects the inductive biases that we suspect children bring to a learning task. Before seeing any data, which hypotheses do we think children are likely to generate? For example, the shape bias (Landau, Smith, & Jones, 1988) suggests that children should readily generate hypotheses linking novel words to the shape of the labeled object as opposed to some other property of the object (e.g., material or color).

The likelihood reflects how we think the data (i.e., instances of referential word use) are generated. Why are people using this word to refer to this object? The intentional model of word learning (Frank, Goodman, & Tenenbaum, 2009) postulates that speaker's intend to refer to an object in the context. Given an intention, people will choose a word in

their lexicon that they believe maps to the intended referent in the context. By modeling word learning as inferring speaker's intentions, this model has qualitatively captured many important phenomena in word learning, including cross-situational word learning, a mutual exclusivity bias, and fast mapping; however, this approach is not without limitation. The intentional model defines the lexicon as a mapping between words and objects, not words and concepts. As a result, the model does not capture how children might generalize a word to similar objects or objects of the same kind. Lewis and Frank (2013) extended this model to include concepts, noting that speakers have a choice of which concept to employ when referring to an object. Here, we adopt a consonant framework, focusing on conceptual development as the mapping between concepts and objects.

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:

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

Hypothesis Space

Constructing the hypothesis space over possible programs involves specifying base functions 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 features (generation, gender), 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 acknowledge that gender and generation are not necessarily observable; nonetheless, we assume that gender and generation can be approximated by children. Given children's knowledge of ownership (e.g., Nancekivell & Friedman, 2017), we assume that children can compute functions over speakers. Given the late timescale of children's acquisition of kinship concepts, we feel these assumptions are appropriate.

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., gender might be represented in terms of primitives that check for perceptual features) or how one could choose to augment this set at a greater level of abstraction (e.g., adding a sibling primitive). For any model of learning, the granularity 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. When reporting the results of the model, we will note the extent to which any finding is dependent on the specific primitives we have chosen. 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 $\xrightarrow{\frac{1}{37}}$ 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, Tenenbaum, Feldman, and Griffiths (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[3]. In addition to defining an infinite space, a PCFG also provides a probability distribution over the 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 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

[3]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.

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) 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 was 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 mixes 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:

$$P(d|h) = \begin{cases} \frac{1}{|h|} & \text{if } d \in |h| \\ 0 & \text{else} \end{cases}, \tag{3}$$

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:

$$P(d|h) = \frac{\alpha}{|h|} + \frac{1-\alpha}{|\mathcal{D}|}. \tag{4}$$

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.

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 "real data" to predict children's word use and to infer the inductive biases and conceptual architectures

supporting conceptual development. Unfortunately, there are no existing data sets that either spans the nine years of a single child's experience with kin and kinship terms with the required detail to fully specify this model or quantitatively measure children's kinship term use. 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 reported in the literature.

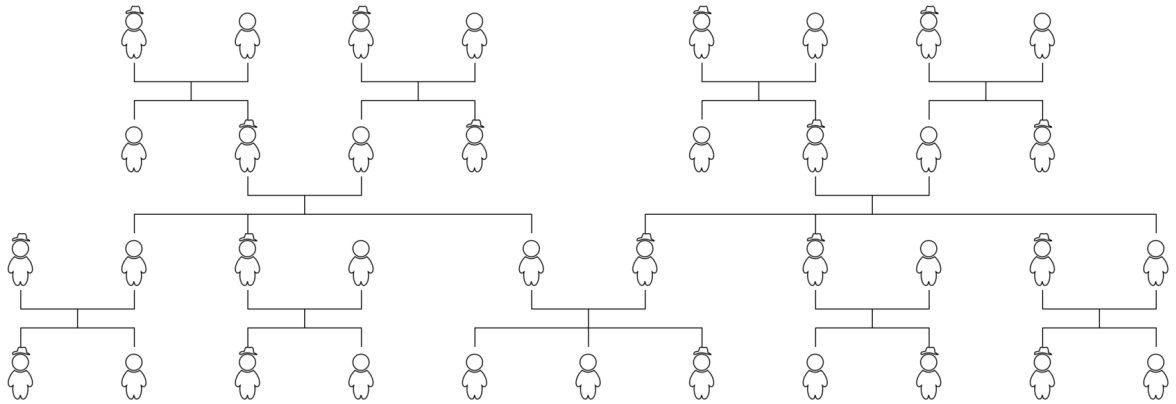

Figure 1. Family tree context for our simulations. Connections above figures reflect parent/child relationships. Connections under figures reflect lateral/spousal relationships. Male denoted with hats.

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.[4] In simulating the data this way, we make two simplifying assumptions. First,

[4]In Supplementary Figure A1, we emulate the simulations conducted by Navarro et al. (2012) to demon-

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 Insights, Order of Acquisition and the Characteristics-to-Defining Shift. In Model Insights, 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 three different kinship systems: Pukapukan, English and Turkish. We then take 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, falls out naturally from this model. Afterwards, we look at how broad patterns of over-generalization fall naturally out of the model.

In Order of Acquisition, we compare the predicted order of English kinship acquisition to 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. Instead, simulations varying assumptions about the data distribution indicates that the order of acquisition is likely driven by naturalistic data distributions (Benson & Anglin, 1987). In Appendix C, we explore how an alternative explanation–i.e., the acquisition of a kinship system instead of independent kinship concepts, might also explain the order of acquisition.

In Section Characteristic-to-Defining Shift, we replicate our analyses using simulations based on naturalistic learning contexts–i.e., informant provided family trees. We then use informant provided characteristic features to augment the model's hypothesis

strate that our main findings are robust under realistic values of α.

space, allowing rules based on characteristic features (e.g., UNCLE : $union(big, strong)$). 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.

Model Insights

The model learns typologically diverse systems as input varies. 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, the structure in the world supporting kinship terms, genealogy, is universal. 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. 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.

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 three kinship systems that vary in their complexity and are common in the languages of the world: Pukapukan, English and Turkish. In the tradition of Morgan (1871), Pukapukan, English and Turkish are from the Hawaiian, Eskimo and Sudanese family of kinship systems respectively. Extensions for the

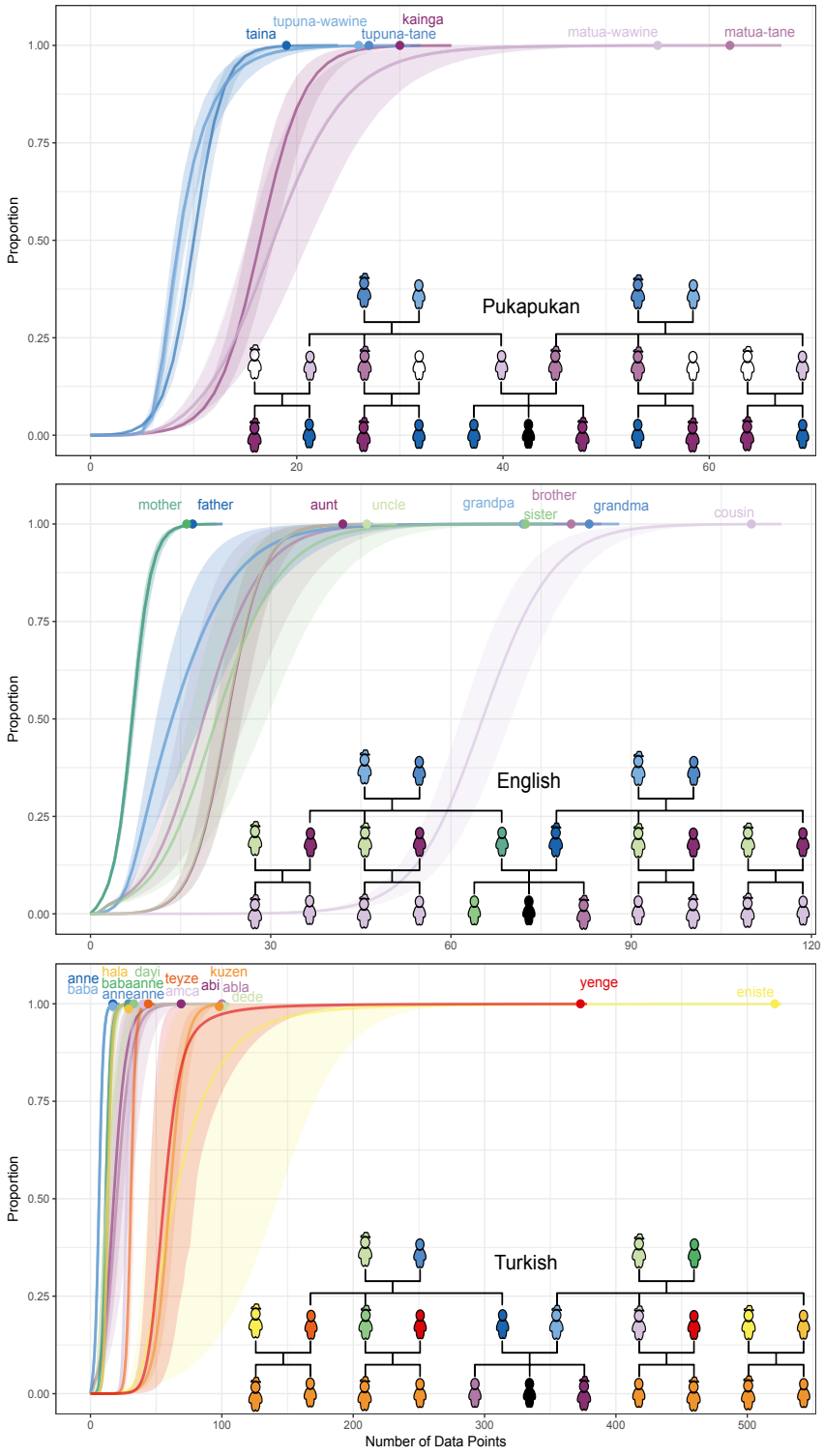

Figure 2. Average lexicon posterior-weighted accuracy for each word as a function of data points of that word. Shaded region denotes 95% bootstrapped confidence intervals. Insets show the color-coded extension of the terms.

	Word	MAP Hypothesis
Pukapuka	*kainga*	difference(generation0(X), sameGender(X))
	matua-tane	male(child(parent(parent(X))))
	matua-wawine	female(child(parent(parent(X))))
	taina	intersection(generation0(X), sameGender(X))
	tupuna-tane	male(child(parent(parent(parent(X)))))
	tupuna-wawine	female(child(parent(parent(parent(X)))))
English	aunt	female(difference(generation1(X), parent(X)))
	brother	male(child(parent(X)))
	cousin	difference(generation0(X), child(parent(X)))
	father	male(parent(X))
	grandma	female(parent(parent(X)))
	grandpa	male(parent(parent(X)))
	mother	female(parent(X))
	sister	female(child(parent(X)))
	uncle	male(difference(generation1(X), parent(X)))
Turkish	abi	male(child(parent(X)))
	abla	female(child(parent(X)))
	amca	male(difference(child(parent(male(parent(X)))), parent(X)))
	anne	female(parent(X))
	anneanne	female(parent(female(parent(X))))
	baba	male(parent(X))
	babaanne	female(parent(male(parent(X))))
	dayi	male(child(parent(female(parent(X)))))
	dede	male(parent(parent(X)))
	eniste	intersection(lateral(child(parent(parent(X)))), male(complement(parent(X))))
	hala	female(child(parent(male(parent(X)))))
	kuzen	difference(generation0(X), child(parent(X)))
	teyze	difference(female(generation0(female(parent(X)))), parent(X))
	yenge	difference(female(generation1(X)), child(parent(parent(X))))

Table 2

The maximum-a-posterior (MAP) hypotheses after learning. For readability, hypotheses were placed into simpler extensionally-equivalent forms.

kinship terms of these languages are provided in the insets of Figure 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. Turkish is even more complex. 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.

Figure 2 shows the predicted learning curves for each kinship term in Pukapuka, English and Turkish. 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 next section. Unfortunately, we could not find empirical data for the order of acquisition of Pukapukan and Turkish kinship terms.

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

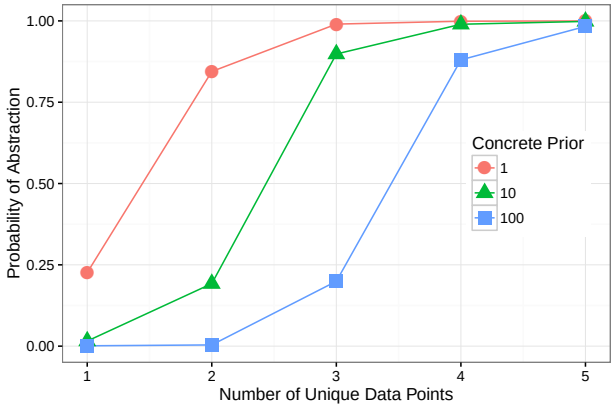

Figure 3. Probability of using abstraction as a function of unique data points at several different prior strengths for concrete reference. At higher prior values of concrete reference, the rise in the probability of abstraction is shifted to require more unique data points.

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.

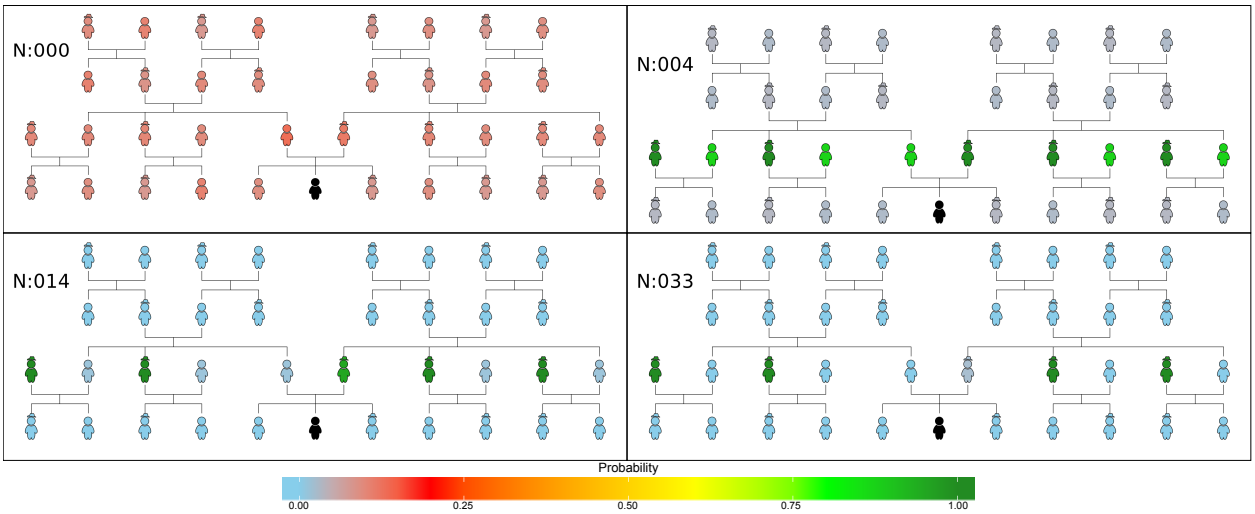

Figure 4. The posterior probability that each person on the tree is an uncle of the learner (in black) at various data amounts. Red indicates high probability and blue indicated low probability.

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 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. The model predicts shifts in the pattern of over-extensions observed as a function of data. For example, the model predicts that UNCLE should first be over-extended to all members of their parent's generation and then to their father. Predictions for the pattern of over-extension for each word is provided in supplemental material. These predictions can be empirically evaluated in future research.

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

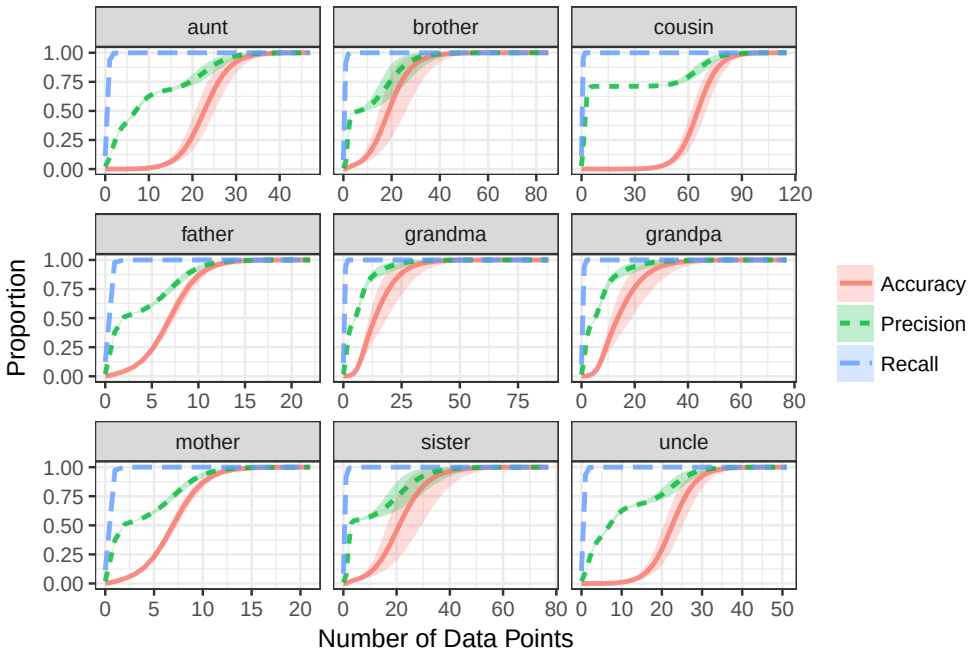

Figure 5. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Recall greater than precision is a hallmark of overgeneralization. Shaded regions represent 95% bootstrapped confidence intervals.

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[5].

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

Order of Acquisition: Simplicity and Data Distributions

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. Surveys of childhood interactions/experience with kids support input favoring the observed order of acquisition. In our model, we can directly pit these two factors against each other and see how they compare.

In this section, we will take a look at different priors and explore what data distributions would give rise to the empirically observed order of acquisition. For each analysis, we simulate 1000 data sets and run the learning model to measure the probability that kinship terms are acquired in a specific order. There are four patterns that we might see with these simulations (illustrated in Figure 6): 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). The x-axis in each panel of Figure 6 reflects the ordinal position in which words were learned. The y-axis reflects the probability that a word was acquired at that time. Words that were never learned in a particular position (e.g., *cousin* is never learned first) are omitted from the graph. The most likely order of acquisition is colored blue. If the order of acquisition is reliable, the probability of the most likely order of acquisition should be much greater than the other words learned at that ordinal position (top panels of Figure 6). Whereas, if the order of acquisition is unreliable, all of the words learned in that ordinal position should have similar probabilities (bottom panels of Figure 6).

Our initial simplicity prior (i.e., the PCFG in Table 1) mostly aligns with Haviland

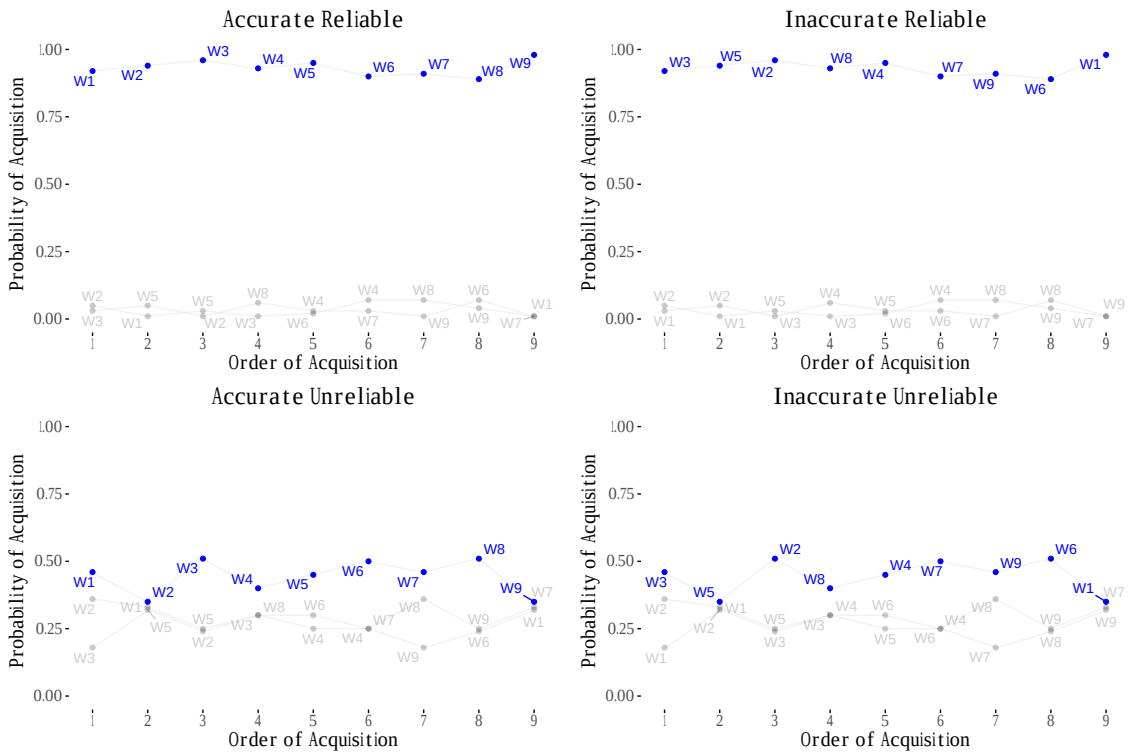

Figure 6. Possible patterns of order of acquisition. The x-axis reflects the ordinal position of acquisition. The y-axis reflects the probability of acquisition for each word. Words that have zero probability at a given ordinal position are omitted. The most likely order of acquisition is colored blue.

and Clark (1974)'s original formulation, as seen in Table 3. If data comes at a uniform rate for each word, we would expect to recover this order of acquisition; however, CHILDES frequencies (MacWhinney, 2000) suggest that the frequency distribution for kinship terms is not uniform. Surprisingly, the order of frequencies in CHILDES for kinship terms also doesn't align with empirical order of acquisition as suggested by Benson and Anglin (1987). The top left panel of Figure 7 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 predicts non-canonical order of acquisition.

That being said, CHILDES frequency estimates differ from the surveys of Benson and Anglin (1987). As a larger point, children do not utilize every instance of a word in their

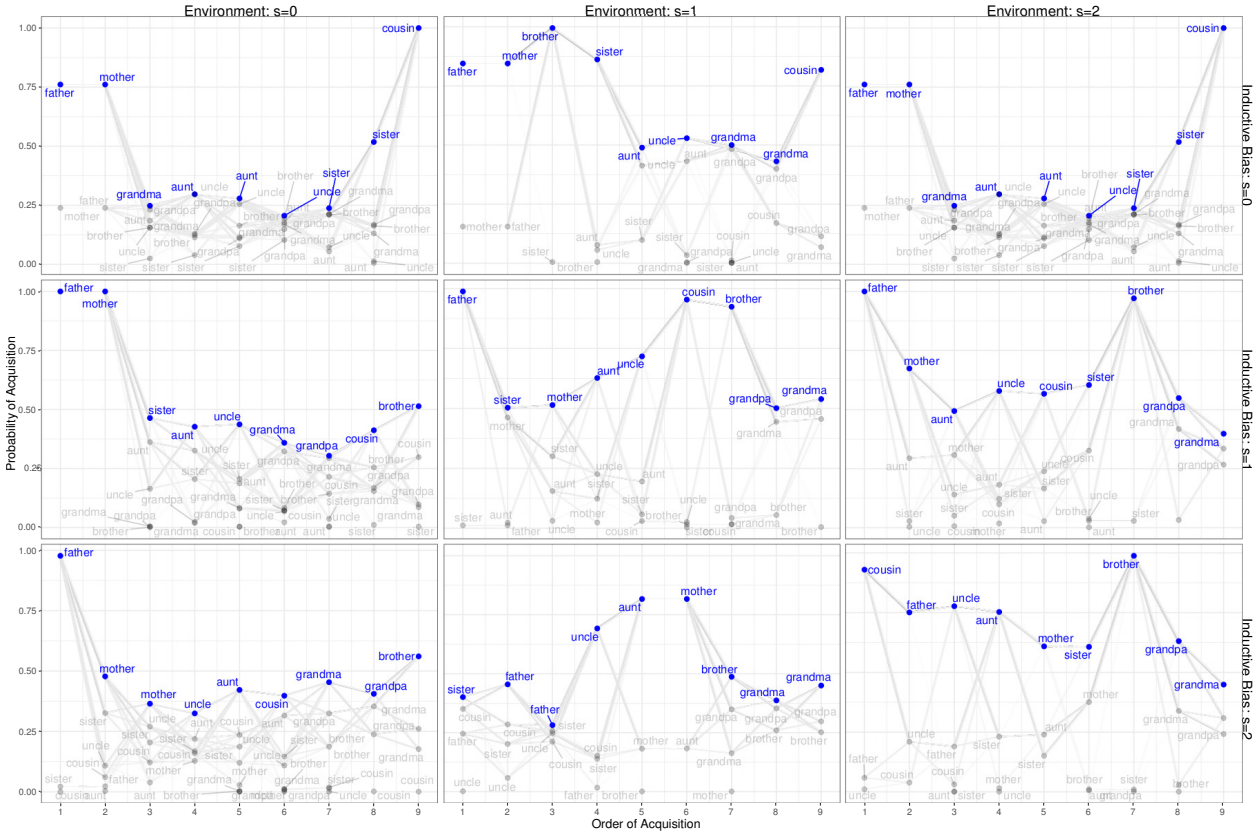

Figure 7. Simulations of the order of acquisition of kinship terms as a function of changes in environmental data distributions and the inductive biases of the learner. The strength of these biases are reflected in the s parameter with $s = 0$ reflecting no Zipfian bias. We find the most reliable patterns of acquisition when the environmental bias is weighted more than the learner's inductive bias. Note the apparent preference for *father* over *mother* is an artifact of how ties were handled (see Methods).

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 3

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.

environment as an effective learning instance (Mollica & Piantadosi, 2017). There is good 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 focus on the intuitions inspired by 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, and through assumptions about the environment. We can add these assumptions to the learner's inductive biases by adopting a weighted size principle likelihood, or Zipfian likelihood:

$$P(x|h,p) = \alpha \frac{d_x^{-s}}{\sum_{x \in h(p)} d_x^{-s}} + (1-\alpha)\frac{1}{|X|}, \tag{5}$$

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.

We can add 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 4.

$$P(p|w) \sim \alpha \text{ Zipf}(p|w,s) + \frac{(1-\alpha)}{|X|} \tag{6}$$

$$P(x|p,w) \sim \alpha \text{ Zipf}(x|w,p,s) + \frac{(1-\alpha)}{|X|} \tag{7}$$

For both implementations of these assumptions, the strength of this 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 about. Similarly, the learner does not expect to see these "black sheep" family members and discounts data including them. For our simulations, we assign distances to family members loosely based on Euclidean distance to the learner in Figure 1.

In Figure 7, 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, 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 $s = 1$ or Inductive Bias $s = 1$ and Environment $s = 2$). The order is closest to empirical order when the environment $s = 1$ and inductive bias $s = 0$, 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 assignment of distances (see Methods) and how we handle ties (i.e., alphabetically). If aunt/uncles were further from the learner than grandparents, we would expect grandparents to be acquired earlier. Differences between words of the same complexity (e.g., mother and father) are influenced by ties such that the alphabetical order appears dominant in Figure 7 where there is likely no bias.

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. 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). Further research is needed on how children filter their linguistic input.

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). Previous research has explained this pattern either as a by-product of shifts in representational capacity or learning mechanism (Werner, 1948; Bruner, Olver, & Greenfield, 1966; Kemler, 1983) or through the development of abstraction (Piaget & Inhelder, 1969). Here, we illustrate that the characteristic-to-defining shift manifests even without discrete changes in representation, processing or abstraction ability. Under our

model, the characteristic-to-defining shift is an epiphenomenon 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. 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.

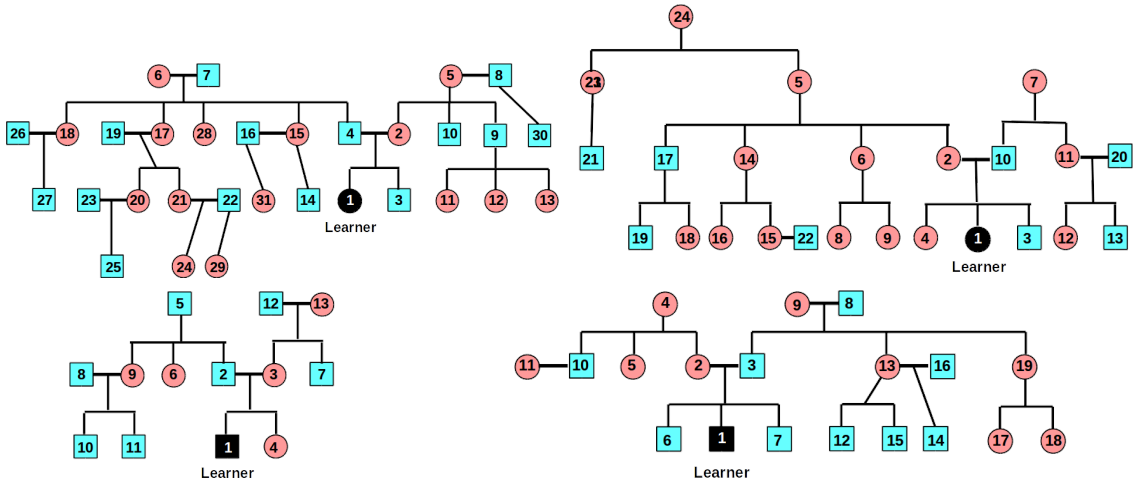

Figure 8. Distance-ranked family trees from informants. Circles represent females; squares males. Bold lateral lines denote spousal relationships. Informant 1 (top left) provided 107 unique features; Informant 2 (top right) 88; Informant 3 (bottom left) 92; and Informant 4, 59.

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 4

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 8), and provided ten one-word adjectives for each family member. For each informant, the unique adjectives were used to construct a binary feature matrix (adjective by family member). 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*)[6]. We used these features to augment the hypothesis space with the rules in Table 4.

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. As a result of the impoverished data/context, the model sometimes fails to learn the conventionally-aligned upon extension of a kinship term; however, it does always learn a program that selects the individuals consistent with the

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

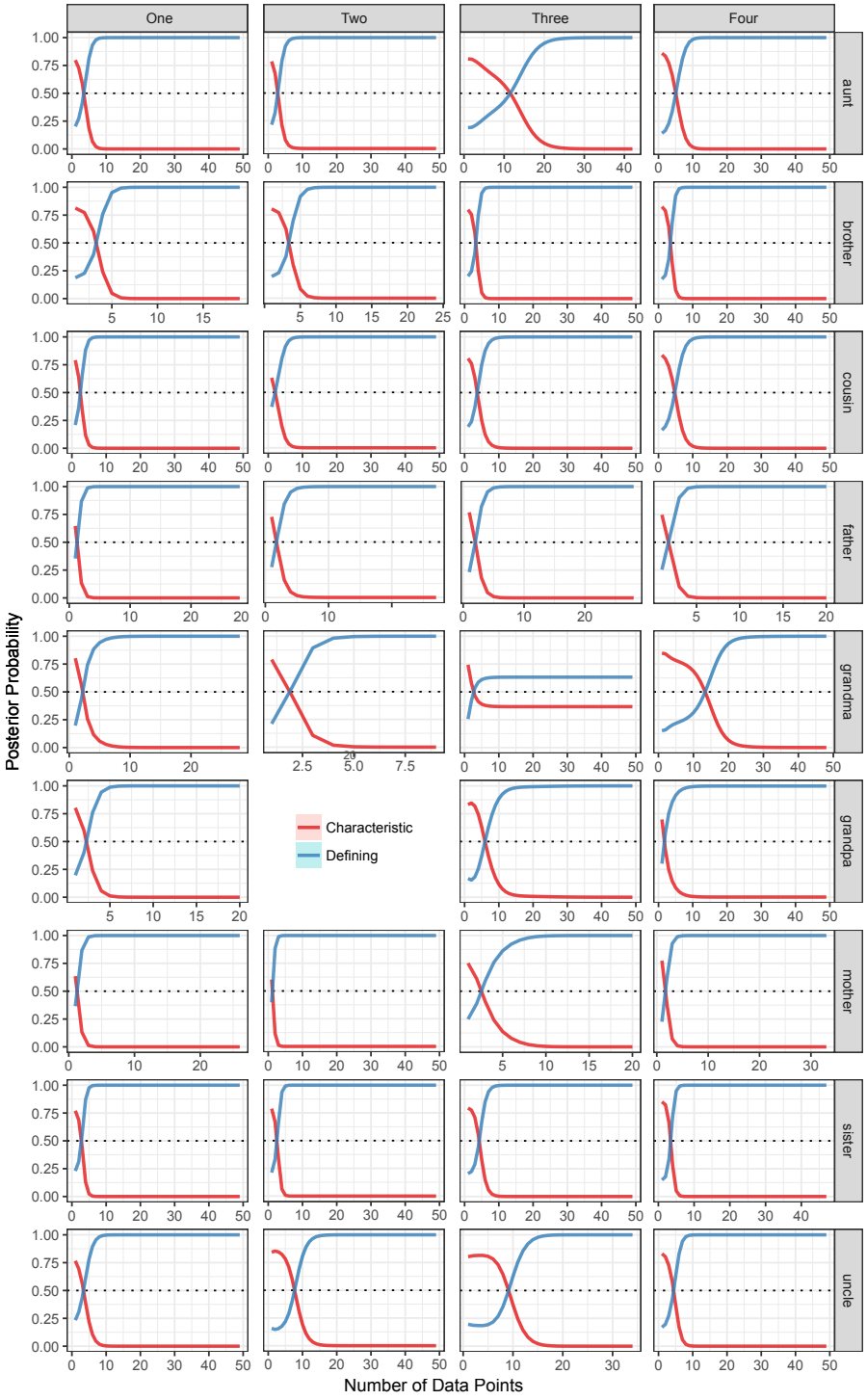

Figure 9. Average posterior probability of using a characteristic or a defining hypothesis (y-axis) as a function of the amount of data observed (x-axis) for words (rows) and informants (columns). Shaded regions reflect 95% bootstrapped confidence intervals. For all words, there is a characteristic-to-defining shift.

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. In modelling individual informants, we also return to the unweighted size principle likelihood, which is faster to compute and should not influence how the expected pattern of word use changes as a function of data. Therefore, we do not analyse the order of acquisition for individual informants.

To visualize the characteristic-to-defining shift (Figure 9), 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[7], 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 just like children purely due to the learning context–i..e, realistic data about logical relations and characteristic features.

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 Insights 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. Only with a precise, formal model of conceptual development like ours could one demonstrate that abstraction could be present from the outset of learning and a rational learner would still undergo a characteristic-to-defining shift.

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

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

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.

Discussion

By framing concept induction as logical program induction, we have demonstrated that an ideal learner model predicts many of the empirical phenomena seen in word learning. The model, like children, learns the kinship system consistent with its input offering a cross-linguistic proof of learnability. The model has a decent qualitative fit to the order of acquisition of kinship terms in English. The model illustrates both an early preference for concrete reference and patterns of over-generalization consistent with children's behavior. More importantly the model explains these phenomena in terms of the local distribution of data at the outset or learning, an inductive bias for simplicity and the relevant cognitive primitives that best predict the data. Additionally, our model provides a novel explanation for the characteristic-to-defining shift seen in children's early understanding of words, highlighting the role of the environment instead of proposing discrete changes in representation and processing. Lastly, Appendix C makes predictions for the patterns of children's word use for learning an inter-related conceptual system.

It is important to highlight several links between this model approach and past approaches. First, the model framework is compatible with similarity based approaches to early concept acquisition. For example, a program could capture similar features, feature correlations or both. 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 based approaches 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.

Lastly, 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). Based on these points, we suggest that our approach might provide an answer to the challenges for conceptual representations outlined by Murphy and Medin (1985).

Our work differs from past work in several ways. First our model is the first rational constructivist model (Xu, 2007), that captures the behavioral phenomena observed in kinship learning. Beyond kinship, our model derives the first 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[8], 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:$5 - 30$)[9]. 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 how children come to wield adult like concepts.

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

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

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

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 essentially must have 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. 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 MCMC 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 and Pukapukan 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:

$$P(\hat{h} = h|d) = \sum^{\mathcal{H}} \delta_{\hat{h}h} P(h|d). \tag{8}$$

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

$$P(x|d) = \sum^{\mathcal{H}} P(x \in |h|) P(h|d), \tag{9}$$

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

$$P(x \in |h|) = \begin{cases} 1 \text{ if } x \in |h| \\ \\ 0 \text{ else} \end{cases}. \tag{10}$$

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:

$$P(r_{SET \to p} \in h) = \begin{cases} 1 \text{ if } r_{SET \to p} \in h \\ 0 \text{ else} \end{cases} . \tag{11}$$

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

$$P(r_{SET \to p}|d) = \sum^{\mathcal{H}} P(r_{SET \to p} \in h) P(h|d). \tag{12}$$

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.

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.

$$w \sim \text{Multinomial}(\theta) \tag{13}$$

$$(x, p) \sim \text{Uniform}(|(x.p)|) \tag{14}$$

For the Zipfian weighted order of acquisition analyses, 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 6 and 7.

For both schemes, we calculate the posterior accuracy of each hypothesis as a function of data following Equation 8 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.

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 4 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

$$P(r_{FSET\rightarrow\text{feature}}|d) = \sum^{\mathcal{H}} P(r_{FSET\rightarrow\text{feature}} \in h)P(h|d), \tag{15}$$

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

$$P(r_{FSET\rightarrow\text{feature}} \in h) = \begin{cases} 1 \text{ if } r_{FSET\rightarrow\text{feature}} \in h \\ 0 \text{ else} \end{cases}. \tag{16}$$

References

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

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.

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

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).

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.

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.

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

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., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Company.

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

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.

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*.

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. (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.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Lewis, M., & Frank, M. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

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

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

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

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*.

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

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

Nancekivell, S. E., & Friedman, O. (2017). "because it's hers": When preschoolers use ownership in their explanations. *Cognitive science*, *41*(3), 827–843.

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.

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*.

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.

Rescorla, L. A. (1980). Overextension in early language development. *Journal of child*

language, *7*(02), 321–335.

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.

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.

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., & 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

`colala.bcs.rochester.edu/people/FrankMollica/kinship.html`.

Appendix A

Alpha Analysis

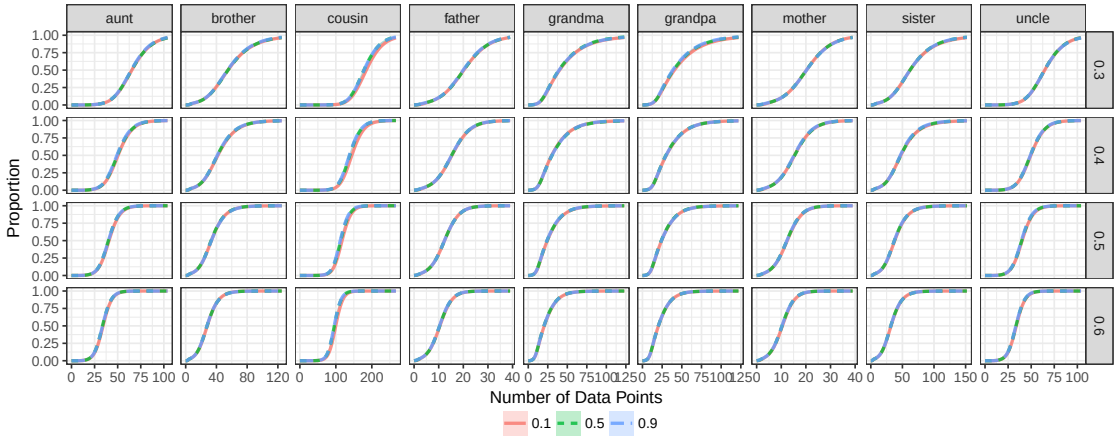

Figure A1. Posterior weighted accuracy (y-axis) as a function of data (x-axis) for models with different sampling assumptions (linetype and color) for different words (columns) and environmental reliability values (rows). The virtually invisible shaded regions reflect 3 standard errors of the mean.

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 guague 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:

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

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:

$$\frac{\sum_{x \in \hat{h}}[x \in h]}{|\hat{h}|}. \tag{18}$$

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

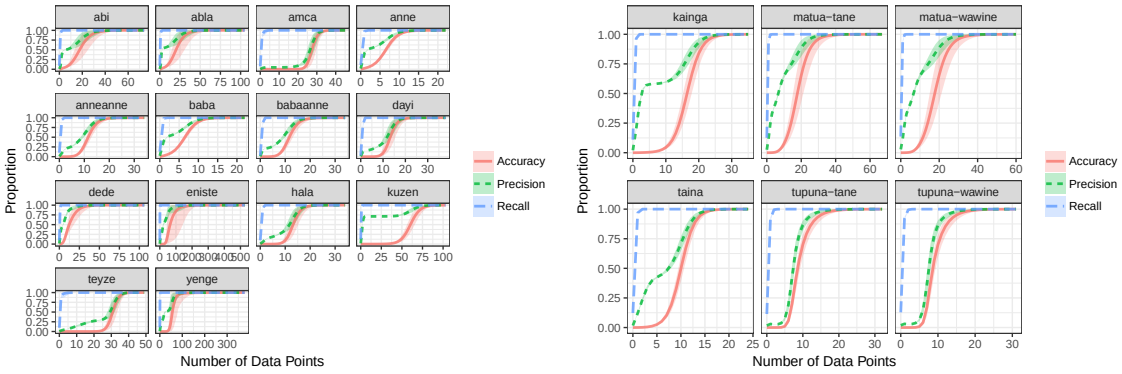

Figure B1. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

Figure B1 displays the F_1 plots for Pukapuka and Turkish. 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.

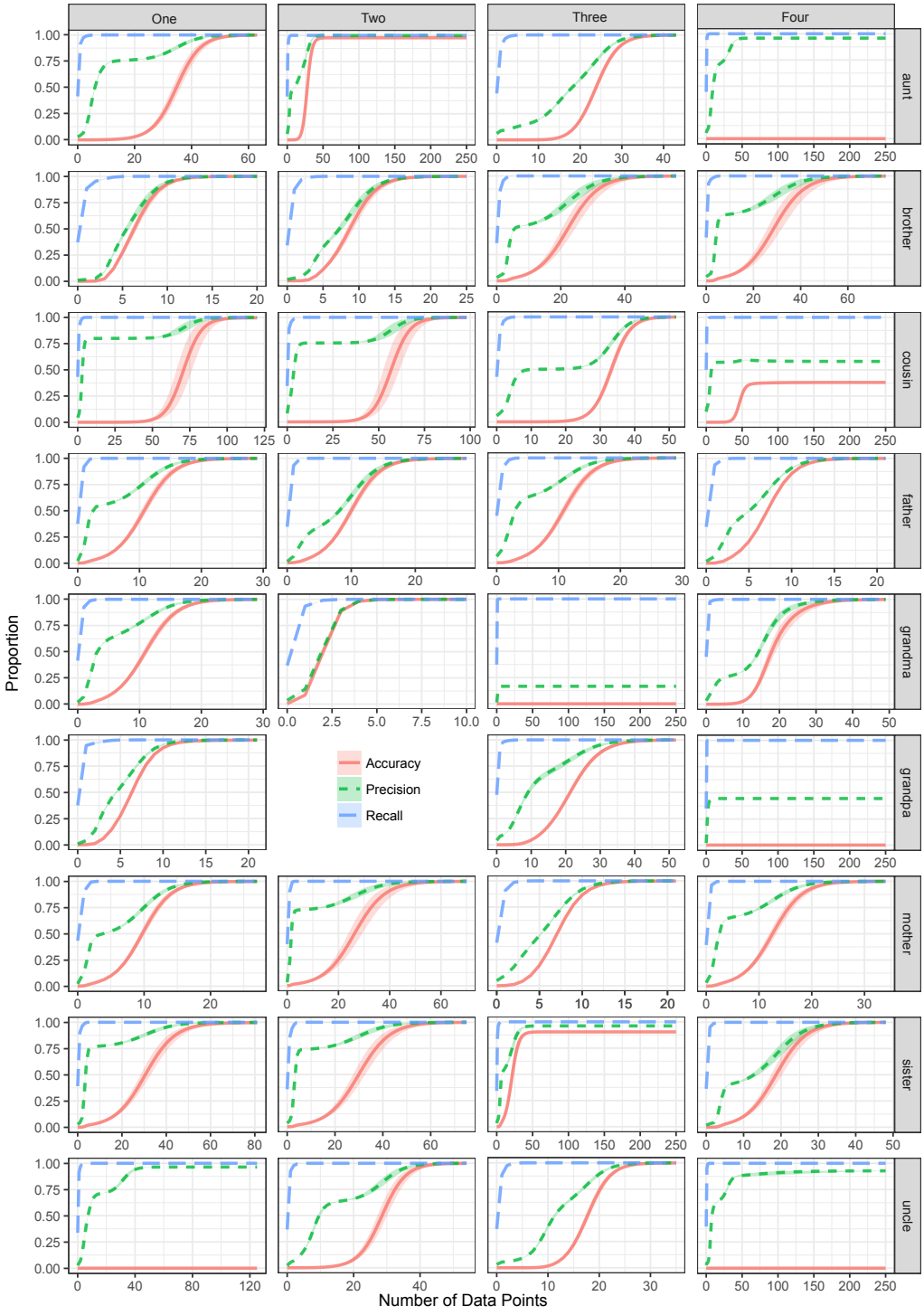

Figure B2. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

Appendix C

Learning an inter-related system: Lexicon

Up until now, we have been assuming that kinship terms are learned independently of each other. In this appendix, we consider the advantages of learning an inter-related system, or lexicon. The simplest way to introduce cross-word dependencies in our model is to change the likelihood from operating over hypotheses h generating data for an individual word to lexicons $\mathcal{L}$ generating data for all words in the system. The simplest prior for a lexicon is the product of the PCFG prior for each hypothesis in the lexicon:

$$P(\mathcal{L}) = \prod_{h \in \mathcal{L}} P(h). \tag{19}$$

The likelihood still follows a noisy size principle:

$$P(d|\mathcal{L}) = \frac{\alpha}{\sum\limits_{h \in \mathcal{L}} |h|} + \frac{(1-\alpha)}{|\mathcal{D}|^2}. \tag{20}$$

Formalizing the problem as lexicon learning has an interesting consequence for how probability mass moves over the hypothesis space for individual words. Probability mass always moves over the hypothesis space along the Pareto-front, or the curve reflecting the optimal trade-off of prior and likelihood (see Supplementary Material). Borrowing the analogy from economics, we can look at this as data purchasing complexity. A hypothesis can only afford to be complex if it explains a lot of data. With very few data points, the highest posterior hypotheses are not very complicated because simpler hypotheses can explain the data. As more data is observed, the pattern of the data can justify more complex hypotheses. In the limit of observing data, the data pattern stabilizes and the highest posterior hypothesis is at the Pareto-front–i.e., the simplest hypothesis that explains all the data. At this point, observing more data will not change the posterior mass over the hypothesis space. When a noisy size principle likelihood operates over lexicons instead of hypotheses, probability mass travels along the Pareto-front slightly faster (as can be seen in Supplementary Materials).

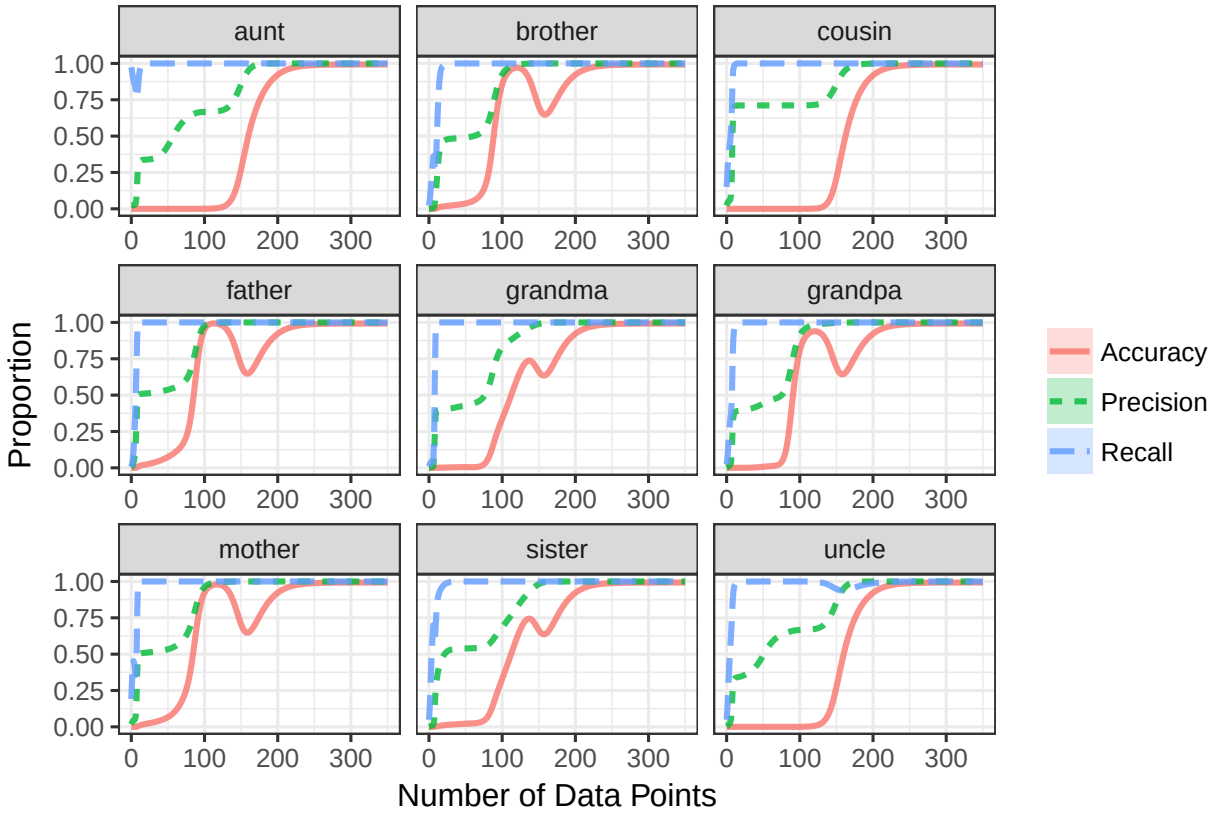

Figure C1. Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

We implemented these changes to the model and conducted the same analyses in the Model Insights and Order of Acquisition sections of the main text. The F_1 plot in Figure C1 illustrates the same patterns of over-generalization found when words are learned independently; however, the lexicon formalization learns all of the kinship terms with fewer data points than the independent formalization. As a result, model comparison between the independent hypothesis and lexicon formations could reveal whether children approach kinship as learning a system as opposed to independent hypotheses for kinship terms. Future research is needed to collect the appropriate dataset for such a model comparison.

Turning to order of acquisition, Figure C2 shows the order of acquisition of the lexicon model given 1000 different CHILDES weighted data distributions. Interestingly, the

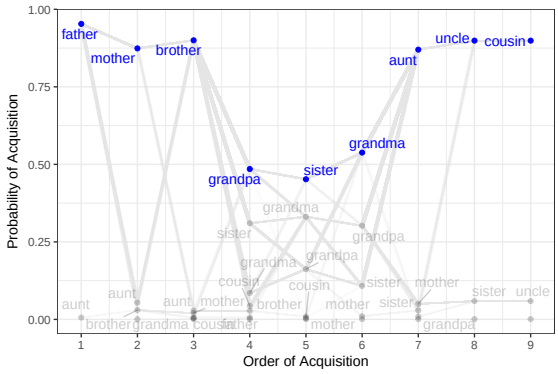

Figure C2. The order of acquisition under a lexicon (left).

pattern is consistent across data distributions and fairly consistent with empirically observed order of acquisition. As shown in the main text, this result cannot be attributed to the simplicity prior or the CHILDES weighted environment. Before we attribute this result to the lexicon's likelihood, there is one trivial alternative worth addressing: the distribution of correct/incorrect words across lexicons might be biased. Our lexicon space is a finite approximation to the infinite space of lexicons specified by the PCFG in Table 1 and thus, is incomplete. If our finite space happened to contain more hypotheses approximating the correct order of acquisition than chance, we would not be able to tell if this result is an artifact of our approximation instead of learning. We think this is unlikely for two reasons. First, we took measures to balance the correct/incorrect word correlations across lexicons by mixing words across lexicons using Gibbs sampling (see Methods). Second, examination of the correlations of our finite approximation to the lexicon space are incongruous with the dominant order of acquisition in Figure C2.

If the proper dataset for model comparison is collected, future implementations of in this model framework can adopt priors that reward reuse of primitive functions (as in Goodman et al., 2008), implement recursion (Mollica & Piantadosi, 2015), memoize combinations of primitives that are useful (O'Donnell, 2015) or analogize from already learned knowledge (Cheyette & Piantadosi, 2017) to learn about alternative conceptual architectures children might adopt when learning inter-related systems.