The Human Learning Machine: Rational Constructivist Models of Conceptual Development

by

Francis Mollica

Submitted in Partial Fulfillment of the

Requirements for the Degree

Doctor of Philosophy

Supervised by Professor Steven Piantadosi

Department of Brain and Cognitive Sciences

Arts, Sciences and Engineering

School of Arts and Sciences

University of Rochester

Rochester, New York

2019

Dedicated to Dr. Lalita Krishnamurthy

Table of Contents

Biographical Sketch

Frank Mollica was born in Brooklyn, NY. He attended the University at Buffalo, and graduated with a Bachelor of Arts degree in Linguistics and a Bachelor of Science degree in Psychology in 2014. He began doctoral studies in Brain and Cognitive Sciences at the University of Rochester in 2014. He was awarded the Donald M. and Janet C. Bernard Fellowship in 2017 and received a Master of Arts degree from the University of Rochester in 2017. He pursued his research in cognitive science under the direction of Steven Piantadosi.

The following publications were a result of work conducted during doctoral study:

- Mollica, F., Piantadosi, S. T., & Tanenhaus, M. K. (2015). The perceptual foundation of linguistic context. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1613–1618).

- 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).

- Martí, L., Mollica, F., Piantadosi, S. T., & Kidd, C. (2016). What determines human certainty? In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 698–703).

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

- Mollica, F., & Piantadosi, S. T. (2017b). An incremental information theoretic buffer supports sentence processing. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 805–810).

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

- Yan, S., Mollica, F., & Tanenhaus, M. K. (2018). A context constructivist account of contextual diversity. In *Proceedings of the 40th annual meeting of the cognitive science society.*

- Oey, L., Mollica, F., & Piantadosi, S. T. (2018). Adults use gradient similarity information in compositional rules. In *Proceedings of the 40th annual meeting of the cognitive science society.*

- Martí, L., Mollica, F., Piantadosi, S. T., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind*, *2*(2), 47-60. doi: doi: 10.1162/opmi_a_00017

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

- Rubio-Fernández, P., Mollica, F., Oraa Ali, M., & Gibson, E. (2019). How do you know that? automatic belief inferences in passing conversation. *Cognition*, *193*, 104011.

- Mollica, F., & Piantadosi, S. T. (in revision). *Logical word learning: The case of kinship.*

- Register, Y., Mollica, F., & Piantadosi, S. T. (under review). *Semantic verification is flexible and sensitive to context.*

- Rubio-Fernández, P., Mollica, F., & Jara-Ettinger, J. (under review). *Why searching for a blue triangle is different in english than in spanish.*

- Mollica, F., & Piantadosi, S. (in prep). *Universal and cultural-specific processes in exact number word acquisition.*

Acknowledgments

I am indebted to the following individuals for support, great conversations and memories that I will cherish: Steven Piantadosi, Michael Tanenhaus, Naomi Feldman, Aaron White, Celeste Kidd, Gail Mauner, Jürgen Bohnemeyer, JP Koenig, Rui Chaves, Erica Chung-i Su, Andreas Brocher, Cody Hashman, Lalita Krishnamurthy, Evelina Fedorenko, Ted Gibson, Paula Rubio-Fernández, Shirlene Wade, Yim Register, Louis Martí, Sam Cheyette, Holly Palmeri, Maddie Pelz, Carla Macias, Emily Sumner, Lauren Oey, Claude Galette, Matthew Hochberg, Sarah Elderkin, Habiba Azab, Richard Lange, Carol Jew, Elizabeth Shay, Stephen Ferrigno, Maya Wang, Zach Burchill, Linda Liu, Maryam Seifeldin, Shraddha Shah, Shaorong Yan, Kathy Corser, Jennifer Gillis, Chris Dambra, Aki, Amanda, Wednesday, Shanna, Kevin, Sebastian, David R., Ben, Chris, Chelsea, Thais, Dave K., Esteban B., Sarah B., Keturah, the Raizada lab, the HLP lab, the Haefner Lab, Rachel R., Richard F., Josh R., Idan, Matt S., Yev, Till, MH, Anna, Kasia, Melissa, Helena, Jayden, Ray, Brian, Roger, Amy, Charles, Simon D., Simon, Simon L., Vanessa, Josh, Hyungwook, Ben, Keith, Kevin, Samarth, Hannah, Alex, Hugh, Dan Y., Tim, Julian, Elika, Dick, Jessica, Brad, Chigusa, Florian, the other Matt S., Hilary, Gabriella, Brendan, Ryan, Alexis, Frank and Robin, and, finally, the coffee providers.

Abstract

This thesis develops the hypothesis that the systematic patterns of children's word use over the course of development are the natural consequence of a sophisticated inductive learning mechanism operating with insufficient data. In this thesis, we sketch out a first-principles account of lexical-conceptual development and implement this model framework for the case of children learning kinship. Kinship is a valuable semantic domain to investigate because children show the same developmental trajectory for early word (mis)use, as in their first year of life, spread out over nine years. A major limitation of evaluating this model and all models of conceptual development is that we have poor intuitions about how children make use of data. To remedy this, we build a data analysis model to investigate the profile of data usage in word learning; although this technique will be broadly applicable to developmental science. We then illustrate how this technique can be used to check the first principles model of inductive learning and investigate the learning process by compiling a large cross-cultural dataset assessing children's knowledge of exact number words. We then take a step back from the learning mechanism and use Fermi-estimation and information theoretic techniques to quantify the scale of language learning tasks and highlight the likelihood of sophisticated learning mechanisms for word meanings.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professors Steven Piantadosi (advisor) and Michael Tanenhaus of the Department of Brain and Cognitive Sciences at the University of Rochester and Professor Naomi Feldman of the Department of Linguistics and the Institute for Advanced Computer Studies at the University of Maryland. Chapters 1, 2, 4 and 6 were written by Francis Mollica under the supervision of Steven Piantadosi. Chapter 2 is currently in revision as Mollica and Piantadosi (in revision). Early versions of the work presented in Chapter 2 have been published as Mollica and Piantadosi (2015) and Mollica, Wade, and Piantadosi (2017). Chapter 3 was written with Steven Piantadosi and published as Mollica and Piantadosi (2017a) in *Open Mind.* Chapter 4 benefited from data collected by Julian Jara-Ettinger and Isabelle Boni and discussions about numerical analysis with Ankani Chattoraj and Sabyasachi Shivkumar. It is in preparation for publication. Chapter 5 was written with Steven Piantadosi and published as Mollica and Piantadosi (2019) in *Royal Society Open Science.*

Graduate study was supported by a the Department of Brain and Cognitive Sciences at the University of Rochester. The research and this dissertation were funded by an NIH Grant (R01HD085996) to Jessica Cantlon and Steven Piantadosi,

start up research funds awarded to Steven Piantadosi, and a Donald M. and Janet C. Barnard Fellowship awarded to Francis Mollica.

List of Tables

List of Figures

Chapter 1

The life of a word

It is a truth universally acknowledged that a child in possession of a new word might not have the adult-like knowledge of how to use it (e.g., P. Bloom, 2000; E. V. Clark, 1973; Brown, 1973). Children exhibit many systematic patterns of incorrect word use. Sometimes young children fail to generalize word meanings to correct referents (e.g., Kay & Anglin, 1982). Other times, they generalize word meanings to incorrect referents (e.g., Rescorla, 1980). Over the course of development, children's understanding of a word shifts between superficial generalizations to the application of diagnostic criteria (e.g., Keil & Batterman, 1984; Keil, 1989). For some words, it can take children years after they have started producing a word to master the non-obvious criteria governing its use (e.g., Borer & Wexler, 1987). Other words require the development of rich conceptual structures before children acquire adult-like understanding (e.g., Carey, 2009; Piantadosi, Tenenbaum, & Goodman, 2012). The goal of this thesis is to evaluate the extent to which a rational constructivist learning mechanism would predict these systematic patterns of word use as natural process of lexical/conceptual development. In the process, we develop broadly-applicable, statistically-principled data analysis tools in order to learn about learning. We begin by outlining the time course of lexical-conceptual development.

The development of the lexicon and conceptual development can be thought of as two separate problems. Lexical acquisition is the process by which learners map words to concepts. Conceptual development is the process by which learner's construct hypothetical word meanings from conceptual machinery. There is strong evidence to suggest that there is conceptual change occurring over the course of development (Carey, 2009). In these cases, conceptual development necessarily precedes the acquisition of adult-like word meanings. The relevant question thus becomes how much of children's lexical acquisition, and patterns of word use, can be viewed as conceptual development as opposed to constructing mappings between words and already acquired concepts (L. R. Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Snedeker, Geren, & Shafto, 2012)?

While lexical acquisition is a lifelong process, the vast majority of our lexicon is acquired fairly quickly. At the age of six months, infants have some comprehension of concrete nouns (Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999). Understanding of basic, abstract non-nouns follows about four months later (Bergelson & Swingley, 2013, 2015). Children start producing their first words around 12-16 months of age (B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015; Schneider, Daniel, & Frank, 2015). Within the next six to eight months, children figure out the meanings for a large number of words in what is referred to as the *word spurt* (Benedict, 1979; Carey, 1978; Goldfield & Reznick, 1990). The initial words learned are mostly nouns and verbs with concrete referents, before abstract words and function words

(Bates et al., 1994). By the time children are in college, their vocabulary approaches around 40,000 words across all lexical categories (Brysbaert, Stevens, Mandera, & Keuleers, 2016) and 25,000 idioms (Jackendoff, 1997).

Informed by the early shift in coverage from concrete to abstract words before acquiring words reflecting algorithmic operations or functional relations, many researchers have proposed that language acquisition is largely constrained by cognitive/conceptual development (e.g., Smiley & Huttenlocher, 1995; K. Wexler, 1999; Shore, 1986; Bowerman, 1974): Without an ability to abstract away from experiences, children would only acquire words with concrete properties. As children acquire the ability to abstract, they will acquire more abstract nouns and verbs and function words. Words that require the development of complex conceptual structures (e.g., number and time words) will take even longer for children to acquire. In this view, the growth patterns of the lexicon are driven by conceptual development.

Alternatively, this pattern of lexicon growth can be explained by positing that children have access to fully formed conceptual representations, but have trouble extracting the linguistic representations that these concepts should map to (L. Gleitman, 1990; L. R. Gleitman et al., 2005; Snedeker, Geren, & Shafto, 2007; Snedeker et al., 2012). Using international adoption as a case study, Snedeker et al. (2007) showed that children who already acquired one language still show this developmental trajectory of words despite having conceptual structures for these words in another language. One exception to this pattern is the adopted children's precocious use of time words, which possibly suggests that conceptual development is occurring in the background. This thesis starts with the assumption that conceptual development must occur in order to learn the meaning of words; however, we do not use the growth patterns of the lexicon as evidence for our claim, nor do we require conceptual development to cause the the growth patterns of the lexicon. Instead, we use the systematic patterns of children's early word usage to test the hypothesis that children's word meanings are constrained by conceptual development.

1.1 Meaning at the word spurt

While infants have knowledge of some word meanings (Bergelson & Swingley, 2012; Bergelson & Aslin, 2017) and the capacity to use words as invitations for concepts (Waxman & Markow, 1995; Waxman, 1999; Waxman & Booth, 2001), it is difficult to infer the extent to which their knowledge matches adult-like competence. The majority of our understanding of children's early word meanings come from diary studies documenting the first years of a child's language production (e.g., Brown, 1973; E. V. Clark, 1973). These studies report two distinct patterns of incorrect word usage, which have been confirmed in experimental settings (Rescorla, 1980; Kay & Anglin, 1982; Fremgen & Fay, 1980): children *under-extend* a term and children *over-extend* a term. For example, children might under-extend the word *doggo*, using it to refer to their dog and no other dog. Later on children might over-extend the word, using it to refer to every four legged thing, including horses, cows and cats.

The earliest explanation for over-/under-extensions was that children do not have the adult-like conceptual knowledge corresponding to the word but hastily map a word to their partial understanding (E. V. Clark, 1973; Kay & Anglin, 1982). Philosophically, children are faced with Quine (1960)'s problem of the indeterminancy of translation—i.e., any simple utterance in isolation has *multiple possible translations* and any complex utterance in context is not necessarily a mere combination of the component utterance translations. Any account of learning that faces this problem has the potential to over-generalize and under-generalize. In the context of word learning, researchers have proposed three different classes of accounts that could reasonably produce the over-/under-extension observed in children: exemplar accounts (e.g., Ambridge, 2018), abstraction accounts (e.g., Doumas, Hummel, & Sandhofer, 2008), and construction accounts (e.g., Xu, 2007, 2016, in press). To date, these proposals have never been formalized on the same learning problem to distinguish between them.

Under an exemplar account, children store the entirety of their experience and extend a word to novel situations based on some similarity computation between stored instances of the word and the current situation (Ambridge, 2018). Learning is holistically storing experiences. In this scenario, the order in which a learner experiences referents of a word can produce over-/under-extension based on the weighting of conceptual features when computing similarity—i.e., the spurious correlations discussed by Murphy and Medin (1985). For example, most bananas are genetically modified but being genetically modified is not necessary or sufficient for an object to be a banana. Yet, in an exemplar account *banana* could be under-extended to non-genetically modified bananas. On the other hand, *banana* might also be over-extended to other genetically modified foods.

The abstraction account also assumes that children have access to a rich conceptual representation of their experience. Instead of storing experience, the problem of conceptual development is framed as learning the relevant abstraction over some part of this rich conceptual representation (Doumas et al., 2008; Rogers & McClelland, 2004). If a learner abstracts over some particular subset of their limited experience, there is a high probability that they will pick the wrong subset. For example, most instances of *birthday* co-occur with cake and so generalizing over CAKE would select the valid birthday events well, but might over-extend the term to include holidays. On the other hand, if a learner encounters a highly valid cue such as the happy birthday song, they might fail to include the boring, song-less birthdays in the events denoted by *birthday*. Similar to the exemplar account, these models would be sensitive to spurious correlations[1]. In general, the exemplar and the abstraction accounts can be viewed as two different approaches to representing a single computational system.

In contrast to the exemplar and abstraction accounts, which each emphasize the central role of the child's access to rich conceptual representations of experience, the construction account assumes that learners have an inventory of primitive computations/representations that are used to compute a word's meaning (Xu, 2007, 2016, in press). Over-/under-extensions occur as the result of the rational construction of a hypothetical word meaning, or concept. When a child only hears a word (e.g, *blanket*) to denote a single referent (e.g., their blanket), they should essentially memorize the word-referent mapping, resulting in under-extension. However, when a child hears a word (e.g., *apple*) used with multiple referents (e.g., a red delicious and

[1] That being said, it is not clear that the spurious correlations are as problematic as described by Murphy and Medin (1985).

a granny smith), they should construct a concept that extends to include everything that they've seen (e.g., rounded with a stem), which can result in over-extension (e.g., a pear is also rounded with a stem). The main difference between the exemplar/abstraction accounts and the constructionist account is a focus switch from *how information is being represented* to *what information is being represented.* In the exemplar and abstraction accounts, the learner's rich conceptual experience of the environment (albeit represented differently in each account) is used to derive hypothetical word meanings and correlations across experiences give rise to extension errors; whereas, in the constructionist account, the conceptual primitives are used to build hypothetical word meanings and extension errors are governed both by the primitive computations and the correlations in the input. It is important to note that the construction account does not preclude the availability of rich conceptual representations of the learner's current experience.

While the above learning accounts would produce the observed over-/under-extension behavior, researchers have noted several alternative explanations that do not appeal to learning. If a rational pragmatic agent needs to refer to something that they do not have a word for, the optimal solution is to over-extend a term that they already know (Fremgen & Fay, 1980). For example, if you don't know the word *lime* but you do know *lemon*, a lime is basically a green lemon and a cooperative listener would understand *lemon* to refer to a lime in a context with no lemons or *green lemon* in a context with a lemon. Researchers have also proposed that over-/under-extension might be due to performance errors, noting that the timing of children's (mis)use immediately follows the rapid acquisition of their first 100-150 words, without giving them much time to practice retrieving words from their growing lexicon (Gershkoff-Stowe, 2001).

To properly contextualize these competing accounts, it is important to note that diary studies often describe the over-/under-extension patterns in terms of children's incomplete conceptual knowledge (Brown, 1973; E. V. Clark, 1973); whereas, controlled experiments primarily highlight the role of pragmatic processing and performance limitations in explaining these patterns (Fremgen & Fay, 1980; Gershkoff-Stowe, 2001). One possibility for this discrepancy is that diary studies have the measurement resolution of behavior on the order of days; whereas, experimental tasks measure behavior averaged over months. It's not controversial that children's early word use should be influenced by pragmatic processing and performance errors; nonetheless, by conducting experiments by measuring at month intervals when children are rapidly acquiring multiple words every day, the experimental studies might not be sensitive enough to detect patterns of word use driven by incomplete conceptual development. Undoubtedly, pragmatics, retrieval processes and learning contribute to children's early word use. To determine the extent to which these three components explain behavior, we need formalized accounts of these mechanisms that make targeted and precise empirical predictions about how behavior might differ for both the time-scale and patterns of word use. Ideally, these models would be implemented on a learning task with a protracted developmental domain so that researchers can measure behavior over the course of development.

1.2 Protracted developmental domains

Fortunately for researchers, children's acquisition of some words, including space, time, number, color, kinship, morality, motion, prepositions, quantifiers and other function words, exhibit a protracted developmental trajectory (Wynn, 1990, 1992; Bowerman, 2007; Haviland & Clark, 1974; Keil, 1989; Borer & Wexler, 1987; Tillman, Marghetis, Barner, & Srinivasan, 2017; Wagner, Chu, & Barner, 2018; Tillman & Barner, 2015). Even though these words make up a small portion of the lexicon, the majority of research on conceptual development has focused on them, likely because there is reliable variance in children's word use and it can be observed with convenient measurements on the order of months. As a further consequence, the delayed trajectory is normally interpreted as conceptual development delaying acquisition. In further support of this view, we see savings on re-learning for some of these domains in international adoption studies (Snedeker et al., 2012), consistent with one-time conceptual development. In this light, the fact that we see the same patterns of over-/under-extension seen in the acquisition of common nouns in some of these protracted developmental domains (e.g., kinship Benson & Anglin, 1987), where children have other means of establishing reference and when retrieval process is fluid for a large portion of their vocabulary, further suggests that children's early word (mis)use might be due to partial conceptual knowledge.

Words that exhibit a protracted developmental trajectory normally have abstract meanings or reflect logical/algorithmic computations. Within these domains, children often exhibit additional systematic patterns of non-adult-like word use beyond over-/under-extension. For example, for abstract concepts involved in morality (e.g., JAIL) or concepts with clear rule governed systems (e.g., kinship), children go through a *characteristic-to-defining shift* (Keil & Batterman, 1984). For example, young children endorse the idea that a building in a slum with bars on the window is a *jail* even though the residents are free to come and go as they please. At the same time, they do not endorse the notion that a beautiful castle with delicious food and a swimming pool that residents are free to use could be a *jail*; despite the residents never being allowed to leave without strict permission, or enter unless they have done something wrong (Keil & Batterman, 1984). These shifts are usually explained as a fundamental change in the learning process or representations (Werner, 1948; Bruner, Olver, & Greenfield, 1966; Kemler, 1983; Shultz, Thivierge, & Laurin, 2008) or the development of abstraction ability (Piaget & Inhelder, 1969); although, a single shift in mechanism or abstraction ability does not predict the differences in the timing of the shift observed across different domains (Keil, 1983).

In more complex conceptual domains, children's systematic patterns have been used to motivate stage theories of development for particular conceptual domains. For example, when children learn the meanings of exact number words, they proceed through a succession of sub-knower stages even though they produce several number words (Wynn, 1990, 1992). Children start out as non-knowers, counting out a random amount when asked for a specific number amount. They progress to a one-knower stage, counting out exactly one correctly, but failing for larger numbers. Similarly, they progress through a two-knower, three-knower and four-knower stage, where they can count out amounts up to their knower level, but fail on larger amounts. Then, children master the cardinal principle and can count out an amount for every number they can count. The systematic patterns of development for complex conceptual domains are important for distinguishing between the construction and exemplar/abstraction accounts of learning as the rapid shifts in generalization patterns are often not reducible to correlations between features in the input and conceptual representations.

1.3 Approach

The approach in this thesis is to formalize a model framework for conceptual development from the first-principles of concept construction. We then implement computational models to identify and predict the patterns of children's early word use throughout the entire learning trajectory. As noted earlier the vast majority of children's early word learning occurs within a short time-scale (Benedict, 1979; Carey, 1978; Goldfield & Reznick, 1990), which complicates observation of their early word use. Therefore, we focus on protracted developmental domains—specifically number and kinship, and use these models not only to provide implemented theories but also as explanatory vehicles through which we can understand the problem of conceptual development. Formal models of development offer an explanatory framework for the patterns of behavior over the course of learning and for the final state. Using developmental models, questions about the information that is stored by our concepts can be redefined as questions of learning mechanisms, the environmental availability of data, the utility of data to the learner and inductive biases. Learning mechanisms constrain the kinds of information learners can abstract away from data. The environmental availability of data determines the regularities and structures in the world that could be inferred. The utility of data to the learner—i.e., both their interest in the data and the perceived value in learning the structure behind the data for their current goals, shapes what conceptual structures learners build and how learners approach future learning problems. Inductive biases guide how we make inductive leaps and construct novel conceptual systems. For example, simple explanations are preferred over complex ones (Lombrozo, 2007; Bonawitz & Lombrozo, 2012).

In addition to questions of how conceptual development may guide early word use, the development of the concept-language interface allows us to address many important theoretical questions in cognitive science broadly: What are the trade-offs between nativism and empiricism? What are the trade-offs between maturation and learning? How do we tease apart emerging competence (i.e., what children actually know) from performance issues (i.e., how they use their knowledge)? Where appropriate, the culturally and developmentally informed computational models developed in this thesis address these questions. To foreshadow, these models place important constraints on how and if these problems can be solved (Chapter 5), serve as data analysis tools to provide convergent evidence in support of empirical findings (Chapter 3), and make precise empirical predictions about children's word use (Chapters 2 and 4).

In Chapter 2, 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 it's learning data using cross-linguistic simulations from English, Pukapuka, Turkish and Yanomamö. More importantly, the behavioral patterns observed in children learning kinship terms, under-extension and over-generalization (Benson & Anglin, 1987; Haviland & Clark, 1974), 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 with a discussion of how this model framework generalizes beyond kinship terms and the limitations of our model.

In Chapter 3, we fill a large gap in the work on models of development by investigating the linking hypothesis between the amount of data children use and the time in which they acquire word meanings. Developmental theories posit that word learning is delayed by "maturational processes" as well as by the "poverty of stimulus" in the environment (Newport, 1990). Experimental results reveal that children can rapidly learn words after a single exposure (Carey & Bartlett, 1978; Heibeck & Markman, 1987; Markson & Bloom, 1997; Spiegel & Halberda, 2011) as well as by aggregating ambiguous information across multiple situations (Smith & Yu, 2008). The challenge of this chapter was to quantify the trade-off between maturational and data-driven processes in word learning by inferring the profile of children's data usage while learning words. Perhaps the most under-appreciated consequence of any account of learning is that a learner must wait for data. Chapter 3 models the acquisition of children's earliest learned words (as measured by cross-sectional parental reports spanning 14 languages; Frank, Braginsky, Yurovsky, & Marchman, 2015) using waiting time models from survival analysis (following Hidaka, 2013). Compared to standard and theoretically informed baseline models, our waiting time model better explains and predicts children's word acquisition. More importantly, our model parameters are interpretable under a generative process reflecting a maturational delay before children start attending to data and a period of data-driven learning. We find that the majority of variance and total time in word learning is explained by data-driven processes as compared to maturational processes. These findings suggest that, despite empirical evidence showing that children can learn words from a single instance and computational models suggesting words require on the order of hundreds or thousands of instances, the typical early-learned word involves keeping track of information across on the order of ten informative instances.

In Chapter 4, we leverage the power of the inductive learning models described in Chapter 2 and the waiting time models investigated in Chapter 3 to learn about learning, specifically quantifying the influence of culture by explaining cross-cultural differences in exact number word learning. We chart out the culturally-specific and universal influences on the acquisition of exact number words in a Bayesian data analysis. Like kinship terms, the acquisition of exact number words has a robust developmental trajectory (Wynn, 1990, 1992); however, as anthropologists are keen to point out, the timing of number word learning varies across cultures (Barner, Libenson, Cheung, & Takasaki, 2009; Sarnecka, Kamenskaya, Yamana, Ogura, &

Yudovina, 2007). The rampant diversity in both the mathematical problems that different cultures face (e.g., basket weaving, tailoring clothes, financial transactions) and the algorithmic solutions cultures adopt to solve them (e.g., Saxe, 1988a) makes number the ideal domain to look at cultural influences on conceptual/semantic acquisition. Besides obvious potential differences in the amount of environmental input, differences in mathematical goals could potentially shape how learners combine conceptual primitives and the distribution of types of learning instances. We compiled a large ($N > 1700$ children) eight-culture dataset of children's number acquisition to infer universal and culturally-specific influences on the learning process. With the largest dataset aggregated to date, we find strong evidence for an influence of culture both on how learners combine conceptual primitives to learn exact number words and on how frequently children experience effective learning instances. Importantly, a universal bias for simplicity underlies hypothesis construction, in line with most models of conceptual development. Further, the rate of effective learning instances are on the same order of magnitude across cultures seen in Chapter 3, which hints at universal constraints on how children use data.

In Chapter 5, we zoom out from focusing on word learning to evaluate whether such a powerful inductive learning mechanism is warranted for the task of lexical development. The majority of work on language acquisition has focused on sophisticated learning mechanisms required to acquire syntactic or phonetic information (Chater & Vitányi, 2007; N. H. Feldman, Griffiths, Goldwater, & Morgan, 2013; Goldwater, Griffiths, & Johnson, 2009; Frank & Tenenbaum, 2011; Perfors, Tenenbaum, & Regier, 2011). How does the amount of information children must store about lexical semantics scale in comparison? Chapter 5 adopts Fermi-estimation and information theoretic techniques to quantify the amount of information about English linguistic representations a learner must store to use English. Given the vast work on the learnability of syntax, it is surprising that we find the majority of information stored about language is allocated for lexical semantics. We see this as support for a sophisticated learning mechanism for lexical semantics, as outlined in the previous chapters.

In the last chapter, we summarise the findings of this work, contextualize the contribution in terms of the literature and highlight future directions.

Chapter 2

Logical word learning: The case of kinship

In order to acquire a language, learners have to map words to objects and situations in the world. From these mappings, they must then learn the underlying concept of the word that will generalize to new objects and situations. The mappings between words and concepts, acquired over a lifetime, will constitute the majority of information a language user stores about linguistic representations (Mollica & Piantadosi, 2019). While there is a vast literature on how children might solve the problem of mapping words to the world (e.g., Carey & Bartlett, 1978; Smith & Yu, 2008; Frank, Goodman, & Tenenbaum, 2009; Medina, Snedeker, Trueswell, & Gleitman, 2011), we know less about how children use these mappings to inform their concepts in order to generalize words to new contexts. Research on children's early word generalization has focused on uncovering biases in children's generalizations (e.g., shape, Landau, Smith, & Jones, 1988) and explaining the mechanism and types of input children need to overcome these biases (e.g., Gentner & Namy, 1999; Graham, Namy, Gentner, & Meagher, 2010); however, research has yet to precisely predict children's behavior across the developmental trajectory. Inspired by the recent findings that children maintain and update a limited number of hypotheses about a word's meaning over the course of development (Medina et al., 2011; Yurovsky & Frank, 2015), we propose a theoretical model from first principles, to scale up our understanding of how children's word meanings should change as they observe more data. In the process, we demonstrate that several seemingly unrelated patterns in children's early word use can be explained by the process of induction in naturalistic learning contexts.

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

While these behavioral patterns are consistently observed in children's early word use, it's unclear whether they reflect partial conceptual knowledge (E. V. 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)? The model we develop here will serve as a baseline for future research to tease apart performance from competence in children's early word use.

In this paper, we describe a rational constructivist framework (Xu, 2007, 2016, in press) of conceptual development formalized as logical program induction. We evaluate our framework against the literature on children's patterns of generalization over time, specifically under-extension, over-generalization and the characteristic-to-defining shift. For demonstrative purposes, we implement a model based on this framework to learn kinship terms, providing the first computational learning model for kinship term acquisition. The paper is organized as follows: First, we review the empirical literature on kinship term acquisition and computational models of kinship. We then flesh out our model framework and implementation. In presenting the results, we first demonstrate that the model is powerful enough to learn any kinship system consistent with its input data. We then provide simulations based on informant provided learning contexts to show that the general patterns of children's word use described above fall out naturally from framing conceptual development as program induction in naturalistic environments. In the process, we present evidence suggesting that children's early word use might be informative about conceptual development and derive a novel account of the characteristic-to-defining shift. To demonstrate how this model can be used to entertain important theoretical questions about how inductive biases and children's input drive children's behavior, we examine the roles of simplicity and environmental input in determining the order of kinship term acquisition. Lastly, we conclude with a discussion of novel predictions and limitations of our account.

2.1 Children's Acquisition of Kinship Terms

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

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

A second line of kinship research lies at the merger of componential analysis in anthropology (Goodenough, 1956) and the semantic feature hypothesis for word learning proposed by E. V. Clark (1973). Componential analysis takes up the task of identifying the minimal set of features required to distinguish relevant distinctions in meaning. For example, gender is a required feature of the English kin system because gender is required to distinguish, for instance, MOTHER from FATHER. The semantic feature hypothesis posits that children acquire the semantics of a concept "component-by-component" (E. V. Clark, 1973). Thus, developmental studies of kinship acquisition could inform theoretical anthropological studies of componential analysis, especially when multiple sets of components are equally as expressive. As Greenfield and Childs (1977) points out, the pattern of children's mistakes in an elicitation task is informative about the actual features of meaning children have acquired. For example, 4-5 year old Zinacantan children's mistakes never violate the feature that siblings have common parentage; however, half of their mistakes violate gender. Whereas, 8-10 year olds never violate common parentage and gender, but violate relative age. Therefore, componential analyses that include features for common parentage and gender are more likely than componential analyses that do not. For our purposes, the developmental evaluation of componential analyses potentially highlights the dimensions on which children might generalize.

The semantic feature hypothesis has also been used to predict the order of acquisition of kinship terms. Haviland and Clark (1974) proposed and found evidence for simplicity to be a driving force in the order of acquisition of English kinship concepts.

In their analysis, a relationship between two individuals was considered to be one feature. Relations that could be explained 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 by differences in experience with relatives and input frequency of kinship terms rather than simplicity. While experience seems to explain differences in adopted children, at least one study has found no effect of household size on kinship acquisition (Price-Williams et al., 1977). In general, the extent to which simplicity and experience contribute to the order of acquisition of kinship terms is still an open question. We return to this question in our analysis of order-of-acquisition effects from model simulations.

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

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

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

2.2 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 an ideal test-bed for how inter-related conceptual systems are learned, as adult kinship knowledge suggests inter-related, not independent, concepts for kinship terms. That being said, most of the previous computational models of kinship had other motivations.

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

Early connectionist models have used learning kinship as a test case for distributed models of relational concepts. Hinton et al. (1986)'s family tree task focused on learning an encoding for the family members on a given tree and the relationships between them. The connectionist model received input vectors reflecting an individual on the tree (e.g., *Simba*) and a kinship relationship (e.g., *father*) and output the individuals on the tree who completed the kin relation (e.g., *Mufasa*). The model learned interpretable embeddings for people on the tree, such that semantic features (e.g., gender) could be easily extracted. However, the relationship embeddings were not interpretable and the generalization performance of the model was poor. Using linear relational embedding, Paccanaro and Hinton (2001) greatly improved the generalization performance. Their model learned relationships as rotational transfers between point vectors in a space reflecting individuals. The model was successful at completing the implicit structure behind the training data especially when incorporating held out people into the system; however, the model did not fare as well when incorporating held out relations to the model. The model learns the family members and all of the relations on the tree without learning the actual tree structure. Therefore, it's unclear how well the relations learned will generalize to an entirely new family tree. Importantly, neither distributed model makes any claims about children's behavior while learning. Though, Paccanaro and Hinton (2001) did point out that the most common generalization error was over-extension of sibling terms to include the speaker—i.e., the common failure of Piaget (1928)'s logic problem.

More recent computational models have approached the acquisition of kinship knowledge through a Bayesian relational-learning or theory-learning perspective. The Infinite Relational Model (IRM; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006) uses the presence or absence of relations between individuals and kinship term use to learn groupings of these individuals and properties shared by the groups, which are diagnostic of the relationship. For example, applying the IRM to data from a complex Australian kinship system results in groups of individuals that share "diagnostic" kinship relevant feature dimensions such as age and gender. Katz, Goodman, Kersting, Kemp, and Tenenbaum (2008) proposed a generative model similar to the IRM but with a richer representation system based in first order logic, Horn Clause Theories. Their model learns each individual's kinship relevant properties and the abstract rule governing how those properties give rise to the kinship relation. Katz et al. (2008)'s representation scheme has two advantages over the IRM. First, Horn Clause Theories are compressible probability models that license deductive inference, inductive inference and deductive inferences based on inductive inferences. Second, Horn Clause theories are context independent, which allows one's knowledge of kinship to easily generalize beyond the observed/training data. Similar first order logic representation schemes have been used to analyze the space of all possible kinship systems to identify the pressures that influence which kinship systems are extant in the world (Kemp & Regier, 2012). Surprisingly, extant kinship systems are found at the optimal trade-off between simplicity and communicative efficiency. Yet again, while these computational models of kinship provide proof of learnability, they do not make claims about children's behavior during learning.

Our model builds on the intuitions of the Bayesian models. Following Katz et al. (2008), we adopt the use of a context independent representation scheme. Our model also incorporates a pressure for simplicity, which is line with Kemp (2012) and other studies of kinship acquisition (e.g., Haviland & Clark, 1974). Our approach will depart from past models in two ways. First, our representation scheme is inspired by set theory instead of horn clauses[1], which provide poor fit to adult's induction and generalization behavior (Piantadosi, Tenenbaum, & Goodman, 2016). Operating over sets is a more functional representation scheme that emphasizes generating members of those sets, or possible word referents, as opposed to computing the truth of a logical expression. Second, we aim to provide not only a proof of learnability but an evaluation of the full developmental trajectory of concepts (illustrated here with kinship), including the the common behavioral patterns of mistakes children display. Whereas previous learning models with logic-like representation schemes have not been evaluated against behavior, we use our model to explain the patterns of children's behavior while learning these concepts, while formalizing the contributions of the data, the learning context, and inductive biases.

2.3 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

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

moving entities. In this regard, a concept's ability to denote entities is like a program that takes as input a context of potential referents and returns a set (possibly empty) of referents consistent with that concept. We formalize this metaphor by defining concept induction as probabilistic program induction (e.g., Lake, Salakhutdinov, & Tenenbaum, 2015; Piantadosi & Jacobs, 2016; N. D. 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 to other concepts (e.g., DOG and WOLF share common origin). 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)[2]. 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

[2] While Carey (2009) discusses placeholder structures in relation to conceptual change, we only mean to highlight the ability for conceptual change in these models (e.g., Piantadosi et al., 2012) without implying that all conceptual learning requires conceptual change

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., Bonawitz & Lombrozo, 2012; Johnston, Johnson, Koven, & Keil, 2016).

We flesh out our framework at the computational level of analysis (Marr, 1982) as an ideal learner model to illustrate how a rational learner might solve the problem of program induction given properties of the environment and prior inductive biases (Tenenbaum, Griffiths, & Kemp, 2006). This approach is also a rational constructivist approach in that we are looking at how data drives the construction of a program (Xu, 2007, 2016, in press). In the past decade, research in this tradition has provided rich accounts of causal learning (e.g., N. D. Goodman, Ullman, & Tenenbaum, 2011), language learning (e.g., Chater & Vitányi, 2007), number learning (Piantadosi et al., 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).

2.4 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 *whole object bias* (Markman, 1990) suggests that children should readily generate hypotheses linking novel words to the entirety of the labeled object (e.g., the whole shoe) as opposed to a particular part (e.g., the laces or aglets).

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 et al., 2009) postulates that speaker's intend to refer to an object in the context. Given an intention, people will choose a word 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. M. 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:

$$P(h|D) \propto P(h) \prod_{d \in D} P(d|h). \tag{2.1}$$

2.4.1 Hypothesis Space

Constructing the hypothesis space over possible programs involves specifying primitive 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 properties (generation, gender, co-residing adult[3]), and variables—the speaker (denoted X) and

[3] We only added co-residing adult as a primitive when modeling an Iroquois kinship system as this primitive could be constructed out of the other primitives but greatly decreased computational

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 early understanding 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.

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

search time. That being said, co-residing adults are also easily noticed by children and would serve as a strong cue for relevant genealogical relationships in some complex kinship systems.

$\text{SET} \xrightarrow{1} \text{union(SET,SET)}$	$\text{SET} \xrightarrow{1} \text{parent(SET)}$	$\text{SET} \xrightarrow{1} \text{generation0(SET)}$	$\text{SET} \xrightarrow{1} \text{male(SET)}$
$\text{SET} \xrightarrow{1} \text{intersection(SET,SET)}$	$\text{SET} \xrightarrow{1} \text{child(SET)}$	$\text{SET} \xrightarrow{1} \text{generation1(SET)}$	$\text{SET} \xrightarrow{1} \text{female(SET)}$
$\text{SET} \xrightarrow{1} \text{difference(SET,SET)}$	$\text{SET} \xrightarrow{1} \text{lateral(SET)}$	$\text{SET} \xrightarrow{1} \text{generation2(SET)}$	$\text{SET} \xrightarrow{1} \text{sameGender(SET)}$
$\text{SET} \xrightarrow{1} \text{complement(SET)}$	$\text{SET} \xrightarrow{1} \text{coreside(SET)}$	$\text{SET} \xrightarrow{\frac{1}{37}} \text{concreteReferent}$	$\text{SET} \xrightarrow{1} \text{all}$ $\quad$ $\text{SET} \xrightarrow{10} \text{X}$

Table 2.1: 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.

to augment this set at a greater level of abstraction (e.g., adding a sibling primitive). For any model of learning, the granularity and span of a hypothesis space depends on the characterization of the learning problem. For our purpose, we are not interested in how children develop their function for gender or even the family tree itself. We are focused on how one learns relations over a structure and these primitives are an appropriate set to investigate this learning problem. Our general findings will not strongly depend on any particular base function inventory; however, inventories can make different predictions about the precise pattern and timing of children's behavior over learning (see Piantadosi et al., 2016, for a method to evaluate different primitive inventories in a similar model framework). Currently there is insufficient empirical data and qualitative reports are too inconsistent to properly evaluate the precise predictions of different primitive inventories; however, we can still evaluate the coarse-grained predictions of the model. For a more detailed discussion of hypothesis spaces see Perfors (2012).

We compose the base functions using a probabilistic context free grammar (PCFG; see Table 2.1) following N. D. 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[4]. Here, we use a PCFG as a tool to generate a finite approximation to an infinite hypothesis space and not as a model of cognition. In addition to defining an infinite space, a PCFG also provides a probability distribution over that space. In this distribution, we weight each rule as equally 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).

2.4.2 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 fewer rules. We motivate this bias towards simplicity in several ways. First, adults learn simpler concepts faster than complex concepts (J. Feldman, 2003, 2000). Second, children prefer simpler explanations over more complex explanations (Lombrozo, 2007; Bonawitz & Lombrozo, 2012)–although see (Walker, Bonawitz, & Lombrozo, 2017). In language learning, simplicity has been suggested as a guiding principle (Chater & Vitányi, 2007). Further in kinship, simplicity has been proposed as the

[4] 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.

driving factor behind the order of acquisition of kinship terms (Haviland & Clark, 1974). In a global analysis of all possible kinship systems, simplicity is a good predictor of which kinship systems are actually observed in the languages of the world (Kemp & Regier, 2012). Therefore, we believe simplicity is an important inductive bias to be incorporated in our model. The prior probability of a hypothesis, h, according to our PCFG is:

$$P(h) = \prod_{r \in h} P(r), \tag{2.2}$$

where r reflects a single use of a base function following the rules in the PCFG (Table 2.1).

2.4.3 Size Principle Likelihood

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

We use a noisy size principle likelihood, which mixes two possible ways a learner might think the data were 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:

$$P(d|h) = \begin{cases} \frac{1}{|h|} & \text{if } d \in \{h\} \\ 0 & \text{else} \end{cases}, \tag{2.3}$$

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 are randomly generated. Combining both of these processes, our likelihood function is given by:

$$P(d|h) = \delta_{d \in \{h\}} \frac{\alpha}{|h|} + \frac{1-\alpha}{|\mathcal{D}|}. \tag{2.4}$$

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

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

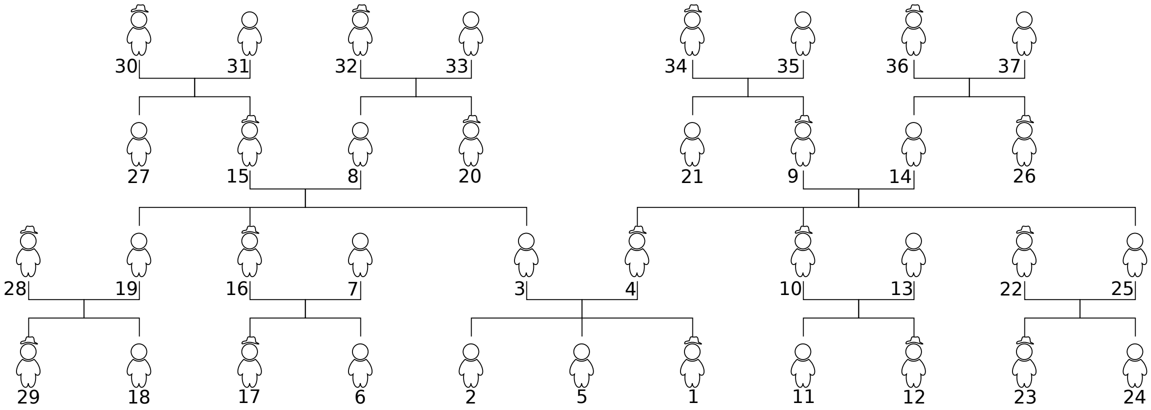


Figure 2.1: Family tree context for our simulations. Connections above figures reflect parent/child relationships. Connections under figures reflect lateral/spousal relationships. Male denoted with hats. Numbers reflect the rank order of the amount of interaction a learner (i.e, 1) has with the other individuals on the tree.

2.4.4 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 span 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 in children’s behavior reported in the literature.

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 2.1). 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.[5] In simulating the data this way, we make two simplifying assumptions. First, we are only sampling the data from one family tree and it is likely that children are exposed to multiple family trees. 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 set of potential referents is always defined with respect to the speaker.

2.5 Results

We divide the results into three sections: Model Insights, the Characteristics-to-Defining Shift and Order of Acquisition. In Model Insights, we first check that the model successfully learns the conventionally agreed upon extension for each kinship term with finite amounts of data. We conduct this analysis using four different kinship systems: Pukapukan, English, Turkish and Yanomamö. 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, naturally follows from the process of induction with few data points. We then look at how

[5] In Supplementary Figure 2.10, we emulate the simulations conducted by Navarro et al. (2012) to demonstrate that our main findings are robust under realistic values of α.

broad patterns of over-generalization fall naturally out of the process of induction when trading off simplicity and fit to the data.

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

In Order of Acquisition, we return to an open question in the kinship acquisition literature: is the order of acquisition driven by experience or the conceptual complexity of the kinship relations? We evaluate the order of English kinship acquisition predicted by the model against the empirically observed order of concept acquisition in children. We illustrate that while the simplicity of the minimal description length correct kinship concepts aligns with the observed order of acquisition in children, the model does not predict that order of acquisition. Inspired by accounts of children's experience with kin relations (Benson & Anglin, 1987), we simulate several plausible data distributions based on kin experience and find that order of acquisition is likely driven by naturalistic data distributions. In Appendix 2.A.3, we propose an alternative explanation for the order of acquisition in learning an inter-related kinship system instead of simultaneously learning independent kinship concepts.

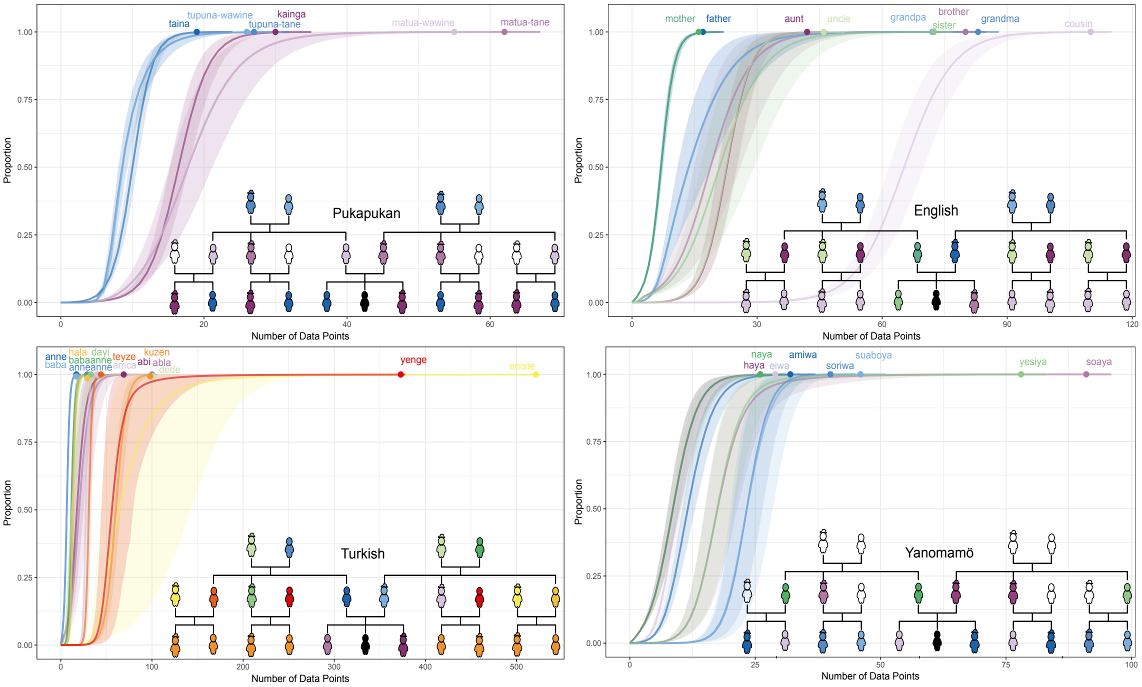


Figure 2.2: Average lexicon posterior-weighted accuracy for each word as a function of data points of that word. Shaded region denotes 95% bootstrapped confidence intervals. Insets show the color-coded extension of the terms.

2.5.1 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 four kinship systems that vary in their complexity and are common in the languages of the world: Pukapukan, English, Turkish and Yanomamö. In the tradition of Morgan (1871), Pukapukan, English, Turkish and Yanomamö are from the Hawaiian, Eskimo, Sudanese and Iroquois family of kinship systems respectively. Extensions for the kinship terms of these languages are provided in the insets of Figure 2.2 and Table 2.2. 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 with high specificity in the first generation. In addition to requiring tree moving functions, the fourteen kinship terms reflect increased specificity in referents, separating paternal and maternal brothers and sisters and their spousal relationships. Without coresidence base functions, Yanomamö is the most complex, requiring both

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

Table 2.2: The maximum-a-posterior (MAP) hypotheses after learning. For readability, hypotheses were placed into simpler extensionally-equivalent forms. F:father, M:mother, P:parent, S:son, D:daughter, C:child, B:brother, Z:sister, G:sibling, H:husband, W:wife, E:spouse

tree and set functions to specify cross- and parallel- cousins.

Figure 2.2 shows the predicted learning curves for each kinship term in Pukapuka, English, Turkish and Yanomamö. The x-axis shows the number of data points for each word observed by the child. Note the differences in scale across languages. The y-axis is the probability that a learner has acquired the conventionally-aligned upon meaning of that term–i.e., extends the term appropriately. The shaded region represents the 95% bootstrapped confidence interval. The line for each word is color coded to match the word's extension in the inset. Table 2.2 provides the maximum-a-posteriori hypotheses learned for each kinship term.

Despite varying 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 for each word learned. We discuss the differences between this model's predicted acquisition order and children's empirical order for English in the Order of Acquisition section. Unfortunately, we could not find empirical data for the order of acquisition of Pukapukan, Turkish and Yanomamö kinship terms.

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 (E. V. Clark, 1973; Kay & Anglin, 1982). This pattern naturally falls out of our model's attempt to explain the data when there are few unique data points, suggesting that the

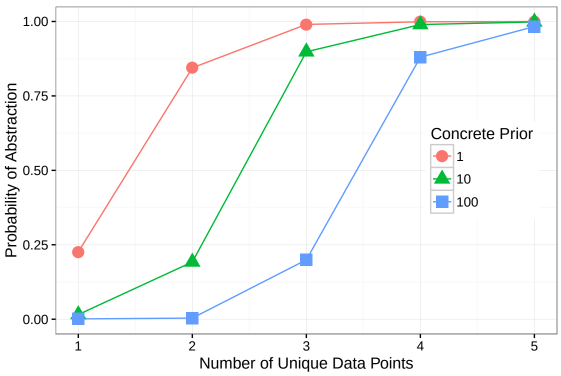


Figure 2.3: Probability of using abstraction as a function of unique data points at several different prior strengths for concrete reference. At higher prior values of concrete reference, the rise in the probability of abstraction is shifted to require more unique data points.

preference for using concrete reference is driven by the data observed rather than by inductive biases of the model. To look at the model's preference for concrete reference, we highlight a single concept, UNCLE, and focus on the first five unique data points that the model observes (see Figure 2.3). The x-axis in Figure 2.3 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 with small amounts of data, 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 too 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 choice of referential form 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 2.3, 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.

The model predicts over-extension as seen in children

While older children embrace abstraction, the rules they learn often over-extend a word to include incorrect referents (E. V. 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 2.4 shows the model's predicted pattern of use for the term *uncle*

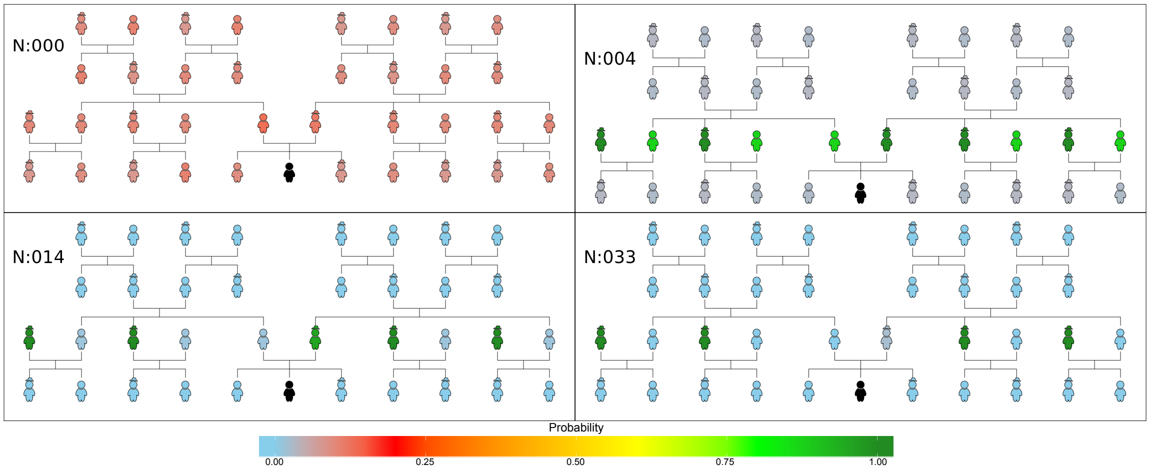


Figure 2.4: The posterior probability that each person on the tree is an uncle of the learner (in black) at various data amounts. Red indicates high probability and blue indicated low probability.

conditioned on a learner, represented in black. Having observed few data points, 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 the time the model has encountered 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 incorrect predictions if the context was to vary. In fact, the model does not 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. A noisy size principle likelihood posits that it is better to predict additional, unseen data than to

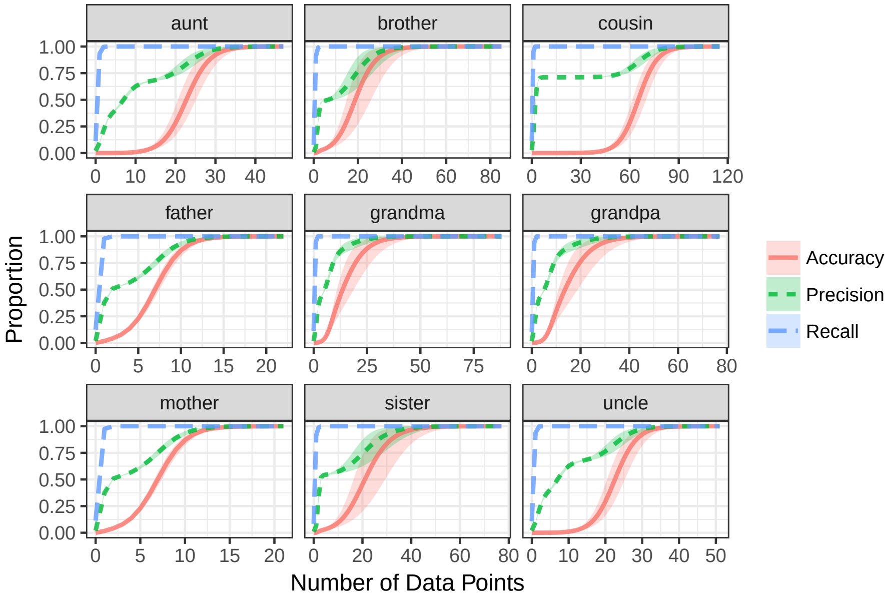


Figure 2.5: Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Recall greater than precision is a hallmark of overgeneralization. Shaded regions represent 95% bootstrapped confidence intervals.

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

For a bird's eye view of over-extension in the model, we can compare the model's posterior weighted *recall* and *precision.* Recall is the probability of comprehending a word when it is used correctly. With a wide enough hypothesis, a learner will accept all of the correct uses of a word; however, 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 all and only 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 2.5 illustrates this signature pattern of over-extension for each word in English[6].

2.5.2 The Characteristic-to-Defining Shift

As introduced earlier, the characteristic-to-defining shift is a prevalent pattern of children's over-extension. Young children are more likely to over-extend using characteristic features (e.g., robbers are *mean*) as opposed to defining features (e.g., robbers *take things*). While the characteristic-to-defining shift is commonly observed in concept acquisition, the process by which this occurs is unclear. One possibility is that the characteristic-to-defining shift is a stage-like transition that occurs in the

[6] Appendix 2.A.2 contains F_1 score plots for every language and context simulated in this paper.

representational system (Werner, 1948; Bruner et al., 1966). For example, the shift could be explained by a transition from representing concepts holistically—i.e., using all the features of objects, to representing concepts analytically—i.e., narrowing the representation to include only specific relevant features of objects (Kemler, 1983). Neural network models of conceptual classification and their exentsions inherently capitalize on this idea when demonstrating a shift (e.g., Shultz et al., 2008; Doumas et al., 2008). Another possibility is that there is a change in the mechanism by which one learns concepts. For example, concept learning might change from storing exemplars to constructing prototype or rule-based representations. These hypothetical changes in representation or processing might be maturational in nature, such as the development of abstraction (Piaget & Inhelder, 1969). Alternately, they may be driven by inductive inference mechanisms operating over observed data, as in rational constructivist accounts (Xu, 2007, 2016, in press).

From the outset we can narrow down this space of theoretical hypotheses. The conceptual to defining shift is most likely a function of data, not maturation (Keil, 1983). One prediction of a maturational-shift is that at a single time-point, children should represent all words using characteristic features or defining features, whereas a data-driven shift predicts that both adults and children should have more characteristic-based representations in unfamiliar domains, and more rule-based representations in familiar domains. The maturational and shift hypotheses do not explain children's behavior—children seem to possess characteristic representations and defining representations of different words at a single time point. In contrast, the prediction of the data-driven hypothesis, namely that individuals have more characteristic-based representations in unfamiliar domains and more rule-based representations in familiar domains, is what we observe both in children (Chi, 1985) and in adults (Chi, Feltovich, & Glaser, 1981).

All of the aforementioned explanations for the characteristic-to-defining shift require a discrete shift in representation or process. However, it is unclear whether a representational or mechanistic shift is entirely warranted. To date, no model has tested whether a characteristic-to-defining shift could be a natural by-product of the continuous data-driven construction of concepts. Here, we illustrate that the characteristic-to-defining shift could emerge 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. If, however, the characteristic features are uninformative, and thus poor at capturing a concept, our model should favor defining hypotheses, predicting either no shift or an implausibly rapid shift from characteristic to defining hypotheses. The extent to which individual features apply beyond the learner's family tree context will also influence its utility in explaining the shift. As a result, the feature landscape across contexts could influence

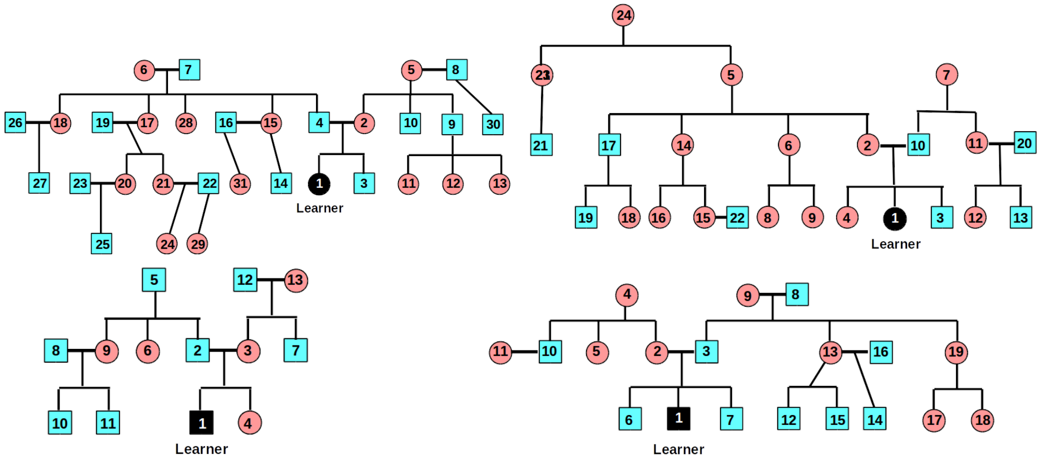


Figure 2.6: Distance-ranked family trees from informants. Circles represent females; squares males. Bold lateral lines denote spousal relationships. Informant 1 (top left) provided 107 unique features; Informant 2 (top right) 88; Informant 3 (bottom left) 92; and Informant 4, 59.

the timing of both the shift and acquisition of the term. For example, if characteristic features explain the learner's data equally as well as defining features, a learner would require more disambiguating data points before learning the term than if they hadn't considered any characteristic hypotheses. Similarly, if the characteristic features that best explain family relations on a learner's own tree apply broadly to individuals outside the family context, the features are not informative enough and the characteristic-to-defining shift should occur sooner. Therefore, it is crucial that we collect data about the characteristic and logical relationships of real people to test if natural data will contain features within the range of informativity that will show a characteristic-to-defining shift.

We asked informants to provide us with information about their family trees. Four informants, who were blind to the purpose of the task, drew their family tree,

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 2.3: Additional rules for the PCFG in Table 2.1. Now, each hypothesis starts with a START symbol.

ranked each family member in terms of how frequently they interacted with them as a child (see Figure 2.6), 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*)[7]. We used these features to augment the hypothesis space with the rules in Table 2.3. One limitation of our design is that across feature matrices there was no requirement for shared features. In our matrices, there is little overlap in the solicited features, which prevents us from simulating data for a learner from other contexts. The main consequence for our analysis is that we can only predict the upper limit for the number of data points required to observe a shift as features applying more broadly or incorrectly across contexts would hasten the shift. Features applying less broadly or correctly across contexts would not introduce a bias.

The informant provided contexts are smaller/sparser than the context used in

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

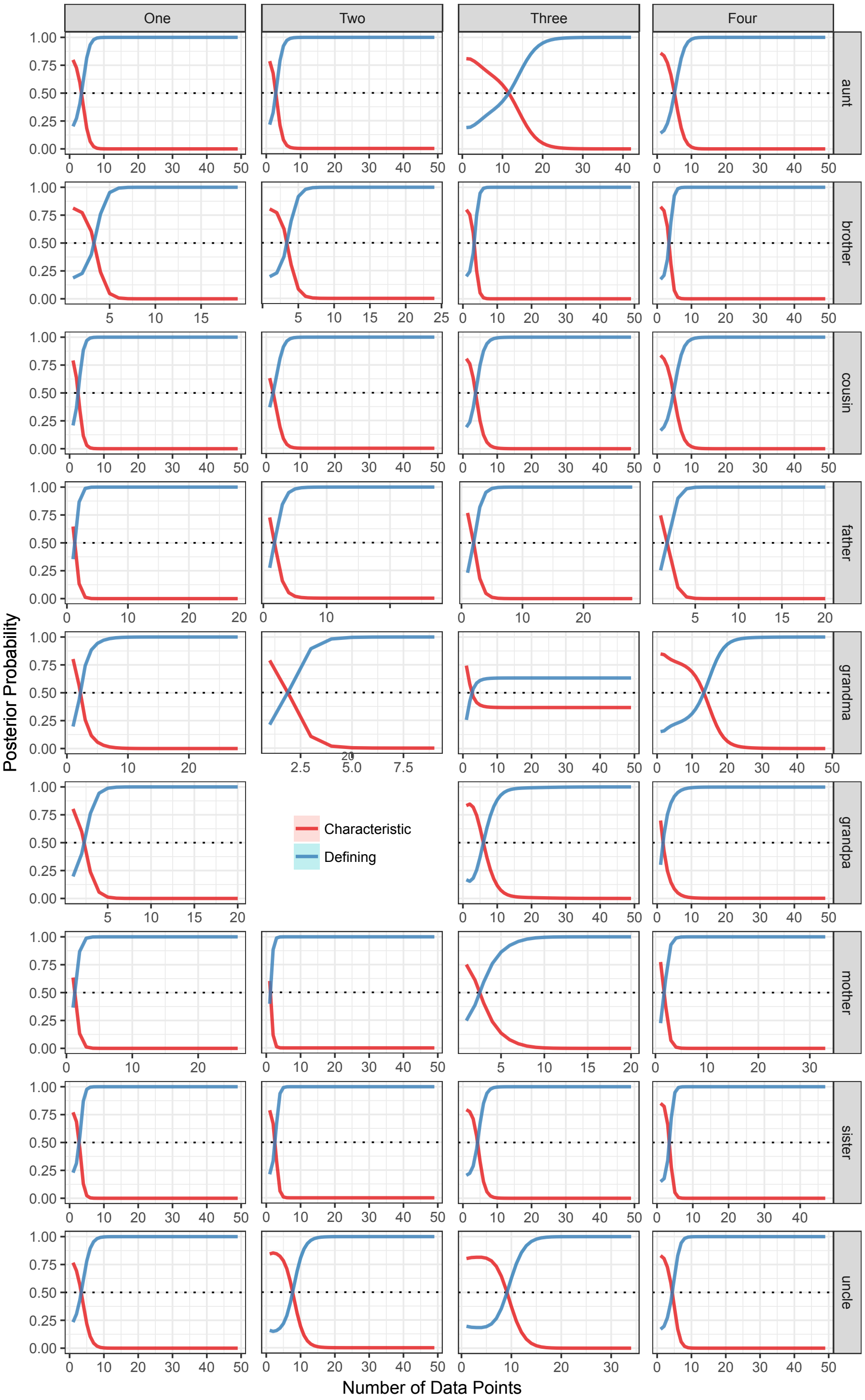


Figure 2.7: Average posterior probability of using a characteristic or a defining hypothesis (y-axis) as a function of the amount of data observed (x-axis) for words (rows) and informants (columns). Shaded regions reflect 95% bootstrapped confidence intervals. For all words, there is a characteristic-to-defining shift.

our previous analyses (Figure 2.1). Consequently, the types of data points the model is given in our informant analyses are restricted to a subspace of all possible types of data points, which could impede learning. The model could accommodate for this limitation by sampling across multiple contexts; however, this is computationally expensive to do for each of our informants. For computational efficiency, we only sample data for each informant within their context, which does not influence our ability to observe a characteristic-to-defining shift. That being said, the impoverished data/context sometimes prohibits the model from learning the conventionally-aligned upon extension of a kinship term. Nonetheless, the model does always learn a program that selects the individuals consistent with the observed data. In Appendix 2.A.2, we provide F_1 plots for all informants and English kinship terms, and discuss the situations in which the model does not learn the "correct" concept for a kin term. Our failure to learn all terms from these simulations suggest that egocentric kinship data is not always sufficient for learning kinship terms.

To visualize the characteristic-to-defining shift (Figure 2.7), 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[8], we observe the characteristic-to-defining shift–i.e., the probability of entertaining a characteristic hypothesis is initially greater than the probability of entertaining a defining hypothesis. This means that a simple conceptual learning model shows a characteristic-to-defining shift purely due to the learning context–i..e, realistic data about logical relations and characteristic features. As these graphs average over the exact data

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

points a learner observes, they hide the early preference for concrete referents; however, when plotted in terms of unique data points the early preference for concrete referents holds.

It's important to note that our model does not have a discrete change in processing or representation as appealed to by previous accounts (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 2.3). Therefore, Piaget and Inhelder (1969)'s explanation, that the characteristic-to-defining shift reflects the development of abstraction, is not supported. With a precise, formal model of conceptual development like ours, it is possible to demonstrate that a rational learner would still undergo a characteristic-to-defining shift even if they had perfect access to the data and the ability to abstract from the outset of learning.

Compared to previous accounts of the characteristic-to-defining shift, our model proposes a new explanation: characteristic features are useful because they are simple and explain children's initial data well. As children observe more data, children can justify more complex defining hypotheses but only 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 using defining features.

2.5.3 Order of Acquisition: Simplicity and Data Distributions

Previous research has found that English-speaking American children tend to acquire kinship terms in a specific order: mother/father, brother/sister, grandpa/grandma, aunt/uncle and cousin. Haviland and Clark (1974) first explained this in terms of simplicity, measured as the number of predicates in first order logic required to define the kinship term. They later revised their account to additionally penalize reusing the same relational predicate (e.g., [X PARENT A][A PARENT Y] is more complicated than [X PARENT A][A CHILD Y]). Other researchers have argued that data and the environment drive the order of kinship term acquisition. Benson and Anglin (1987) had parents rank order how frequently children spend time with, hear about or talk about twelve different kinship terms. They found that children's experience with different kinship relations correlated with their observed order of acquisition. The extent to which simplicity, as opposed to experience, drives the order of acquisition of kinship terms is an open theoretical question. In our model, we can directly pit experience against simplicity and evaluate these theoretical hypotheses. In Appendix 2.A.3, we propose an additional possibility for the observed order of acquisition.

In this section, we will explore how different data distributions and inductive biases about the environment influence the order of acquisition of kinship terms. For each analysis, we simulate 1000 data sets from the tree in Figure 2.1 and run the learning model with only the base primitives to measure the probability that kinship terms are acquired in a specific order[9]. There are four patterns that we

[9] We return to the simulated tree for practical convenience and because the sparseness of the solicited trees lead to incomplete learning of kinship terms (see Appendix 2.A.2). While the presence

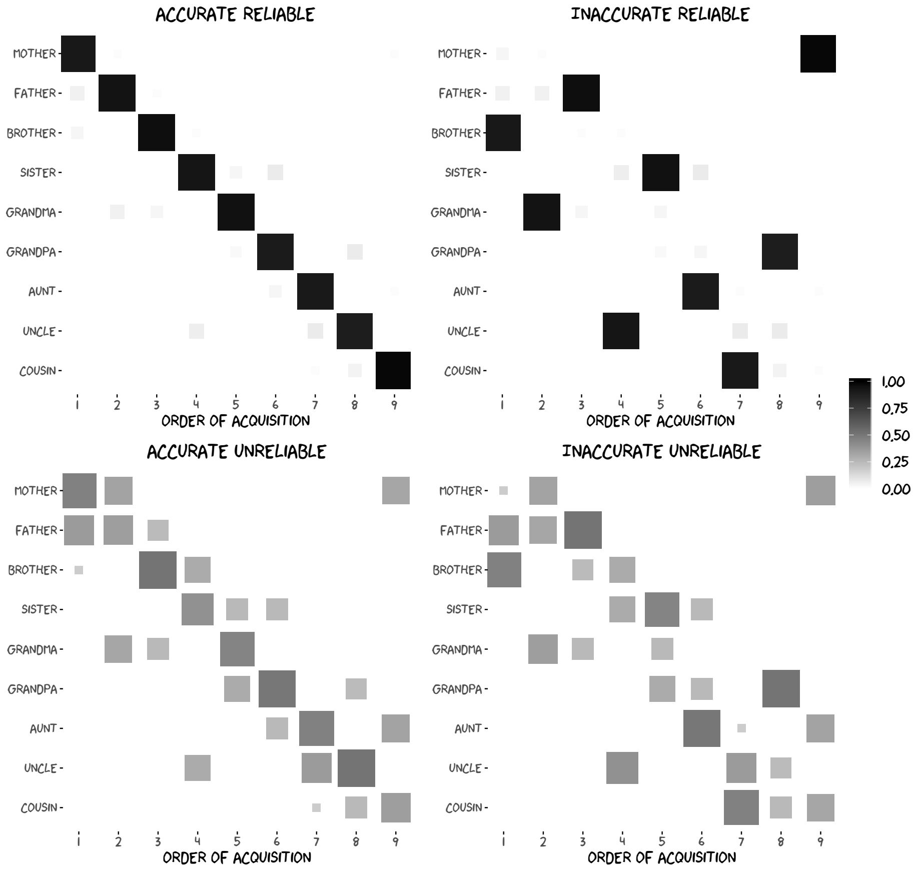


Figure 2.8: Possible patterns of order of acquisition. The x-axis reflects the ordinal position of acquisition. The y-axis represents each word. The shading on the tiles are filled according to the probability of acquisition. Words that have zero probability at a given ordinal position are omitted.

might see with these simulations (illustrated in Figure 2.8): 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 2.8 reflects the ordinal position in which

of characteristic features has the potential to influence the order of acquisition, our analyses in Figure 2.7 suggests the shift would occur before any of the terms are learned and, thus, have little to no effect on the order of acquisition.

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 2.4: 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.

words were learned. The shading reflects the probability that a word was acquired at that time. If the order of acquisition is reliable, there should be only one probable word acquired at each ordinal position (top panels of Figure 2.8). Whereas, if the order of acquisition is unreliable, there should be several probable words at each ordinal position (bottom panels of Figure 2.8).

Our initial simplicity prior (i.e., the PCFG in Table 2.1) mostly aligns with Haviland and Clark (1974)'s original formulation, as seen in Table 2.4. 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. The top left panel of Figure 2.9 shows the order of acquisition for the model given 1000 different data sets from the environmental distribution based on CHILDES frequencies and our simplicity prior. As expected, the model does not predict the empirical order of acquisition. Instead, the model is mainly uncertain about the order.

One possibility for this pattern is that CHILDES frequency estimates are not representative of children's actual input. CHILDES frequency estimates differ from the

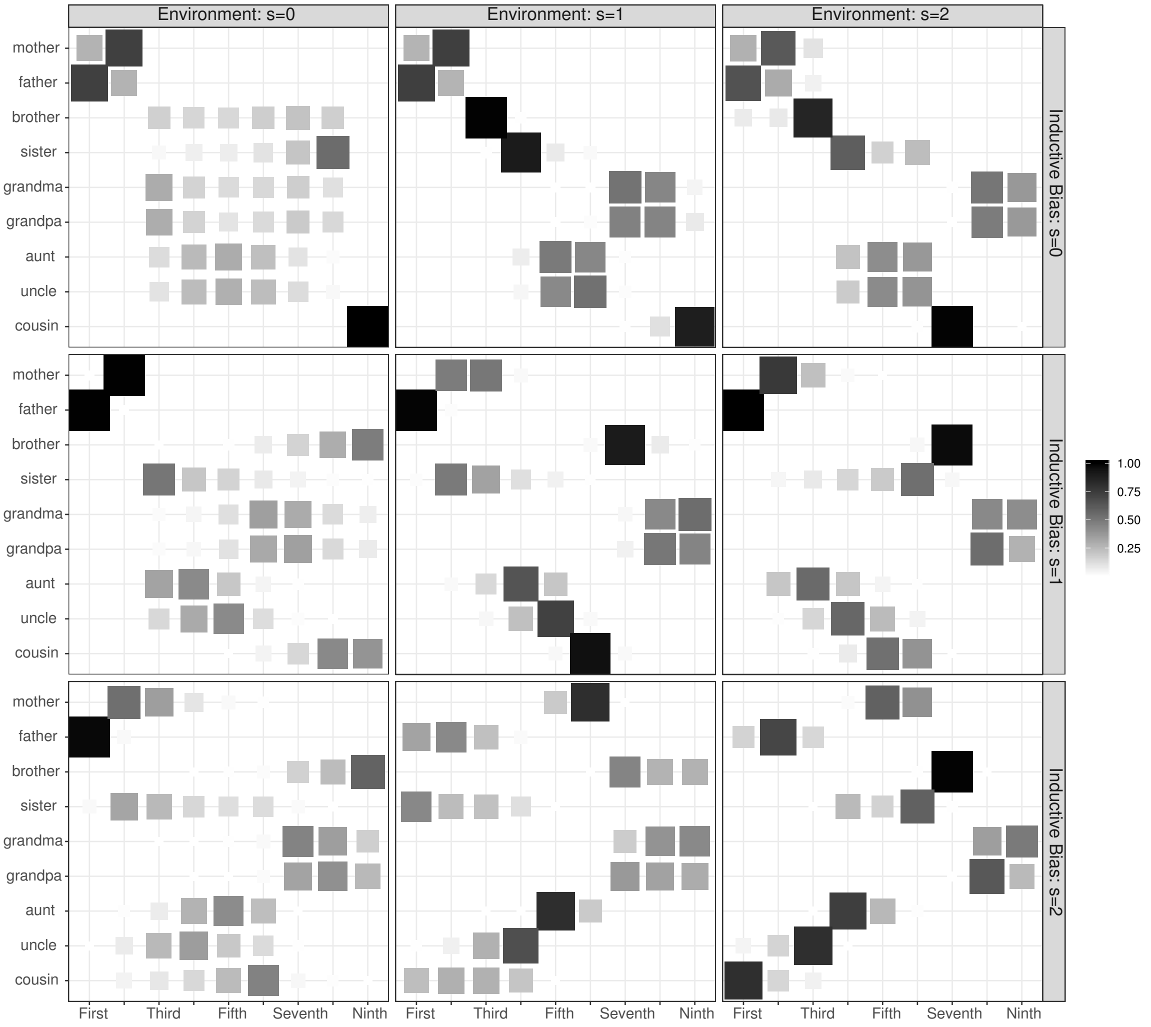


Figure 2.9: Simulations of the order of acquisition of kinship terms as a function of changes in environmental data distributions and the inductive biases of the learner. The strength of these biases are reflected in the s parameter with $s = 0$ reflecting no Zipfian bias. A tiny amount of random noise was added to probabilities in each simulation to settle ties.

surveys of Benson and Anglin (1987) and a larger corpus analysis of kinship term use across Indo-European languages (Racz & Jordan, 2017). As a larger point, children do not utilize every instance of a word in their environment as an effective learning instance (Mollica & Piantadosi, 2017a; L. R. Gleitman & Trueswell, 2018). There is evidence to suggest that children filter their input (Perkins, Feldman, & Lidz, 2017; Kidd, Piantadosi, & Aslin, 2012). To account for the discrepancy between environmental input and the latent distribution of effective learning instances utilized by a learner, we 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:

$$P(x|h,p) = \delta_{d\in\{h\}}\alpha\frac{d_x^{-s}}{\sum_{x\in h(p)} d_x^{-s}} + (1-\alpha)\frac{1}{|X|}, \tag{2.5}$$

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. This can be understood as a child expecting kinship terms to refer to people they frequently interact with as opposed to people they rarely hear about or see.

We 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 2.4.

$$P(p|w) \sim \alpha \text{ Zipf}(p|w, s) + \frac{(1-\alpha)}{|X|} \tag{2.6}$$

$$P(x|p, w) \sim \alpha \text{ Zipf}(x|w, p, s) + \frac{(1-\alpha)}{|X|} \tag{2.7}$$

For both implementations of these assumptions, the strength of the bias is modulated by the Zipfian exponent s. When $s = 0$, the data are randomly generated–i.e., no bias, and the likelihood is equivalent to a size principle likelihood. When $s \sim 1$, the environment is biased to an extent consistent with the distribution of words in English, and the learner expects to see data points reflecting this bias. When $s > 1$, the environment is heavily biased with some “black sheep” family members almost never spoken about. Similarly, the learner does not expect to see these “black sheep” family members and discounts data including them. For simulation purposes, we assigned distances to family members loosely based on Euclidean distance to the learner in the tree context (see Figure 2.1).

In Figure 2.9, 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

$s = 1$), reflecting naturalistic environments where the Zipfian exponent is ~ 1 and an unweighted size principle likelihood. The discrepancies between empirical order of acquisition and our predictions can be explained by our a-priori assignment of distances. If aunt/uncles were further from the learner than grandparents, we would expect grandparents to be acquired earlier. Differences between concepts of the same complexity (e.g., GRANDMA and GRANDPA) are slightly influenced by ties such that the alphabetical order appears dominant in Figure 2.9 where there is likely no bias. Importantly, under this Zipfian environmental distribution the model still shows under-extension, over-generalization and the characteristic-to-defining shift (Mollica et al., 2017).

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

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

Table 2.5: Summary of the empirical behavior, how the model explains this behavior and the behavioral predictions to be generated by the model.

data points as noise and it will still learns kinship terms even when there is considerable noise (see Appendix 2.A.1). Further research is needed on how exactly children filter their linguistic input.

2.6 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 illustrates both an early preference for concrete reference and patterns of over-generalization consistent with children's behavior, including the characteristic-to-defining shift. 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 learning context instead of proposing discrete changes in representation and processing. Lastly, the model has addressed open theoretical questions about the forces driving the order of acquisition of kinship terms in English and how learning an inter-related system influence children's pattern of word use (Appendix 2.A.3).

Table 2.5 outlines each behavioral phenomenon we attempted to explain, the components of the model that explain that phenomenon and the behavioral predictions from the model. There are two ways in which the behavioral predictions of our computational model can be used. First, experiments can be designed to directly assess components of the model, and the learning environment. For example, the children's patterns of generalization could be used in the tradition of componential analysis to empirically ground the primitive functions used by children. Similarly, assumptions about how children use data (i.e., the likelihood function) and the inductive biases they bring to the learning task make different predictions for patterns of generalization and the timing of those behaviors. The model also makes predictions for if and when a learning context should result in a characteristic-to-defining shift. Second, this model can be used as a baseline or normative model for comparison against other theories of conceptual learning and for the development of theories of related processes. For example, this model shows how a learner should behave if their goal were to learn the structure in the world; however, it's *possible* that learners are not trying to learn the structure in the world, but instead the conventions of lexical production through linguistic structure alone. Comparing the predictions of our model with those of formal models built to learn from linguistic structure would give us leverage to tell when, and to what extent, children are learning from world structure and from linguistic structure. Additionally, the model makes predictions of how children's competence should change as a function of data. This has the potential to aid the construction of theoretical models of pragmatic and retrieval processing in children's early word use, theoretical models of children's exploration and information extraction, and theoretical models of the other affordances of children's concepts (e.g., property induction/generalization).

It is important to highlight several links between this model approach and past approaches, links which may be connected more formally in future research. First, the model framework is compatible with similarity based approaches to early concept acquisition. For example, a program could capture 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; N. D. 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 within the data are 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.

Finally, 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 a program is learned, it can be used as a function in another program. Preliminary evidence suggests that people 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, 2016, in press), that captures the behavioral phenomena observed in kinship learning. Beyond kinship, our model derives novel 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 (e.g., 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 their past theories 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[10], 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:$5 - 30$)[11]. Although at this time, we do not provide a mechanism for how children might generate the hypothesis space, we do not mean to suggest that children will be considering the entire hypothesis space. Our goal in presenting this model is not to account for all conceptual change, but rather to provide both convergent evidence for accounts of conceptual development and a tool for predicting children's behavior as they come to wield adult like concepts.

There are many ways interesting directions for future work. For our purposes, we modelled concepts as programs that take as input all entities in a context, and return as output a subset of those entities. This formalization captures one use case for conceptual representations, i.e., extension. It's plain to see that we use our concepts

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

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

to do much more than to pick out things in the world. Conceptual representations must also support a whole taxonomy of inductive problems (Kemp & Jern, 2014), inductive reasoning (Gerstenberg & Goodman, 2012) and simulation (Ullman, Spelke, Battaglia, & Tenenbaum, 2017), including generating possible entities in the absence of context.

Our account also doesn't explain how intensional knowledge that people store about concepts is represented and the way in which people most readily access that knowledge. In fact, in our specification, a more useful end state would not be a single hypothesis always returning the conceptually-aligned upon extension, but rather a posterior over the most useful hypotheses over the course of development. For example, in addition to GRANDPA as $male(parent(parent(X)))$, a better system might also have some probability mass on $union(male, old)$. Having multiple ways to generate the extension of a concept would aid retrieval in resource intensive situations, reflecting rational meta-reasoning (Lieder & Griffiths, 2017). For example, finding grandpa in a room with one man and one woman should not require intimate knowledge of a person's family tree. In addition, having a posterior over useful hypotheses might explain typicality effects (Armstrong, Gleitman, & Gleitman, 1983), as particular entities might be more easily generated than others under a posterior, compared to strong-sampling, which would suggest that all true entities be equally as likely under a fixed hypothesis. Future work should address these possibilities.

As we have argued here, our account, while illustrated with kinship, is broadly applicable to conceptual development in other domains. That being said, certain conceptual domains will provide us with a greater opportunity to understand the inductive learning mechanism than others. We suspect the model will be most useful for conceptual domains where children's understanding unfolds gradually over time. Protracted conceptual domains, like color, locomotion, number, space and time, have measurable variance over the course of development, which allows us to use our model framework as a data analysis tool (Tauber et al., 2015). These domains are also interesting from an anthropological perspective because they show considerable cultural variation. As a result, cross-cultural use of this model framework has the potential to illustrate cross-cultural differences in inductive biases and data usage, possibly reflecting differential utility of conceptual representations across cultures.

We hope to impart two lessons learned from our model. First, programs are a powerful representational scheme to formalize concepts. Programs have the ability to capture both logical and graded/stochastic aspects of conceptual structure. When combined with data-driven learning techniques, programs not only capture the end state representation of concepts but provide rich behavioral predictions across the entire developmental trajectory, capturing phenomena like the characteristic-to-defining shift in a single model. A critical component of our program representation scheme is that our programs are functions of contexts, similar to Katz et al. (2008). Concept deployment and language use are heavily context-sensitive. To generalize across contexts, we 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 allows one to rigorously test theories and questions developmental science has put forward without committing to often necessary data analysis assumptions. For example, fundamental questions in developmental science include: What biases or abilities (e.g., simplicity, compositionality, abstraction, recursion) must be in place for children to learn X? How much data do children need to learn X? Which types of data do children find most useful for learning X? What resource limitations must be in place to explain the developmental trajectory of X? Are cross-cultural differences or differences across populations learning X caused by different biases/abilities, different data availability or different consumption/usage of data? These questions can all be addressed within our model framework through Bayesian data analyses and model comparisons. As a result, formal models of conceptual development provide important and substantial convergent evidence and insight about developmental theories, which might not be possible or, more realistically, feasible to gather from behavioral experiments/observation alone.

2.7 Methods

2.7.1 Generating the Hypothesis Space

To construct a finite lexicon space appropriate for our analyses, we utilized a variety of Markov Chain Monte-Carlo methods to draw samples from the posterior distribution over lexicons at different data amounts. Our model is implemented using the Language of Thought Library for python (Piantadosi, 2014a). 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 2.A.3, 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 (N. D. Goodman et al., 2008; Piantadosi et al., 2012). For each language, we ran 16 chains at each of 25 equally spaced data amounts between 10 and 250. Due to memory limitations, we only saved the top 100 best lexicons from each chain. For English, Pukapukan and Yanomaman lexicons, each chain was run for one million steps. For Turkish, we first ran 5 chains for three million steps on a smaller lexicon—i.e., the search did not include the three words for grandparents or the word for cousin. We then ran 5 chains for three million steps on the full lexicon. Few if any lexicons resulting from this search contained the correct hypothesis for all words; however, across all lexicons the correct hypothesis for every word was learned.

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

2.7.2 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:

$$P(\hat{h} = h|d) = \sum^{\mathcal{H}} \delta_{\hat{h}h} P(h|d). \tag{2.8}$$

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

$$P(x|d) = \sum^{\mathcal{H}} P(x \in |h|) P(h|d), \tag{2.9}$$

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

$$P(x \in |h|) = \begin{cases} 1 \text{ if } x \in |h| \\ 0 \text{ else} \end{cases}. \tag{2.10}$$

2.7.3 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:

$$P(r_{SET \to p} \in h) = \begin{cases} 1 \text{ if } r_{SET \to p} \in h \\ 0 \text{ else} \end{cases} . \tag{2.11}$$

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

$$P(r_{SET \to p}|d) = \sum^{\mathcal{H}} P(r_{SET \to p} \in h) P(h|d). \tag{2.12}$$

We manipulate the prior bias for concrete reference by changing the PCFG production probabilities given in Table 2.1, which influences the prior probability following Equation 2.2.

2.7.4 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 2.1 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 2.3 for $500,000$ steps. Due to memory limitations, we only saved the top 100 best lexicons from each chain. For each informant, the defining and characteristic hypotheses were concatenated to form a single finite hypothesis space. As our analyses collapsed over lexicons, we did not perform Gibbs sampling as above.

We replicate the learnability and F_1 analyses (described in Appendix 2.A.2) 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

$$P(r_{FSET\to\text{feature}}|d) = \sum^{\mathcal{H}} P(r_{FSET\to\text{feature}} \in h)P(h|d), \tag{2.13}$$

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

$$P(r_{FSET\to\text{feature}} \in h) = \begin{cases} 1 \text{ if } r_{FSET\to\text{feature}} \in h \\ 0 \text{ else} \end{cases}. \tag{2.14}$$

2.7.5 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.

$$w \sim \text{Multinomial}(\theta) \tag{2.15}$$

$$(x, p) \sim \text{Uniform}(|(x.p)|) \tag{2.16}$$

For the Zipfian weighted order of acquisition analyses, we assigned distances to the tree context in Figure 2.1 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 2.6 and 2.7.

For both schemes, we calculate the posterior accuracy of each hypothesis as a function of data following Equation 2.8 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 its 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.

2.A Appendix to Chapter 2

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

2.A.1 Alpha Analysis

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

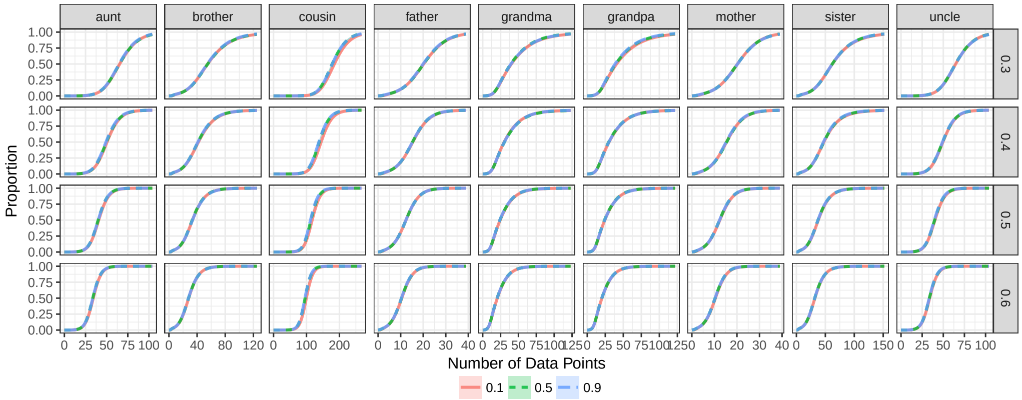


Figure 2.10: Posterior weighted accuracy (y-axis) as a function of data (x-axis) for models with different sampling assumptions (linetype and color) for different words (columns) and environmental reliability values (rows). The virtually invisible shaded regions reflect 3 standard errors of the mean.

time, and checked how distinguishable a model with different sampling assumptions would be from pure strong sampling ($\alpha = 1$). They found that in the limit of data, models with reliability parameters as low as 0.1 converge to the predictions of strong sampling. We parametrically vary the reliability of the environment by simulating data with $30 - 60\%$ reliability and set our model's sampling assumptions to either 0.1, 0.5 and 0.9 to gauge whether learning in our simulations will be robust to unreliable environments and variable sampling assumptions. As can be seen in Figure 2.10, we find no significant differences in learning across sampling assumptions and environments.

2.A.2 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:

$$\frac{\sum_{x \in \hat{h}} [x \in h]}{|h|}, \tag{2.17}$$

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:

$$\frac{\sum_{x \in \hat{h}} [x \in h]}{|\hat{h}|}. \tag{2.18}$$

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

Figure 2.11 displays the F_1 plots for Pukapuka, Turkish and Yanomamö. As shown in the main text, the model learns the correct extension for every word. As expected, the posterior weighted recall is greater than the posterior weighted precision for every word, suggesting that the model over-extends the meaning of kinship terms. Predictions for the pattern of over-extension for each word is provided in supplemental material.

The Characteristic-to-Defining Shift

Figure 2.12 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

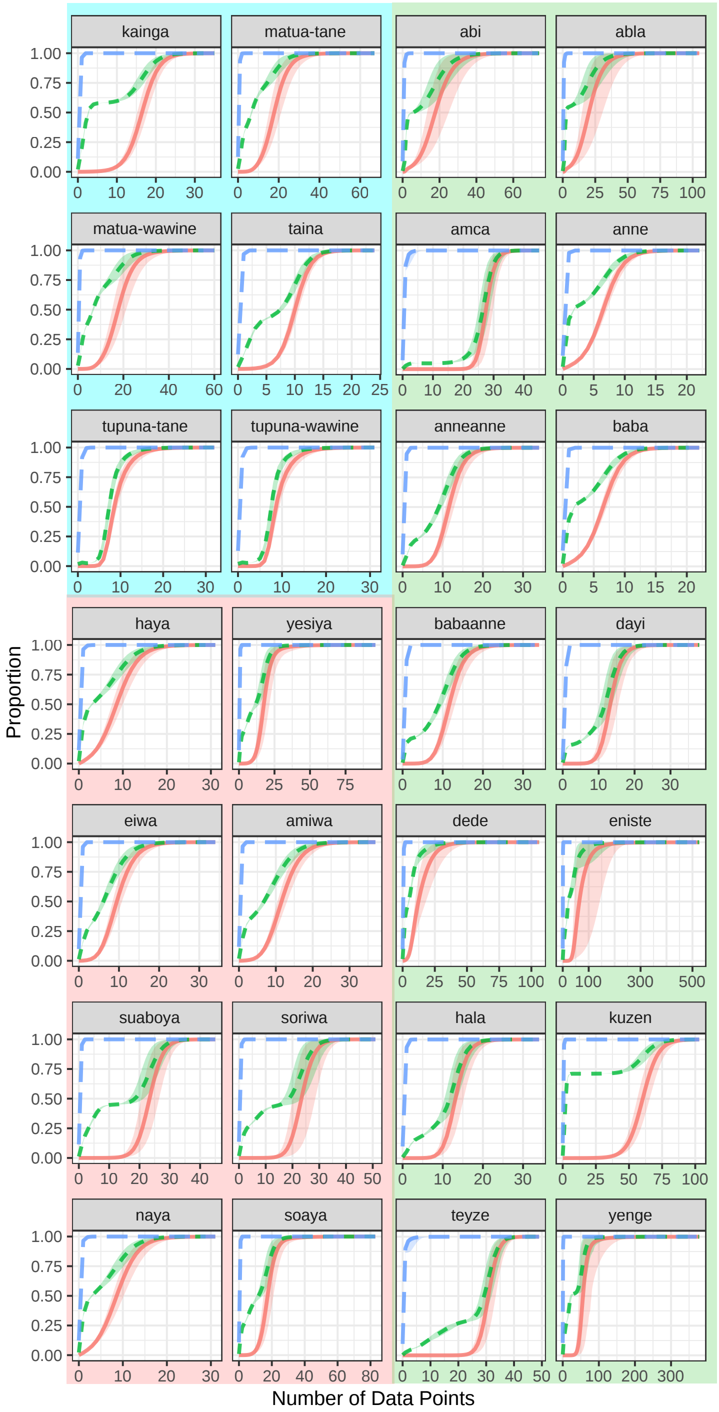


Figure 2.11: Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

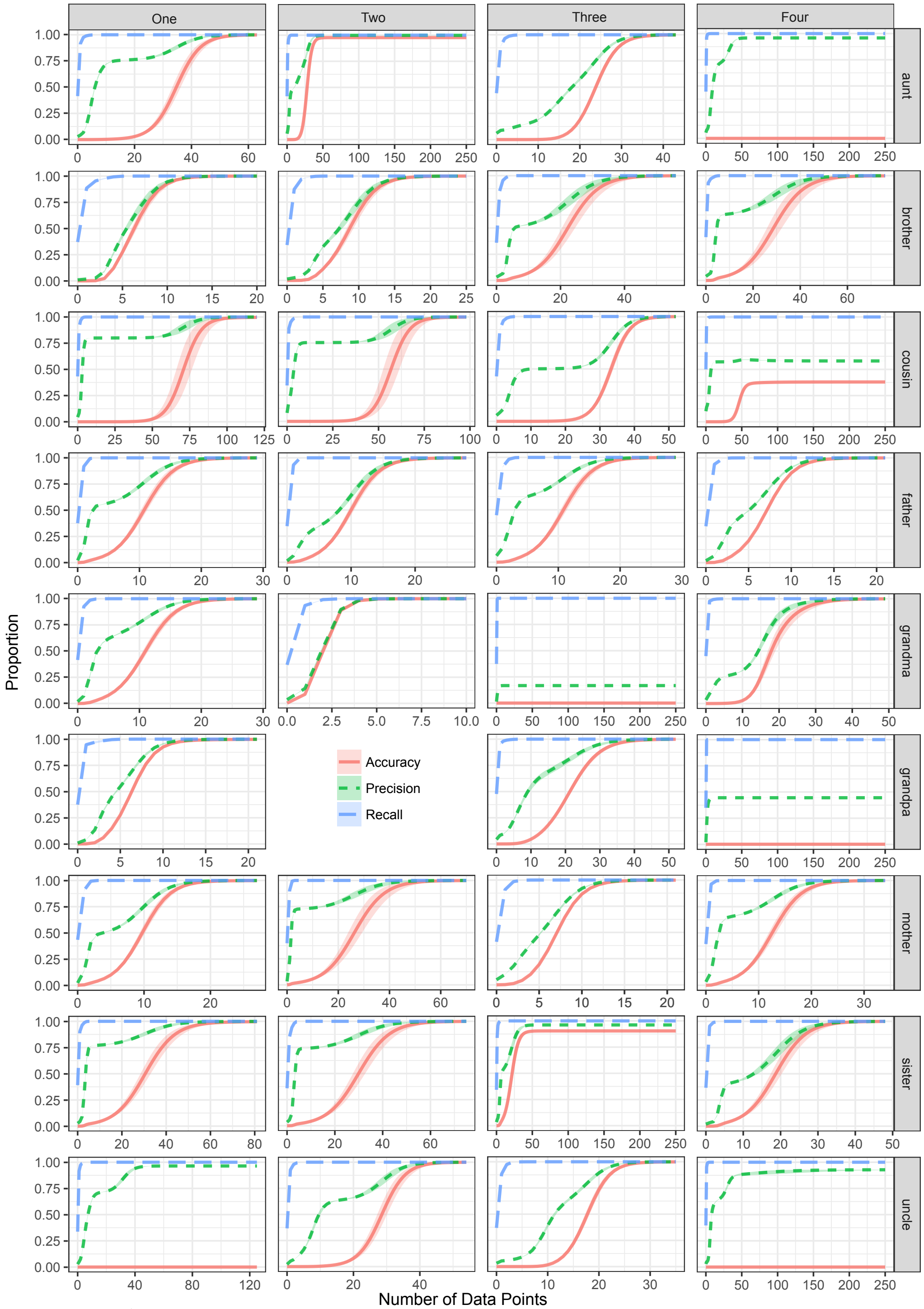


Figure 2.12: Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

being said, the model always learns a hypothesis that is consistent with its 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 2.12 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.

2.A.3 Learning an inter-related system

Learning a 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. In terms of formalization, 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:

$$P(\mathcal{L}) = \prod_{h \in \mathcal{L}} P(h). \tag{2.19}$$

The likelihood still follows a noisy size principle:

$$P(d|\mathcal{L}) = \delta_{d \in \bigcup\limits_{h \in \mathcal{L}} \{h\}} \frac{\alpha}{\sum\limits_{h \in \mathcal{L}} |h|} + \frac{(1-\alpha)}{|\mathcal{D}|^2}. \tag{2.20}$$

The main challenge in evaluating this inter-related system against developmental patterns was searching for an acceptable lexicon space—i.e., a space that contains the correct hypothesis and developmental trajectory for each word and whose lexicons contain all relevant combinations of those hypotheses. We note that this is purely a computational resource limitation. We have done our best to capture this space and note the limitations of our approximation *in situ*.

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).

We conducted the same analyses in the Model Insights and Order of Acquisition sections of the main text. The F_1 plot in Figure 2.13 illustrates the same patterns of over-generalization found when words are learned independently[12]. 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 2.14 shows the order of acquisition of the

[12] The non-monotonicity in the posterior probability of the correct hypothesis is due to the lexicon space not containing all possible combinations of likely hypotheses.

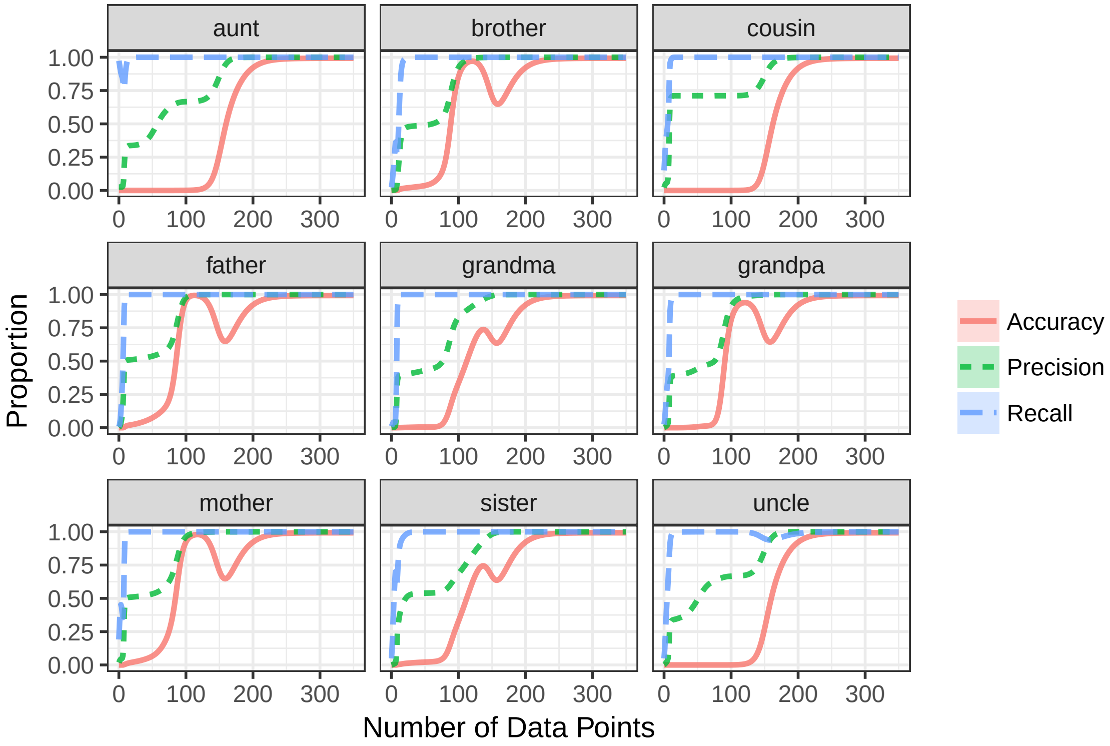


Figure 2.13: Average lexicon posterior-weighted accuracy, precision and recall for each word as a function of data points. Precision greater than recall is a hallmark of over-generalization. Shaded regions represent 95% bootstrapped confidence intervals.

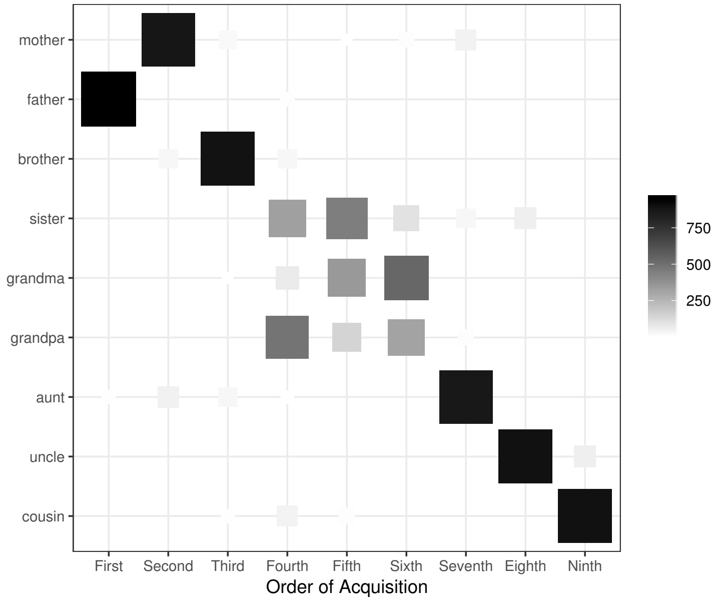


Figure 2.14: The order of acquisition under a lexicon (left).

lexicon model given 1000 different CHILDES weighted data distributions. Interestingly, the 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 2.1 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 rather than a consequence 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 inconsistent with the dominant order of acquisition in Figure 2.14.

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

Recursive calls. Another natural way to think of learning an inter-related system would be to allow for recursive calls. For example, a learner might use their current concept for BROTHER in their concept for UNCLE. We implemented this in the model but we were unable to construct an acceptable lexicon space to evaluate the model against developmental behavior. One issue with including recursive calls is that the model sometimes constructs a useful new function composition, which acts like a primitive in the hypotheses for other words but ultimately blocks a word from being acquired. Additionally, adding recursive calls exacerbates the problem of having all possible combinations of relevant hypotheses in the lexicon space by prohibiting techniques like Gibbs sampling.

Rather than directly implementing recursive calls, we attempted to capture the same intuitions by using the Lempel-Ziv compression of the lexicon in terms of

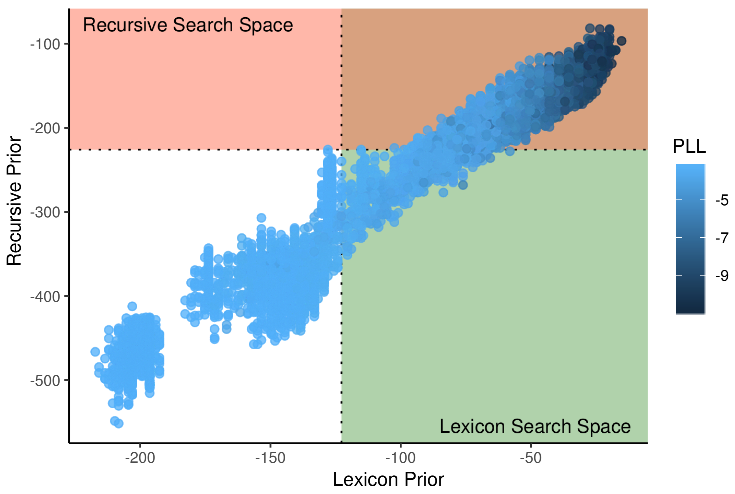


Figure 2.15: The English lexicons are plotted as a function of the recursive (compression) and lexicon prior. The color of each point represents the point log likelihood (PLL) of the lexicon. If the learner searched the space starting from the simplest to the most complex lexicon and terminated at the first correct lexicon, they would have to search a smaller space under a compression prior (red shade) than under a lexicon prior (green shade). Importantly, the developmental trajectory is not predicted under the recursive prior without additional assumptions about the complexity/development of recursion.

the grammar as a prior over lexicons. This compression prior rewards the reuse of primitive compositions across the lexicon in addition to a simplicity bias. We found that when using a compression prior, the model predicts an inductive leap from most of the kinship terms not being properly acquired to all of the kinship terms being learned. We see this leap because the correct lexicon under the compression prior is significantly less complex than the lexicons required in search space to get you there (Figure 2.15). To remove this inductive leap, we could add a parameter that penalizes recursion (as in Piantadosi et al., 2012); however, we think that the better explanation would be through the development and integration of a more cognitively grounded notion of hypothesis generation—i.e., an algorithmic level explanation.

Chapter 3

How data drives early word learning: A cross-linguistic waiting time analysis

The first year of life is an incredibly productive time for language learners. Babies discover which sounds are in their language (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), how speech is segmented (Saffran, Aslin, & Newport, 1996), what common words refer to (Bergelson & Swingley, 2012) and, towards the end of the first year, how to produce their first word (Brown, 1973; Schneider et al., 2015). This growth is a complex endeavor that requires relying on abilities in many domains—social and pragmatic understanding, conceptual representation, joint attention, and acoustic and motor systems. However, little is known about how the development of non-linguistic factors influences language growth. For instance, is the timing of language growth locked to factors like the maturation of cognitive and motor systems (e.g. memory and attention), or to the growth of children's conceptual repertoire? Or, alternatively, is early language learning primarily limited by the amount of data that children receive about language itself?

Evidence for a data-driven view of the timing of language learning comes from studies showing the importance of linguistic input for early learning (Hoff, 2003; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013; Weisleder & Fernald, 2013). However, there are complications for the view that data is all that matters. Maturational constraints are often thought to play an important role in language learning (Borer & Wexler, 1987; Newport, 1990). Many words like function words (e.g. "the") and number words (e.g. "two") are learned surprisingly late for their frequency, suggesting that the number of times a word is heard by a child is not a definitive predictor of learning. This fact has motivated hypothetical processes, including maturational constraints on function words or syntax (Borer & Wexler, 1987; Modyanova & Wexler, 2007) and conceptual or linguistic constraints in the case of number words (Carey, 2009).

At the heart of data-driven accounts is an ambiguity about how much data is required. Experimental studies of word learning have revealed children's ability to acquire word meanings from single instances (Carey & Bartlett, 1978; Heibeck & Markman, 1987; Markson & Bloom, 1997; Spiegel & Halberda, 2011) as well as from the aggregation of word usage across multiple contexts (Smith & Yu, 2008). It is not known which of these regimes governs the majority of lexical acquisition: Are most words learned by aggregation of tens, hundreds, or thousands of examples, or from a single informative instance?

Here, we develop a novel data analysis of word learning across thirteen languages in order to address two questions about early word learning: when does it begin and how much data does it require? These questions turn out to be interrelated—they are

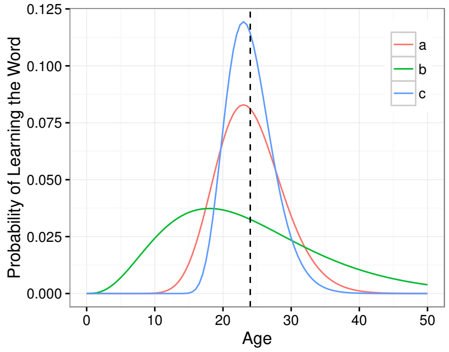


Figure 3.1: Example acquisition ages under three example assumptions: (a) children receive learning instances once a month from birth and require 24 total, (b) children require 4 examples and receive one every 6 months on average, (c) children require 12 instances, coming once every month, but only begin accumulating data at 12 months. Each predicts the same mean of 24 months (dotted line), but different shapes and variances in the timing of acquisition.

coupled together by quantitative predictions that they make about the *distribution* of ages at which children learn a word. To illustrate this, consider a simplified picture of learning: suppose that a word is learned by age two. This could occur under many different situations. Three illustrative examples are: (a) the child could start accumulating data at birth, require about 24 cross-situational examples of the word, and receive them about once a month; (b) the child could start accumulating data at birth, require 4 examples, and receive them on average once every 6 months; (c) the child could start accumulating data at 12 months, require 12 cross-situational examples, and receive them once a month.

The central idea of our approach is that although (a), (b), and (c) predict the same mean age of learning, they critically predict different distributions of ages at which acquisition succeeds due to the statistics of waiting for data (see Figure 3.1). Empirical measurement of the distribution shape could in principle distinguish these hypotheses, informing us about how data influences the process of word learning. For instance, if the distribution supported (b), we might infer that there are few early constraints on learning since data accumulation begins at birth, and that learning required few examples. If the data supported (c), we might infer that cognitive or maturational constraints delayed the accumulation of data substantially, and that word learning required aggregating information across contexts.

The logic of our approach is to formalize the process of learning by accumulating data. Following Hidaka (2013), we assume that learners successfully acquire a word after k Effective Learning Instances (ELIs), or instances of the word that contribute to the learner's accumulating an amount of information about the word. We also assume that ELIs arrive with an average frequency of λ per month[1]. However, unlike previous work, we also infer the age s at which data accumulation begins and implement our analyses in a Bayesian data analysis that is capable of inferring the likely ranges of parameter values from children's data. This Bayesian approach comes with several distinct advantages (Kruschke, 2010; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008), including the ability to determine all three variables simultaneously, with our uncertainty in each correctly influenced by uncertainty in the others. Thus, our inferences about the amount of data required to learn a word are statistically adjusted for our uncertainty over when learning that word began, and vice versa. The analysis also has the potential to reveal that the data is not informative about these variables, in which case we would find high uncertainty in the parameters given

[1] Hidaka (2013) compares three different generative models for Age of Acquisition distributions including one with a changing rate. In this analysis, we extend on his best fitting model for the greatest amount of words, which has a fixed rate. As this might seem counter-intuitive, we summarise the models he suggested and justify our choice of model in Appendix 3.A.1.

children's data. The advantage of our analysis compared to Hidaka (2013)'s model comparisons is that we can confidently focus on interpreting the parameter estimates.

3.1 Probabilistic assumptions

Our model requires three primary assumptions: (i) age of acquisition (AoA) consists of two periods of time: a start time s before learning a word begins and an accumulation time t, during which children are waiting for data; (ii) children learn a word after observing a number k of ELIs of the word; and, (iii) these ELIs occur stochastically, but at a fixed rate λ (measured here in ELIs per month). For instance, $s = 0$, $k = 24$ and $\lambda = 1$ in example (a) above. Note that the model infers these parameters from learning curves, *not* from counting putative ELIs in child-directed data. It is likely that a constellation of factors are involved in determining whether any given instance contributes to learning (counts as an ELI). Similarly start time s could reflect several processes, including when children develop the ability to track and remember the data that they need to learn a word, or when their conceptual repertoire is ready to begin learning a word.

When data is observed stochastically with a rate λ that is uniform in time, the number of ELIs actually received in a month will follow a Poisson distribution with rate λ. Under these assumptions, the distribution of times t children must wait before receiving k ELIs follows a Gamma distribution $\Gamma(k, \lambda)$ with density,

$$f(t; k, \lambda) = \frac{t^{k-1} e^{-t \cdot \lambda} \cdot \lambda^k}{\Gamma(k)}. \tag{3.1}$$

Thus, f describes the distribution of time children must wait before observing enough data to learn a word. The curves in Figure 1 are Gamma distributions with the appropriate values for k and λ. Note that in a Gamma, the mean scales linearly in the variance, meaning that if acquisition is driven by accumulating data, children's variance in learning times should scale with their mean learning time. Gamma-shaped learning time distributions should be taken as a hallmark of data-driven, constructivist accounts of learning (Xu, 2007; Kushnir & Xu, 2012) that applies to any theory of development in which accumulating data is the primary force advancing learners' knowledge.

3.2 The data analysis model

Our data analysis model uses Bayesian techniques to recover k, λ and s from empirically-measured learning curves. To do this, we require one data-analysis assumption that the population of children studied is relatively homogeneous, meaning that we may extend a word's single s, k, and λ across children[2]. In this case, the proportion of children who know a word at accumulation time T will approximate the cumulative distribution function of Equation 3.1 at time T,

$$F(T; k, \lambda) = \int_0^T f(t; k, \lambda)\, dt. \tag{3.2}$$

Figure 3.2 shows a graphical model of the relationships between these variables and the observed data. At each age a, N_a children were measured and x_a of them

[2] Our conclusions hold even if we relax this assumption (see Appendix 3.A.2).

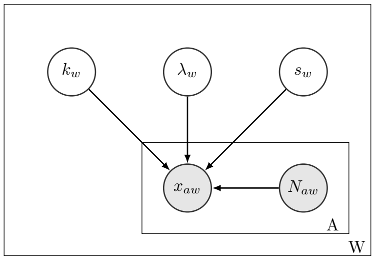


Figure 3.2: Graphical model notation for our model. Nodes denote variables of interest. Shaded nodes are observed variables. Plates denote groups of variables over age (A) and and words (W). In the text, we provide equations for a single word and omit the subscript w.

reported having learned the word to either production or comprehension[3]. We model the number of children producing/comprehending the word x_a as being drawn from a binomial distribution with N_a trials and a probability of success equal to the proportion of children who know the word given by Equation 3.2 at time $t = a - s$:

$$x_a \sim \text{Binom}(F(a - s, k, \lambda), N_a) \tag{3.3}$$

We assume uniform priors on these variables: $k \sim$ Uniform(0, 10000) ELIs, $\lambda \sim$ Uniform(0, 10000) ELI(s)/month and $s \sim$ Uniform(0, 1000) months. Bayesian inference in this generative model allows us to take the empirical acquisition curves and determine posterior distributions for k, λ, and s for each word in each language.

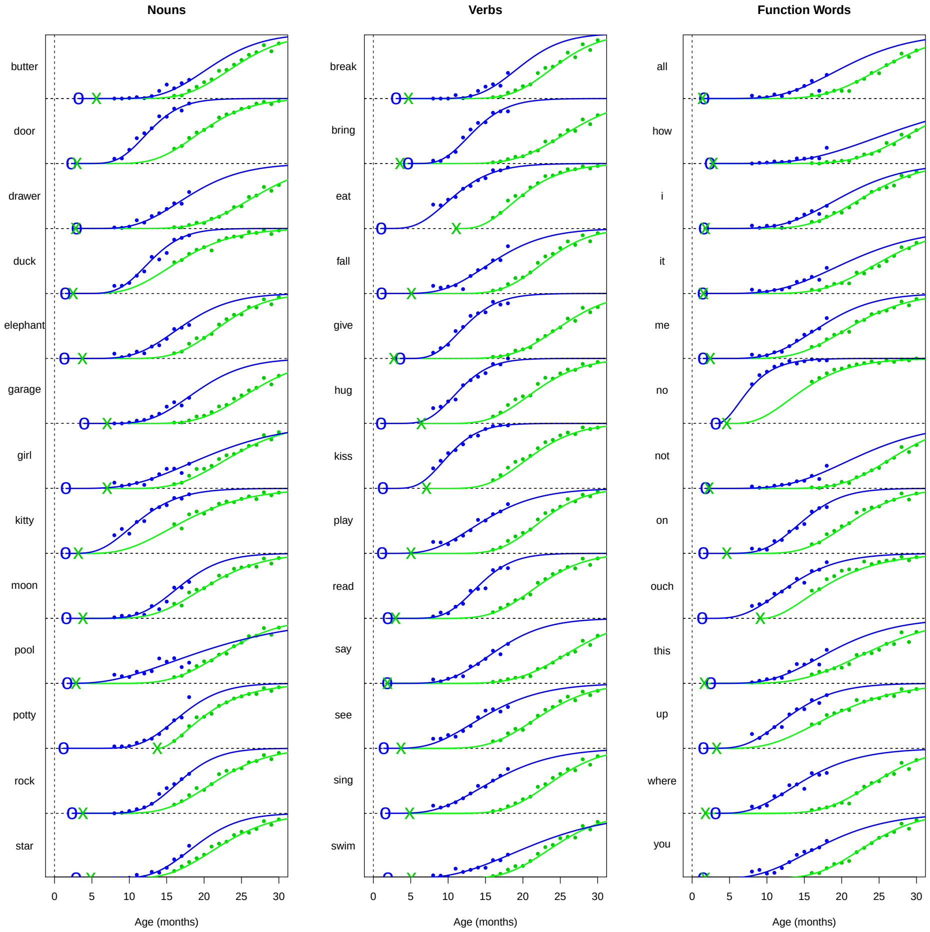


Figure 3.3: Points shows the proportion of English-speaking children (y-axis) who know a word at each age (x-axis) as measured by comprehension (blue) and production (green). Lines show the posterior mean parameters in the model (Equation 3.2), and X and O show the posterior mean start time of data accumulation for each word. This generally shows good model fits, early start times for comprehension, and somewhat later times for production.

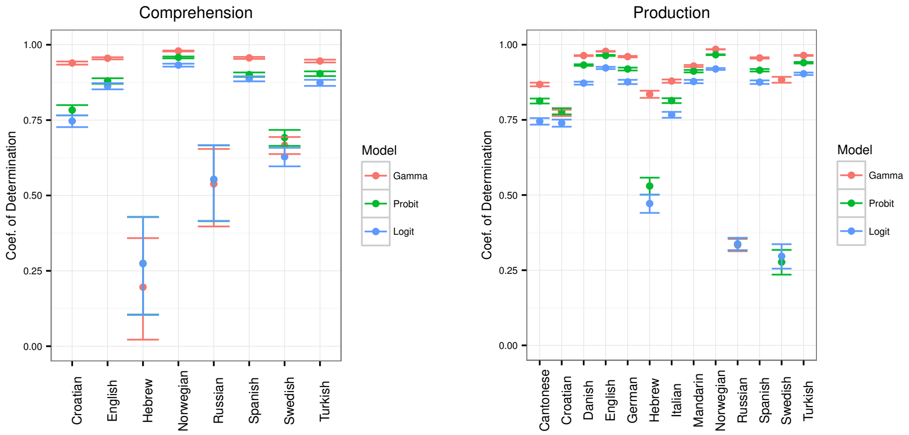


Figure 3.4: Model comparison of the Logit, Probit and Gamma models when trained on the first half of comprehension and production learning curves and tested on the full trajectory. Across words and languages, the correlations between observed data and model predictions for the full curve are close to 1 with the Gamma model showing the best fit.

3.3 Results

3.3.1 The cumulative Gamma matches observed word learning curves

Figure 3.3 shows a general visualization of the model fit across a variety of English words. Despite its simplicity, the model closely accounts for the empirical learning trajectories across word types for both comprehension and production. Quantitatively, correlations between predicted values and the behavioral data are near 1.0 for each language (see Figure 3.8) meaning that the model is able to capture the overall shape of acquisition across languages. More importantly, the model is able to more successfully *predict* learning than more standard alternatives: a probit (McMurray,

[3] We fit the comprehension and production data separately.

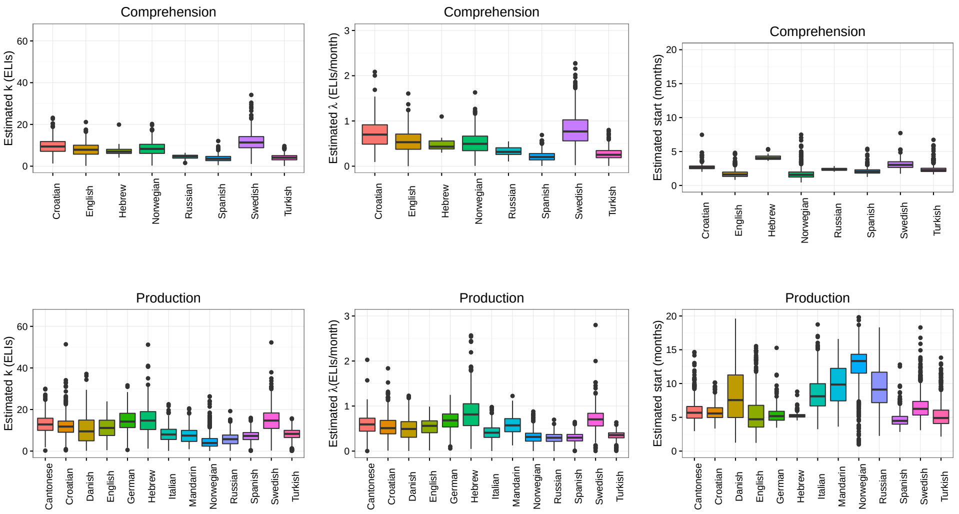


Figure 3.5: Box plots of the distribution of k, λ and s across words in each language.

2007) and a logistic model. To test this, we divided the learning curve for each word into two halves, where we fit k, λ and s for each word on the first half and then computed the correlation between model and human data across words and ages on the full curves. The Gamma distribution fit quantitatively out-performs either the probit or the logit across most languages (see Figure 3.4).

3.3.2 On the order of 10 ELIs are needed to learn a word

The order of magnitude of the estimated parameters are informative about the underlying mechanisms of learning, as they characterize when learning starts (s), how many ELIs are needed (k) and how frequently they occur (λ). Figure 3.5 shows the mean values of k, λ and s for each language. The box plots for English further broken down based on MCDI semantic category are similar (see Figure 3.9).

Figure 3.5a,d shows that, across languages, the order of magnitude of k is around

10 for production, with slightly lower values for comprehension. It is important to focus on the order of magnitude, not the exact numerical values, because the order of magnitude of our parameter estimates are robust to noise (see Appendix 3.A.2). The important issues in language development can still be distinguished based on order of magnitude. We primarily interpret Figure 3.5 as showing that languages agree in order of magnitude of their estimates[4]. Thus, children do not require hundreds or thousands of instances of a word to learn, even for words that may be very frequent, nor do they learn from a single instance. Instead, learning is likely focused around ten critically informative learning instances. These findings demonstrate the importance of cross-situational statistics over single examples and is consistent with the finding that children do not retain fast-mapped meanings (Horst & Samuelson, 2008).

3.3.3 ELIs of a word occur roughly every two months

The variable λ characterizes the estimated rate at which ELIs of a word occur. Figures 3.5b,e show that ELIs of a word occur once every two months ($\lambda \approx 0.5$), indicating that ELIs are relatively infrequent for an individual word. However, because children learn many words simultaneously, ELIs of any word may in fact be quite frequent. For instance, if children track statistics on 1000 early words, and observe an ELI for *each* word on average once every two months, they will receive around 17 ELIs per day.

[4] We suspect that the greater uncertainty around estimates for Hebrew and Swedish is due to data sparsity (see Figure 3.10).

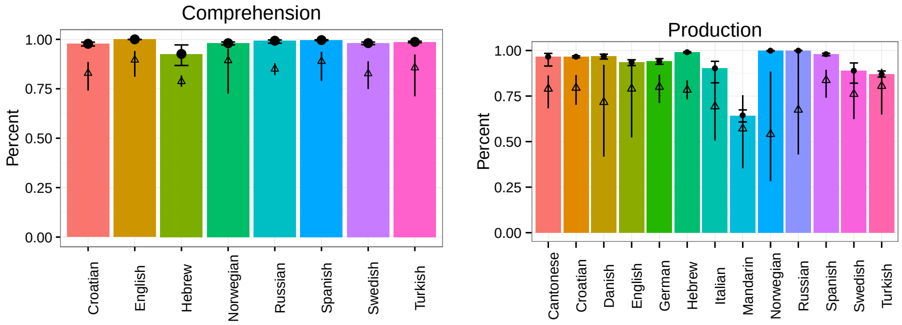


Figure 3.6: The bar plot shows percent of the variance in age of acquisition times explained by accumulation time (suggesting data-driven learning). The triangular points shows the percent of age of acquisition time spent accumulating data. Error bars and point ranges represent bootstrapped 95% confidence intervals. Outliers ($<$ 2.5% of the data) were removed for this analysis (see Methods).

3.3.4 Data accumulation starts around 2 months

The start times in Figures 3.5c,f show that learning begins early: approximately by two months in the case of comprehension measures. The starting age is somewhat later when curves are fit to production measures, possibly because production may require motor and speech systems to be working before production can progress. This may indicate that although maturational factors play little role in learning as measured by comprehension, production depends on the development of other cognitive or motor systems.

3.3.5 Early word learning is primarily data-driven

The model assumes that AoA is the sum of two time periods: start time s and accumulation time t. There are two measures we derive from these parameters to quantify the extent to which early word learning is data-driven: the percent of total

AoA time spent accumulating data, and the percent of variance in AoA explained by variance in accumulation times. If early word learning is primarily constrained by maturation, the majority of acquisition time should not be spent accumulating data and the majority of the variance in acquisition times should be explained by the variance in start times s. On the other hand, a data-driven account of early word learning would expect the majority of acquisition time to be spent accumulating data and the majority of the variance in acquisition times to be explained by variance in accumulation times t. Figure 3.6 shows the proportion of total acquisition time and the variance in acquisition times that is due to t (accumulating data) rather than s (start times). We find that generally the majority of acquisition time is spent accumulating data and the variance in accumulation times explains the majority of the variance in acquisition times. Taken together, this indicates that data-driven factors are the primary drivers of early word learning.

3.3.6 Learning instances are weakly correlated with log frequency

Under a simple view that most usages of a word are informative about its meaning, our estimates of k and λ should be surprising; word frequencies vary over several orders of magnitude (Zipf, 1949), yet the inferred k and λ values do not. This means that ELIs cannot be very strongly correlated with frequency. Most of the time a frequent word is used, it is not an ELI.

To investigate the relationship further, we computed the correlation between the estimated k, λ and s values for each word in English and the log frequency

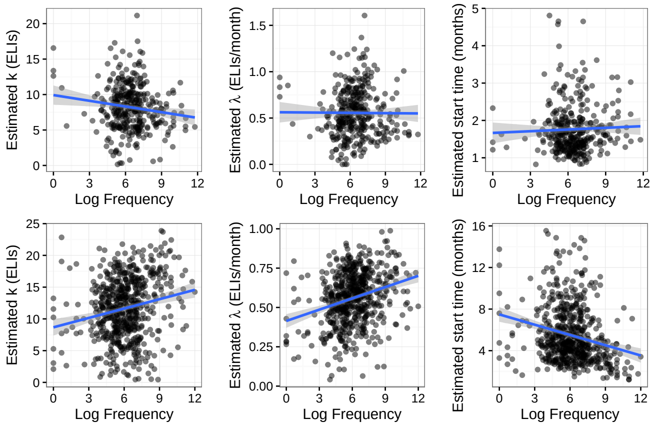


Figure 3.7: Correlations between CHILDES frequency for words in English and estimated parameter values. Top row: For comprehension, there is a small correlation between frequency and k and no correlation between frequency and λ and frequency and s. Bottom row: For production, the correlations between frequency and k, frequency and λ, and frequency and s are very weak and only significant when frequency is log transformed.

as measured in CHILDES (MacWhinney, 2000). For comprehension, there is only a small correlation between the estimated k parameter and frequency ($k : r = -0.14, p = 0.01$). For production, there is a modest correlation ($k : r = 0.19, p < 0.001$; $\lambda : r = 0.32, p < 0.001$; $s : r = -0.22, p < 0.001$) as observed by Hidaka (2013). But what is notable is the *weakness* of the correlation (see Figure 3.7)—it is not as though doubling the quantity of input will double the number of ELIs. This finding is compatible with findings of frequency effects in word learning (Ambridge, Kidd, Rowland, & Theakston, 2015; Hoff, 2003; Huttenlocher et al., 1991; Shneidman et al., 2013; Weisleder & Fernald, 2013), but suggests that frequency will be less important than the frequency of ELIs (see also Hoff, 2003).

3.4 Discussion

We view the Gamma model not as a mechanistic learning account, but instead as a scientific *tool* for understanding the basic forces in early language acquisition. Unlike characterizations in terms of mean acquisition ages, the parameters s, k and λ are *psychologically meaningful* in terms of a causal process that likely supports part of word learning, data accumulation (Hidaka, 2013). Our analysis of empirical learning curves strongly suggests that data accumulation begins very early, that production may be delayed due to maturational factors, and that typical words take on the order of ~ 10 ELIs to learn, not hundreds of occurrences and not a single occurrence or two. The model also suggests that the *informative* data points for word learning occur relatively infrequently, about once every two months, and that these occurrences are not strongly related to a word's overall frequency. Moreover, the mechanisms of data accumulation not only provide the best quantitative fit to learning curves, they explain *nearly all* of the variance in when children learn a word.

This analysis has capitalized on the existence of large corpora of acquisition trajectories across children. In particular, the key variables of interest, data amounts, data rates and the time at which data is first considered, are discovered entirely from children's acquisition trajectory—not from recordings of children's input. While it may seem tempting to address these questions of acquisition with an intensive home recording study (D. Roy et al., 2006) or an evaluation of child-parent interactions (MacWhinney, 2000), these approaches come with the challenge of delineating which instances of a word concretely contributed to learning. For example, a word use might only aid acquisition if the child is attentive and receptive, and the referent is clear, which might not be observable in those data sets. Given that we have found that overall frequency is a weak predictor of the rate of ELIs, the detailed measurement of just parental productions will not fully clarify the relevant data sources for learning. Instead, our work takes a different tack, looking to find evidence of data-driven effects writ large in the *distribution* of learning times for words.

This work leaves open a central question: what makes a usage of a word an ELI? The weak correlation between the parameters and word frequency suggests that ELIs are rare—and perhaps even intentional. It is likely that children actively decide what stimuli they engage and deeply process (Kidd et al., 2012; Kidd, Piantadosi, & Aslin, 2014), which could place an internal yoke on the rate of ELIs. Extrinsic factors probably also play a role though, as seen by the correlations with frequency. Analogously, these analyses raise the question of what determines differences in k and λ across words and languages. Future research should attempt to characterise the impact of external factors, such as semantic content (M. N. Jones, Johns, & Recchia, 2012) and phonotactic probability (Storkel, 2001), on k and λ. Our framework provides the initial step at connecting such factors to the data accumulation process that implicitly supports all existing models of word learning.

It is also important to note the limitations of the MCDI data and our model. First, we restrict all of our conclusions to the early learned words covered by the MCDI. It will be important to extend this model beyond the age range of the existing MCDI. Children are flexible learners and it is probable that an older child adopts a variety of strategies, which may influence the data-driven process. For example, older children might be able to bootstrap from their existing vocabulary/syntactic constructions or their intuitive theories of the world. Additionally, the lack of variability in the MCDI words constrains the empirical testing of many hypothesized constraints on vocabulary acquisition (e.g., Markman, 1990). Applied to the appropriate data, our approach is a suitable tool to evaluate these constraints at the computational level. Further, we chose to encode maturation as a constant offset from birth to address our main questions. This is an appropriate operationalization but a coarse distinction and future research should address this.

3.5 Conclusion

Our results have shown that under a simple model of learning as waiting for data, we may estimate the amount of data required to learn a word, the rate at which useful instances occur and the start time of learning from group-level learning curves. Our results robustly demonstrate that on average words require on the order of ~ 10 ELIs to learn across multiple languages. ELIs appear to occur about once every two months, relatively independent of frequency. Children start accumulating data very early, but their learning may be delayed in the case of production while systems like motor processing mature. Empirically, our model provides close fits on held-out data and suggests that waiting for data is a primary constraint on early word learning, consistent with views emphasizing the important role that data plays in learning and development.

3.6 Methods

We fit k, λ and s within individual words and languages on data retrieved on June 16th 2015 from Wordbank (Frank et al., 2015), a repository for MCDI instruments (Fenson et al., 2007). This yielded cross-sectional data from thirteen languages (see Figure 3.10 for further description). For each word in each language, k, λ and s were fit using JAGS (Plummer et al., 2003) and corresponding R packages, `rjags` and `runjags`. For every word, four chains were run for a total of 1.25 million steps with a thin of 1000 steps between each saved step. The chains converged ($\widehat{R} < 1.2$) for all 2397 words in the comprehension and 9420 words in the production measure. For our data vs. maturation analyses, we removed outliers ($< 2.5\%$ of the data), that were all syntactic constructions as opposed to lexical items. The forward predicting model was trained on the first half of the data using the same method. In these runs, 88 words failed to converge for comprehension and 78 words failed to converge for production and were excluded from further analysis. Code and parameter estimates are available from the first author and our lab's webpage.

3.A Appendix to Chapter 3

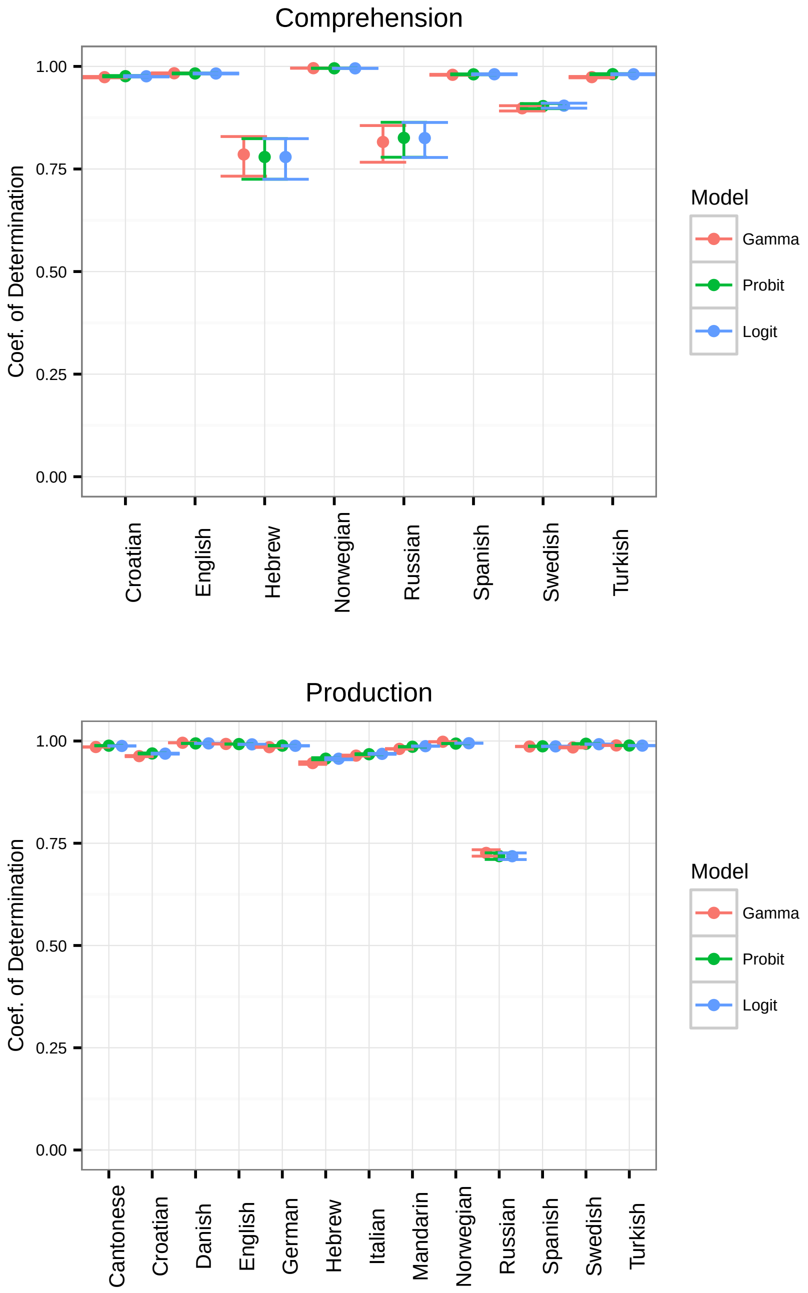


Figure 3.8: Model comparison of the Logit, Probit and Gamma models when trained on the full learning curve. Across words and languages, the correlations between observed data and model predictions are close to 1.

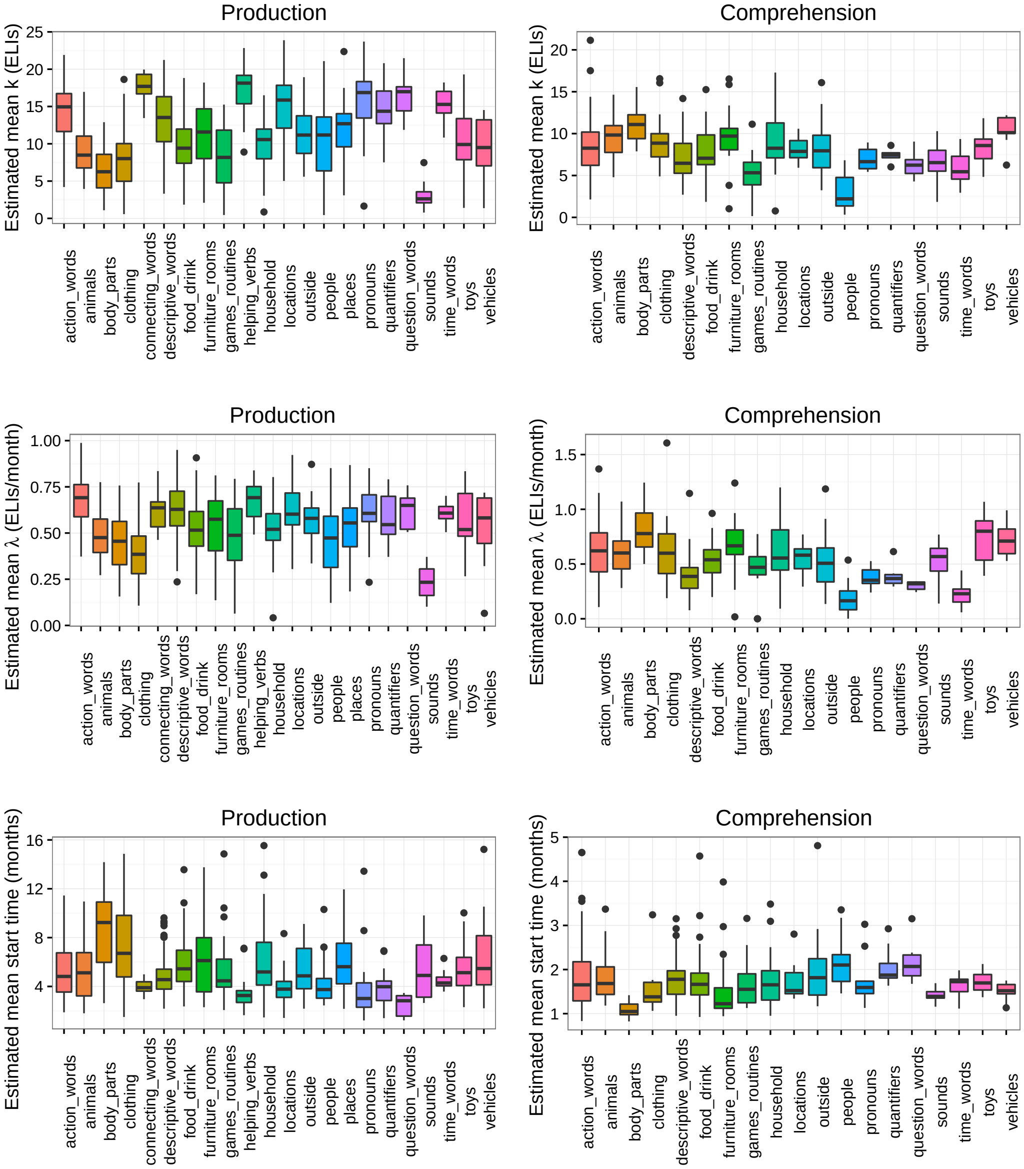


Figure 3.9: Box plots of the mean k, λ and s values measured for English words split by MCDI semantic category.

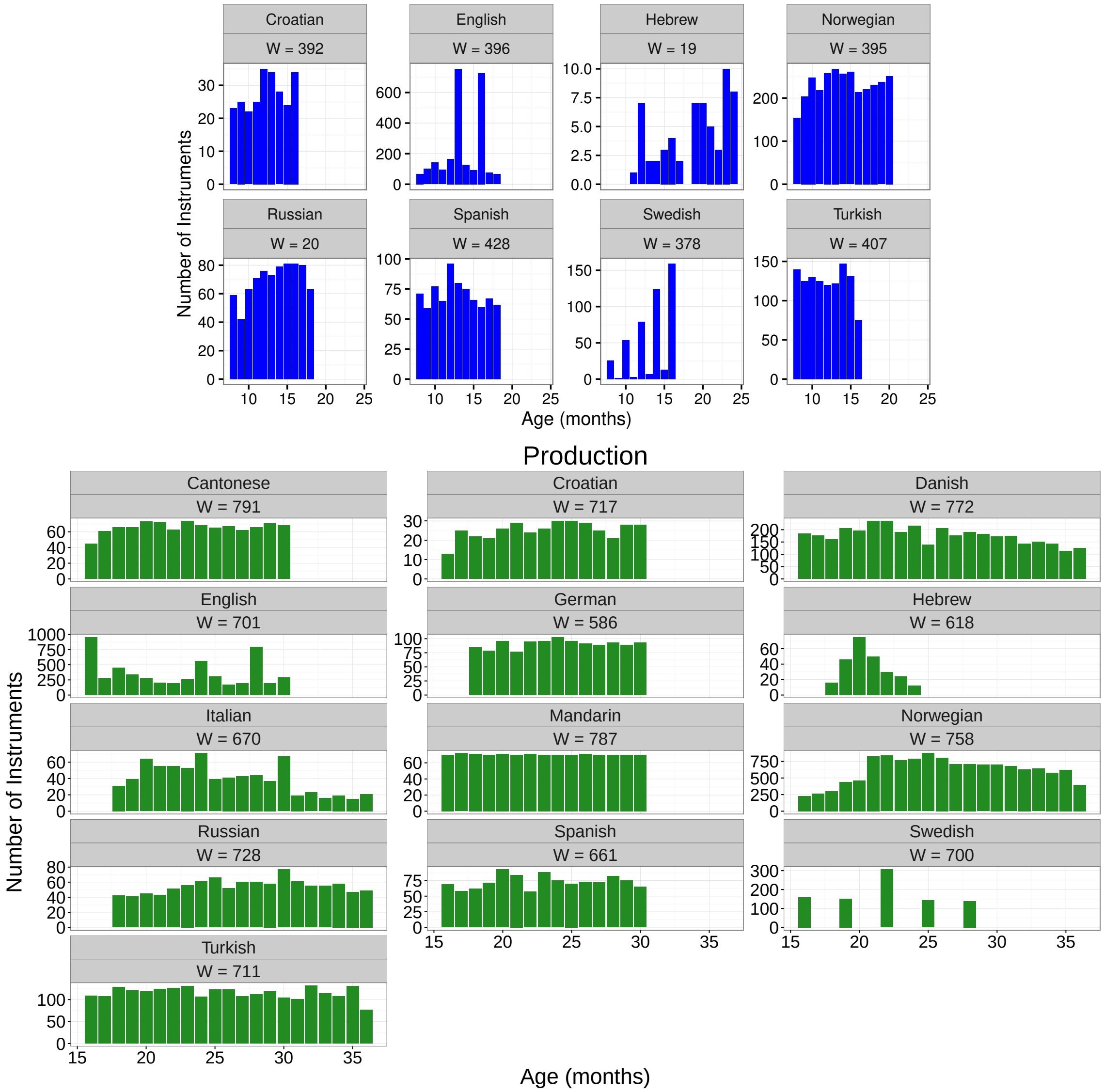


Figure 3.10: Number of completed MCDIs at each age for each language and words for each instrument. Note the y-axes differ in each panel.

3.A.1 Model Justification

Hidaka (2013) conducted a model comparison of three different generative models for the AoA distributions: a rate-change learning model (i.e., a Weibull model), a cumulative learning model (i.e., a Gamma model), and a cumulative-and-rate-change learning model (i.e., a Weibull-Gamma model). In the rate-change model, a learner only requires a single ELI and each month the initial probability of observing an ELI, λ, changes (presumably increases) polynomially, with an exponent of δ. The cumulative learning model is the gamma model we chose to implement (without a start time parameter), i.e., a learner requires k ELIs to learn a word and ELIs come stochastically but at a fixed rate, λ ELIs/month. In the cumulative-and-rate-change model, a learner requires k effective learning instances to learn a word and these instances have a base rate λ which changes by a power of δ each month.

The cumulative distribution function of these three models describes the probability of a child having learned a word as a function of age a. The equations for these three models follow:

Weibull Model

$$F(a; \lambda, \delta) = \gamma(1, (\lambda \cdot a)^{\delta}) \tag{3.4}$$

Gamma Model

$$F(a; k, \lambda) = \frac{\gamma(k, \lambda \cdot a)}{\Gamma(k)} \tag{3.5}$$

Weibull-Gamma Model

$$F(a; k, \lambda, \delta) = \frac{\gamma(k, (\lambda \cdot a)^{\delta})}{\Gamma(k)} \tag{3.6}$$

where $\gamma(k, b)$ is the lower incomplete gamma function,

$$\gamma(k, b) = \int_0^b t^{k-1} e^{-t} dt. \tag{3.7}$$

and $\Gamma(k)$ is the gamma function:

$$\gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt. \tag{3.8}$$

The parameter k is interpreted the same as above—i.e., the number of ELIs required for learning. The parameter b corresponds to the expected number of instances observed at that time.

Hidaka (2013) fit these models for 652 productive vocabulary words in the MCDI. He found that when aggregating over words, the cumulative-and-rate-change model has the best fit as measured by Bayesian Inference Criterion (BIC). However, when he looked at each word individually and compared the BICs for the different models, he found that the cumulative model fits best for 50% of the words. This means that the cumulative-and-rate-change model is a good overall model of early word learning[5], but the generative process that best captures how individual words are learned is the cumulative (Gamma) model.

We chose to use and extend the cumulative (Gamma) model for our purposes for two reasons. First, the majority of the individual words fit by Hidaka (2013) were best fit by the Gamma model, making the option of choice for capturing individual word curves. Second, the Gamma model has a more straightforward interpretation than

[5] This is not too surprising considering both the cumulative (Gamma) model and the rate-change (Weibul) model are special cases of the cumulative-and-rate-change model.

the rate change models. While the k parameter retains the same units (number of ELIs) across all models[6], the interpretation of the parameter b differs across models. For the gamma model, the unit for b—i.e., expected ELIs, does not change across time. However, in the rate change model, b is raised to an exponent, giving it a much less clear interpretation. It effectively gives rise to some polynomial of time, but not one which is to our knowledge motivated by independent considerations. These two factors lead us to build off the gamma model rather than the rate change model, even though the latter fits better in one analysis.

Lastly, we can directly compare the performance of cumulative-and-rate-change models and our model in predicting our data. These results show that the model we focus on, a Gamma model with start time, does a much better job than others in predicting learning curves. Figure 3.11 displays the coefficients of determination (across learning curves) of individual English words[7]. This figure contains four models: Hidaka (2013)'s estimated parameters for the Weibull-Gamma (Hidaka_WG), a Bayesian Data Analysis[8] (BDA) Weibull-Gamma model (BDA_WG), a BDA Weibull-Gamma with a start time parameter (BDA_WGs), and a Gamma model which includes a start time (BDA_Gs) (the primary one in our analysis).

Comparing Hidaka's Weibull-Gamma to the BDA_WG, we see that the BDA_WG

[6] In a Weibull distribution $k = 1$ ELI

[7] Code and parameter estimates available on lab website.

[8] For the Bayesian Data Analysis models, the same inference procedure was used as in the main paper. The prior on the rate change parameter was $\delta \sim Gamma(2.25, 1.25)$. For the BDA_WG model, 319 of the 797 English MCDI words converged and were used in the analysis. Although the number of words successfully fit under this model seems low, Hidaka's WG parameter estimates only yield learning curves for 117 words. The WG parameters for the rest of the words Hidaka fit describe learning curves that are 1 or 0 over the window of data. For the BDA_WGs, 698 of the 797 words converged and were used in the analysis.

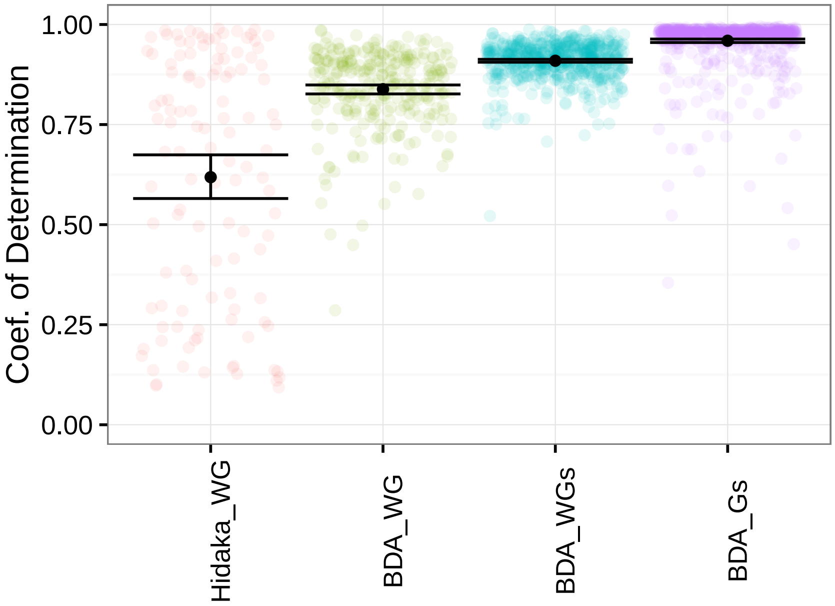


Figure 3.11: Model Comparison of Hidaka's Weibull-Gamma (Hidaka_WG), a BDA Weibull-Gamma (BDA_WG), a BDA Weibull-Gamma with a start time (BDA_WGs), and our Gamma model, which includes a start time (BDA_Gs). The low contrast points represent the coefficient of determination for individual words. The points represent the mean and their error bar's represent bootstrapped 95% confidence intervals.

provides a slightly better fit to the Wordbank MCDI data. Unfortunately we cannot distinguish if this might be due to the increased data amounts the parameters were estimated from or the robustness of Bayesian Data Analysis as a method. More interestingly, the inclusion of a start time parameter to the BDA_WG significantly increases the model fit to the data. Nonetheless, the Gamma model (BDA_Gs) still outperforms all of the cumulative-and-rate-change models. In other words, on these data sets, the model with a start time and no rate change provides the best fit.

However, given the nearness in fit between the BDA_WGs and our model, we compared the parameter values to see if the rate change parameter significantly differed from 1, which would suggest very little to no change in rate. We find that for 98% of the words, the rate change parameter is not significantly different than 1. This explains why BDA_Gs, which has this parameter set to 1, can perform so well.

To summarize, these results justify the use of a gamma model with start time in our primary analyses. However, it is important to remember that other age ranges or data sets may necessitate models with different probabilistic assumptions.

3.A.2 Relaxing Our Data Analysis Assumption

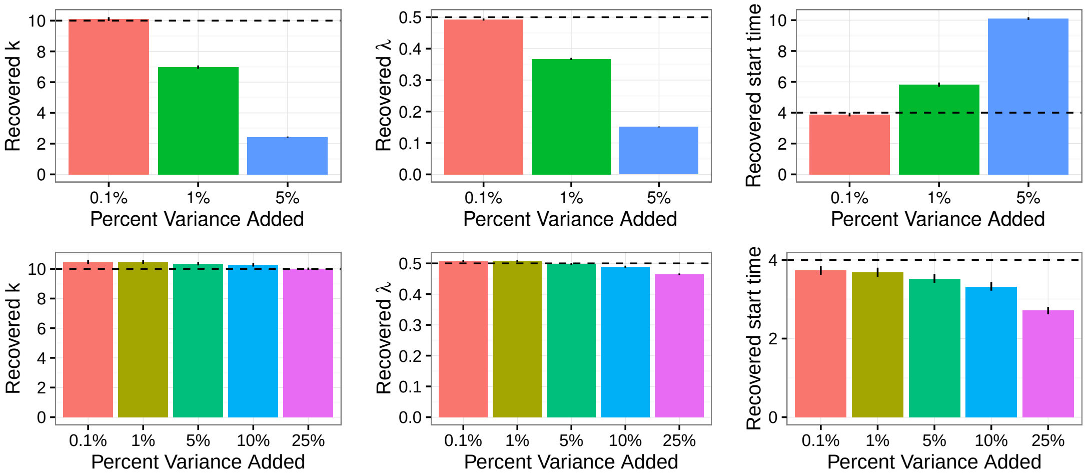


Figure 3.12: Recovered parameters estimates for simulations with varying percent model-internal noise (top row) and model-external noise (bottom row). The dashed line represents the generating parameter value. Point ranges reflects 95% bootstrapped confidence interval.

In our analysis, we make the data analysis assumption that the parameter values we infer will be the same across children; however, it is widely acknowledged that children implement different strategies for language learning (Brown, 1973). To examine how the model works when our assumption is violated, we simulated data with two different types of noise: model-internal noise—i.e, noise in the parameter values, and model-external noise. Model-internal noise might reflect individual differences in the learning process. Whereas, model-external noise might reflect things like measurement error.

We simulate data with model-internal noise by sampling parameters as follows:

$$\begin{aligned} k &\sim N(10, \sqrt{10v}) \\ \lambda &\sim N(0.5, \sqrt{0.5v}) \\ s &\sim N(4, \sqrt{4v}) \\ AoA &\sim \Gamma(k, \lambda) + s \end{aligned} \tag{3.9}$$

where v is the percent noise. To assess internal noise, we added either 0.1%, 1% and 5% noise. As the percent of internal noise increases, the shape of the AoA distribution to be fit changes significantly. For example, adding 1% internal noise increases the standard deviation of the AoA distribution by one month; whereas, adding 5% internal noise increases the standard deviation of the AoA distribution by at least 20 months. For each percent noise, we simulated age of acquisition data for 1000 children. We binned the simulated data across the age range of $15-36$ and ran the model on the binned data. We repeated this process 1000 times.

We expect that the recovered parameters from the model runs should be similar to the generating parameters. The results are shown in the top row of Figure 3.12. First, note that with only 0.1% model-internal noise added, the recovered parameters are virtually the same as the generating parameters. Second, we find that under a reasonable percentage of added model-internal noise, the model recovers parameter values on the same order of magnitude as the generating parameters, suggesting that model-internal noise has a small effect on the order of magnitude of the parameter values. Lastly, we find that as the percent of model-internal noise increases, the recovered parameters for k and λ are under-estimated and the recovered parameter for s is over-estimated.

Given that the data in Wordbank (Frank et al., 2015) , like all data, is inherently noisy, these simulations would suggest that our estimates for k and λ should be interpreted as lower bounds and our estimates for s should be interpreted as an upper bound. In effect, the presence of model-internal noise under-estimates the contribution of data-driven processes to word learning and over-estimates the contribution of maturational processes. Despite this, we still find that the majority of the variance in early word learning can be explained by the simplest data-driven processes, i.e., waiting for data.

We simulate data with model-external noise by sampling ages of acquisition as follows:

$$\begin{aligned} a &\sim \Gamma(10, 0.5) + 4 \\ AoA &\sim N(a, \sqrt{av}) \end{aligned} \tag{3.10}$$

To assess external noise, we added either 0.1%, 1%, 5%, 10% or 25% noise. As can be seen in the bottom row of Figure 3.12, We find that the model is remarkably robust at recovering the generating parameters when model-external noise is added.

Chapter 4

Universal and cultural specific processes in exact number word acquisition

While there is numerical structure in the world, mathematics, itself, is a human invention. Humans have two core conceptual systems for dealing with numbers/math: the object file system (Carey & Xu, 2001) and the approximate number system (Dehaene, 1997). The object file system allows humans to track and manipulate approximately four objects (Feigenson & Carey, 2003; Leslie, Xu, Tremoulet, & Scholl, 1998) but fails as the number of objects increases (Feigenson & Carey, 2005). On the other hand, the approximate number system allows us to represent large magnitudes; however, as the set of objects to maintain increases, the fidelity with which we represent this set decreases (Dehaene, 1997). Neither of these core conceptual systems fully support the rich and precise numerical life prevalent in industrialized societies. Conceptual structures for exact number (e.g., counting systems) are cognitive technologies (Frank, Everett, Fedorenko, & Gibson, 2008) that people develop to handle the need for exact number representations for large quantities. As a result, counting systems are fundamentally cultural innovations shaped by the societal needs of a community. Despite what would appear to be a common problem, counting systems have considerable diversity in, for example, the base of the system, the highest number represented by the system and the iconicity of the writing systems (Epps, Bowern, Hansen, Hill, & Zentz, 2012; Beller et al., 2018). While economic pressures have pushed cultures to adopt the same numerical systems, the problems and need/desire for number still varies across cultures (Everett et al., 2005; Núñez, 2017). Therefore, number acquisition is a prime case study for teasing apart cultures influence on the human learning mechanisms.

The development of exact number is an important case study in conceptual change (Carey, 2009): how do humans create a conceptual system that goes beyond the capacity of our core knowledge? Given the significant diversity in the numerical systems of the world, it is actually surprising that children's development shows the same pattern of acquisition across variable cultures (e.g., Sarnecka et al., 2007; Le Corre, Li, Huang, Jia, & Carey, 2016; Piantadosi, Jara-Ettinger, & Gibson, 2014). Children begin by memorizing their number words. Then, they slowly learn how to map quantities of objects to number words in the count list. They initially gain competence in mapping sets of one to *one*, then sets of two to *two*, three to *three*, and four to *four*. Instead of repeating this process indefinitely to learn the meaning of all exact number words, children appear to master the logic of counting after learning the meaning of four, recognizing that the cardinality of a set of objects is the last number in the count list—i.e., the Cardinal Principle. Children demonstrating mastery of only a subset of the count are often referred to as *subset knowers*; whereas, children who have mastered the cardinal principle are often referred to as

CP-knowers. This pervasive developmental trajectory suggests that the fundamental learning problem drives the acquisition of numerical concepts.

At the same time, anthropological evidence suggests the timing of numerical knowledge acquisition varies across cultures (Saxe & Posner, 1983) as different cultures still place different value on number and the tasks in which number is used. For example, the unschooled children of merchants are often just as adept at numbers as schooled children but far superior to those of unschooled children (Posner & Baroody, 1979). From a different angle, the primary uses of number can determine whether it's worthwhile to learn a full blown system of numbers. For example, Brazillian candy-sellers are primarily concerned about small financial transactions and solve the problem not by learning a full number system but by learning a small set of shuffling algorithms with colored papers (i.e. money) to complement their approximate number system (Saxe, 1988b, 1988a). Looking specifically at the acquisition of exact number words, researchers have also reported cross-cultural differences in acquisition. Specifically, Mandarin speaking children are slower to acquire *one* than English speaking children (Le Corre et al., 2016). Similarly, Japanese speaking children are slower to acquire *one* than English and Russian speaking children (Sarnecka et al., 2007; Barner, Libenson, et al., 2009). Saudi Arabic and Slovenian children are faster to acquire *two* than English and Mandarin speaking children (Almoammer et al., 2013). Slovenian children from urban environments are faster to learn *two* than Slovenian children in rural environments where there is also a dialect difference (Marušič et al., 2016). Tsimané children acquire exact numerical number much later than those other cultures (Piantadosi et al., 2014). The pattern of differences across cultures has been attributed to the differences in linguistic cues, specifically number morphology and quantifiers (Le Corre & Carey, 2007; Le Corre et al., 2016; Sarnecka et al., 2007; Barner, Libenson, et al., 2009; Barner & Bachrach, 2010; Barner, Chow, & Yang, 2009; Almoammer et al., 2013; Marušič et al., 2016), while trying to control for other cultural differences as much as possible.

Here, we compile a large ($N = 1772$) cross-cultural ($n = 8$) data set of exact number acquisition. We map how various cultural differences would influence learning under a computational model of exact number acquisition (Piantadosi et al., 2012). Using a novel linking hypothesis based on Mollica and Piantadosi (2017a); Hidaka (2013), we extend this ideal learner model to make predictions about behavior at a given time, as opposed to at a given amount of data. This allows us to conduct a descriptive Bayesian Data Analysis (Tauber et al., 2015), where we learn the parameters of the model from the empirical data and then interpret the parameters of the model in terms of universal and cultural-specific influences. Specifically, how does culture influence the learning process? We consider two possibilities: culture influences the rate at which children use data or culture influences how children generate hypotheses. Finally, we round out our discussion with limitations of our model and future directions of its application.

4.1 Model

To model children's acquisition of exact number words as a function of data, we start with the ideal number word learning model of Piantadosi et al. (2012). In an ideal learner model of development, we must specify the space of possible hypotheses h a learner might entertain (e.g., *if there is only one object say "one" otherwise guess*), any prior biases they place on hypotheses, and how they evaluate their hypotheses against data D (i.e., likelihood function). The developmental trajectory can be then be modeled as the movement of probability mass over a hypothesis space as a function of data, following Bayes rule:

$$P(h|D) \propto P(D|h)P(h). \tag{4.1}$$

In this model, a data point consists of a word w and a context c (e.g., (*three*, 🐨🐨🐨)). To illustrate learning in a natural environment, Piantadosi et al. (2012) simulated data points according to CHILDES frequencies for number words (MacWhinney, 2000). Larger corpus investigations have found similar frequency ratios in adult speech across several different languages and modalities (Dehaene & Mehler, 1992). A learner evaluates their hypotheses based on the probability that their lexicon $\mathcal{L}$ generates observed data points:

$$P(D|h, \alpha) = P(w|c, \mathcal{L}, \alpha) = \begin{cases} \alpha + \frac{1-\alpha}{N} & \text{if } \mathcal{L}(c) = w \\ \frac{1-\alpha}{N} & \text{else} \end{cases}, \tag{4.2}$$

where α is a reliability parameter and N is the highest number a learner can count[1]. Piantadosi et al. (2012) assumed a simplicity prior over hypotheses. Instead of assuming a prior, a strength of our model is that we infer the prior over hypotheses

[1] For all analyses in the paper α was set at 0.9 and N at 10.

from developmental data, following the rich literature on prior inference (Griffiths & Tenenbaum, 2011; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010).

The primary limitation in connecting ideal learner models to actual developmental data is that we often only observe children's behavior at a given time, without any measure of how much data they have seen or more importantly used. When it has been attempted, the most common approach to linking model predictions as a function of data to behavioral observations at given times is to fit some function of time to the mean age of the behavior (e.g., logistic regression). While this seems reasonable, these models have poor predictive fit compared to models that fit both the mean and variance of a learning curve (Mollica & Piantadosi, 2017a). A considerable amount of information about how data relates to time is conveyed by the variance in acquisition times and can be used to make better predictions for behavior as a function of time and inform the underlying process of data use (Hidaka, 2013; Mollica & Piantadosi, 2017a). This is the approach we will adopt for our model. Other studies trying to gain insight about behavior as a function of time (including those by the first author) have made the false assumption that measured estimates of data points in a learner's environment at a given age (e.g., from a corpus study) reflect the amount of data a child uses to reach some learning outcome. This is often an implicit assumption in computational models that make use of corpora as naturalistic input. These approaches often pool variance caused by maturation with variance in learning. Further, there is evidence to suggest that children filter their input (Perkins et al., 2017; Kidd et al., 2012) and that the data they receive is not always reliable (Medina et al., 2011; Cartmill et al., 2013).

Borrowing from techniques in survival analysis, Chapter 3 describes one of the simplest generative processes for how data amounts, the rate of data accumulation and the onset of learning predict the timing of lexical acquisition for children's earliest learned words. In their model, an age of acquisition consists of two durations of time: maturational time s before attending to data and time spent waiting for data t. Data accumulate at a fixed rate λ. After observing k data points, a child has acquired the same behavioral performance as an adult. Mollica and Piantadosi (2017a) found that across 14 different languages, the age of acquisition curves for children's earliest learned words were best explained and predicted by this *constant rate* generative process.

Our model in Figure 4.1a represents a joint probability distribution:

$$P(t, \lambda, n, \theta, \mathbf{d}, h, k) = P(k|h)P(h|D)\prod_{d=0}^{n} P(d|\theta)P(\theta)P(n|t, \lambda)P(\lambda)P(t). \qquad (4.3)$$

The shading in the figure reflect variables that we observe. The model describes a generative process: At a given age t and rate of effective learning instances (ELIs) λ, the number of ELIs that children observe n is given by a Poisson distribution:

$$n \sim \text{Poisson}(\lambda t). \qquad (4.4)$$

We assume that the type of ELIs (e.g., (*one*, 🐨) or (*two*, 🐨🐨)) will vary according to the measured environmental distribution of number words $\vec{\phi_w}$:

$$D \sim \text{Multinomial}(n, \vec{\phi_w}). \qquad (4.5)$$

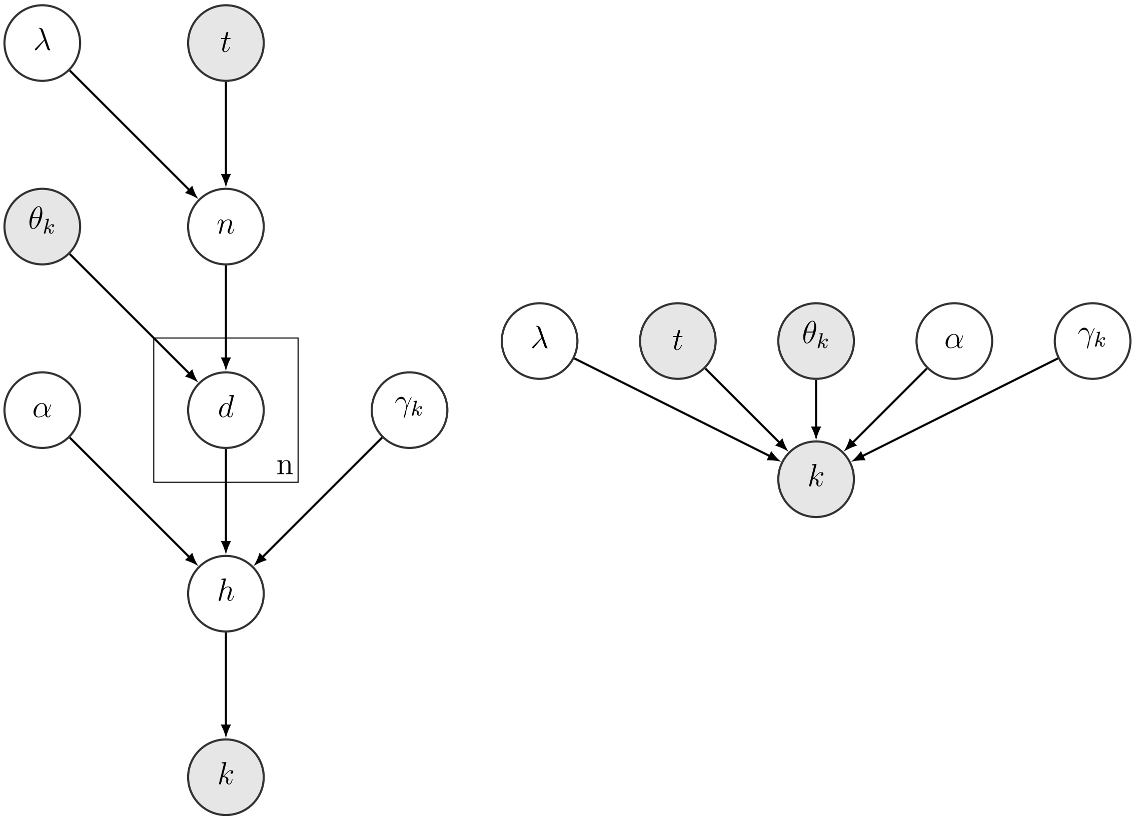


Figure 4.1: The full plate diagram for our data analysis model (left) and the plate diagram after marginalizations (right).

For mathematical convenience, we represent the probability of the environment generating a data point d that would be responded to correctly at particular knower level k as:

$$d \sim \text{Bernoulli}(\theta_k), \tag{4.6}$$

where $\theta_k = \sum_{w \in k} \phi_w$. A learner will then use this dataset D to update their beliefs about number hypotheses following Equation 4.1. For our data analysis, we make two simplifications of the model in Figure 4.1b. First, we marginalize over all possible amounts of ELI n and datasets D:

$$\begin{aligned} P(k|t,\lambda,\theta,\alpha,h) &= P(k|h)P(h)\sum_{n=0}^{\infty}P(n|t,\lambda)\sum_{d=0}^{n}\prod_{i}P(d_i|\theta)P(d_i|h) \\ &= P(k|h)P(h)e^{\lambda t\left[\left((1-\theta_k)\cdot\frac{1-\alpha}{N}\right)+\left(\theta_k\cdot\left(\alpha+\frac{1-\alpha}{N}\right)\right)-1\right]}. \end{aligned} \tag{4.7}$$

Second, we only consider six knower level hypotheses corresponding to the knower level patterns (i.e., Non, One, Two, Three, Four and CP). Marginalizing over hypotheses, we can rewrite Equation 4.7 as:

$$f_k(t;\lambda,\theta_k,\gamma_k) = P(k|t,\lambda,\theta_k,\alpha) = \gamma_k e^{\lambda t\left[\left((1-\theta_k)\cdot\frac{1-\alpha}{N}\right)+\left(\theta_k\cdot\left(\alpha+\frac{1-\alpha}{N}\right)\right)-1\right]}, \tag{4.8}$$

where γ_k represents the prior probability for the hypothesis. The full derivations of Equations 4.7-4.8 can be found in Appendix 4.A.1.

Translating this to a data analysis, at each age, we model the number of children for each knower level $\mathbf{x_t}$ as being drawn from a multinomial distribution with N_t trials and a probability of success given by Equation 4.8:

$$\mathbf{x_t} \sim \text{Multinomial}(f_k(t;\lambda\theta_k\gamma_k), N_t). \tag{4.9}$$

4.1.1 The influence of culture in our model

From the remaining parameters in our model (Figure 4.1b), there are two ways in which we could observe culture influencing the learning process. First, culture could shape the availability of data and thus, the rate of ELIs λ. For instance, Mandarin children grow up in an environment with twice as much number data (Le Corre et al., 2016) as other *Western Educated Industrialized Rich and Democratic* (WEIRD) children, but still with the same type distribution of data as other WEIRD countries (Dehaene & Mehler, 1992). As a result, Mandarin learners might have a faster rate of ELIs λ. Similarly, culture can shape how much attention children pay to number. For example, children in WEIRD cultures receive significant formal instruction in number compared to children in non-WEIRD cultures. The explicit engagement with number might increase the rate of ELIs. Second, cultures could influence how easy it is to generate a knower level hypothesis either through the primitives learners bring to the task or the weight learners place on those primitives—i.e., the prior over hypotheses γ_k. For example, if a culture focuses on numerical tasks that don't require exact number but privilege faster approximate number computations, learners may be more likely to start generating hypotheses based on approximate number computations.

To assess differences in rate and differences in priors, we construct a series of hierarchical Bayesian models and analyse a large cross-linguistic dataset of children's number knowledge. The first model we consider is a baseline multinomial logistic regression with random slopes for knower level and age by culture:

$$\begin{aligned}
B_{0k} &\sim \mathcal{N}(0, 32) \\
B_1 &\sim \mathcal{N}(0, 32) \\
\sigma_{B_{0k}} &\sim \text{Uniform}(0.0001, 10000) \\
\sigma_{B_1} &\sim \text{Uniform}(0.0001, 10000) \\
\beta_{0kc} &\sim N(B_{0k}, \sigma_{B_{0k}}) \\
\beta_{1c} &\sim N(B_1, \sigma_{B_1}) \\
P(k|t) &= \frac{e^{\beta_{0kc}+\beta_{1c}\cdot\text{age}}}{\sum^K e^{\beta_{0kc}+\beta_{1c}\cdot\text{age}}}.
\end{aligned} \tag{4.10}$$

The second model is a universal model (i.e., Constant Rate, Constant Prior—*CR-CP*) in which there are no differences across cultures (see Figure 4.1b). We create a cultural rate model (Varying Rate, Constant Prior—*VR-CP*) by extending the universal model with a hyperprior on the rate of ELIs and letting each culture have a different rate:

$$\begin{aligned}
\Lambda &\sim \mathcal{N}(0, 32) \\
\sigma_\lambda &\sim \text{Uniform}(0.0001, 10000) \\
\lambda_c &\sim \mathcal{N}(\Lambda, \sigma_\lambda).
\end{aligned} \tag{4.11}$$

We create a cultural prior model (i.e., Constant Rate, Varying Prior—*CR-VP*) by extending the universal model with a hyperprior on the knower level prior and letting each culture have a different knower level prior:

$$\begin{aligned}\Gamma_k &\sim \mathcal{N}(0, 32)\\ \sigma_{\gamma_k} &\sim \text{Uniform}(0.0001, 10000)\\ \gamma_{kc} &\sim \mathcal{N}(\Gamma_k, \sigma_{\gamma_k}).\end{aligned} \tag{4.12}$$

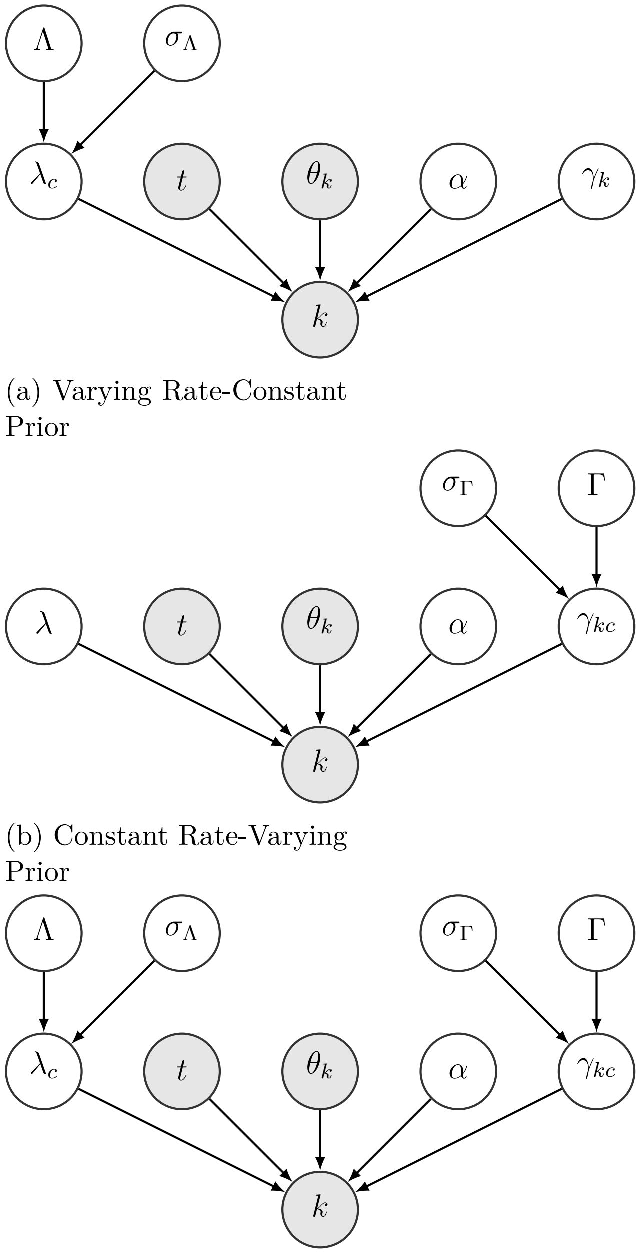
Lastly, we consider a full culture model (i.e., Varying Rate, Varying Prior—*VR-VP*) where culture simultaneously influences both rate of ELIs and knower level priors. The plate diagrams for these culturally influenced models are illustrated in Figure 4.2.

For each of these models, we infer the relevant rate and prior parameters/hyper-parameters using Bayes Rule. Considering the universal (CR-CP) model as an example,

$$P(\lambda, \gamma_k | k, t, \theta) \propto P(k | t, \theta, \lambda, \gamma_k) P(\lambda) P(\gamma_k). \tag{4.13}$$

4.2 Data Pre-processing

Our model serves as an explanatory vehicle linking the rate of ELIs and the knower level prior to the pattern of behavior children demonstrate when learning exact number, which provides us leverage to investigate culture's influence on the learning processes through those factors. Our model takes as input measurements of children's knower level and their age. The standard measure of children's knower level knowledge at a given age is revealed in the Give-N task (Wynn, 1990, 1992). In the Give-N task, children are presented with a pile of objects. On each trial, the experimenter



(a) Varying Rate-Constant Prior

(b) Constant Rate-Varying Prior

(c) Varying Rate-Varying Prior

Figure 4.2: The plate diagrams for our model comparisons.

Study	**Cultures**	**N kids**
Almoammer et al., (2013)	English	77
	Mandarin†	79
	Saudi Arabic†	83
Barner et al., (2009a)	English†	101
Barner et al., (2009b)	Japanese†	104
Boni et al. (unpublished)	Tsimané	100
Jara-Ettinger et al., (unpublished)	Tsimané	401
Krajcsi et al., (2018)	Hungarian	151
Marusic et al., (2016)	Slovenian†	343
Piantadosi et al., (2014)	Tsimané	92
Sarnecka et al., (2007)	English	91
	Japanese	48
	Russian	59
Wagner et al., (2016)	English	43

Table 4.1: Sources of compiled data. † denotes pre-processed knower level as opposed to raw Give-N data.

asks the child to give N of the objects to the experimenter and records how many objects the child hands over. The task is often titrated up so as to minimize the number of trials required to classify a child's knower level. While the rule to determine knower level varies slightly across studies, knower level classification seems to be robust to the form of the task (Lee & Sarnecka, 2011). Overall, the Give-N task is a conservative measure of children's number knowledge (Wagner et al., 2018).

We compiled data from ten cross-sectional studies detailed in Table 4.1, resulting in 1772 children across eight cultures and seven language families. To convert raw Give-N data to knower levels, we used the Bayesian data analysis model developed by Lee and Sarnecka (2010, 2011). See Appendix 4.A.2 for more details. Figure 4.3 shows the distribution of ages at which children are at each knower level for the different languages.

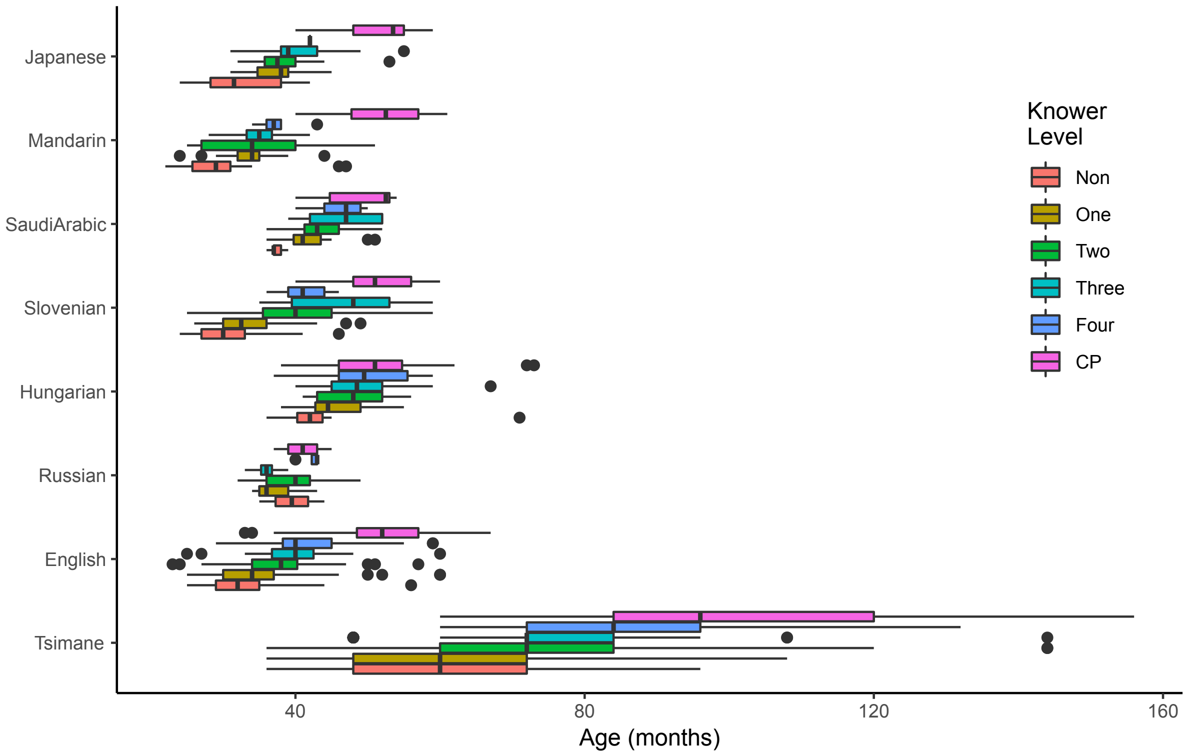


Figure 4.3: Distribution of ages at which children progress through knower level patterns across cultures.

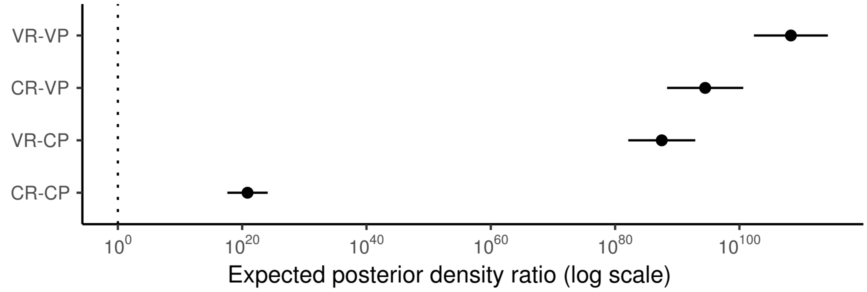


Figure 4.4: Expected log pointwise density ratio for each model (y-axis) relative to a multinomial baseline model (Equation 4.10). Higher ratio values reflect greater predictive ability. Line ranges reflect standard errors.

4.3 Results

To compare models, we use leave-one-out cross-validation with Pareto-smoothed importance sampling (Vehtari, Gelman, & Gabry, 2017). In standard cross-validation methods, the data is divided into k folds. The model is trained k times, leaving out a different fold to serve as a test set each time. The model's predictive performance on each held-out fold is then averaged to serve as a performance score. When k is less than the total number of data points, the resulting prediction estimates are slightly influenced by dependencies between how the data points were grouped into folds. This is avoided by setting k to the number of data points—i.e., leaving one data point out each fold. As training a model is computationally expensive, we do not calculate the leave-one-out score exactly; instead, we use Pareto-smoothed importance sampling to approximate the computation from values saved when initially fitting the model. The resulting predictive score is the expected log pointwise predictive density (elpd).

In Figure 4.4, we plot the ratio of the expected pointwise density for each model relative to the multinomial baseline model. Higher ratio values reflect better performance predicting a child's knower level given their age and culture. There are two comparisons worth particular attention. First, we check the performance of our simplest (6 parameter) number word learning model (CR-CP) against a standard multinomial baseline model (56 parameters) with full access to cultural information. This is illustrated in Figure 4.4 as the difference between the vertical baseline model and the CR-CP model. Compared to the baseline model (elpd: -2786; SE: 26), the universal (CR-CP) model has much greater predictive ability (elpd: -2738; SE: 33), suggesting children's acquisition of exact number words is well explained by our model. Next we evaluate how models with cultural specific rates or priors compare to the universal (CR-CP) model. As can be seen in Figure 4.4, there is strong evidence that culture has an influence on both the rate of ELIs and the inductive biases learner's bring to the task (VR-VP elpd:-2537; SE: 39).

Having validated the model's predictive ability, we can further explore the parameters of the full model. In Figure 4.5, we plot the prior over hypotheses for each culture as the odds of generating a knower level hypothesis relative to a non-knower hypothesis. Consistent with the simplicity prior used by Piantadosi et al. (2012), our inferred parameters suggest it is easier to generate a one-knower hypothesis than a two-knower hypothesis, easier to generate a two-knower hypothesis than a three-knower hypothesis and easier to generate a three-knower hypothesis than a four-knower hypothesis. Interestingly, our estimates suggest that learners are about equally as likely to be a four knower as a CP knower.

Turning to the rate of ELIs, Figure 4.6 plots the the posterior distribution of

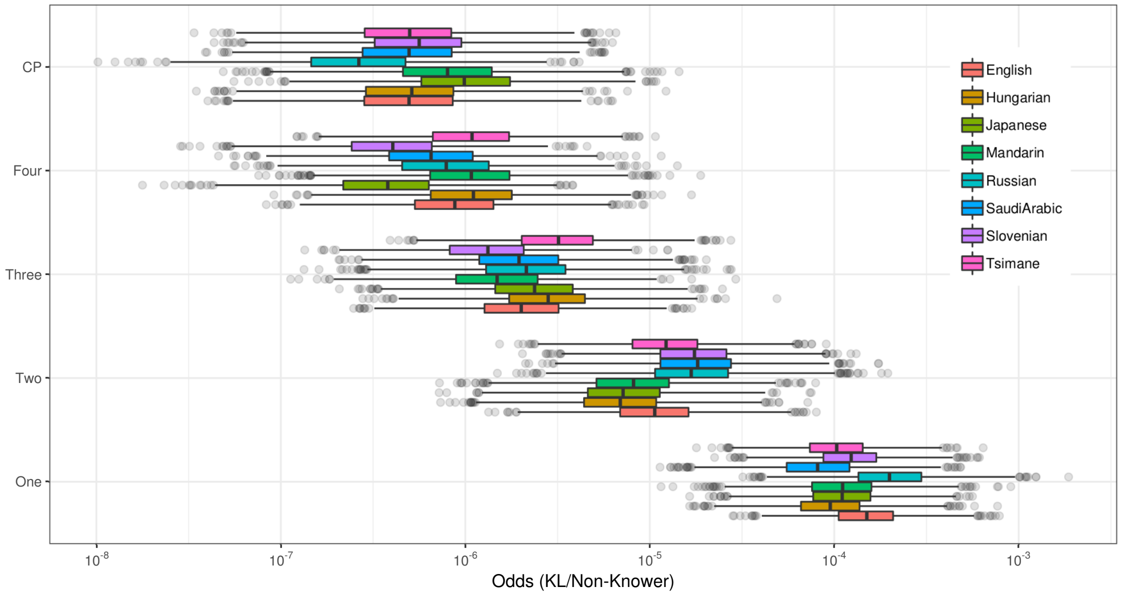


Figure 4.5: Inferred odds of generating a knower level hypothesis relative to a non-knower hypothesis.

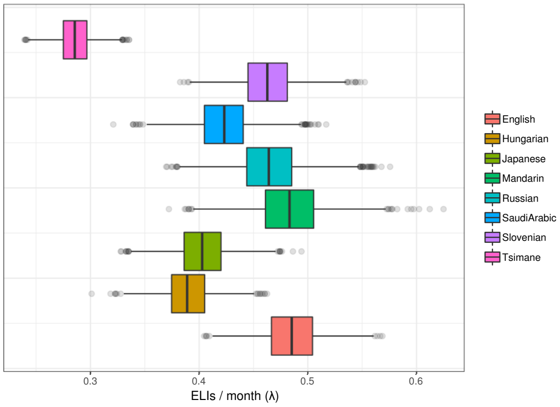


Figure 4.6: Inferred rates of effective learning instances for each culture.

rates for each culture. Surprisingly, the inferred parameters suggest that ELIs for number word learning are fairly infrequent occurring between once every four months to once ever other month. Perhaps less surprisingly, the WEIRD cultures pattern together at the upper end of the range; whereas, the Tsimané have a lower rate of ELIs. Future research is needed to understand how exactly culture influences the rate. That being said, the order of magnitude of rates is very similar to the rates of ELIs required for children to learn their earliest learned words (Mollica & Piantadosi, 2017a), potentially suggesting universal cognitive constraints on children's data usage.

4.4 Discussion

There is a whole wide world of environments and cultural needs. Our data analysis model provides us leverage to identify where culture might influence the learning process. Here, we found a role for culture in both how people generate hypotheses in the absence of data (i.e., priors) and how frequently people have ELIs. It's also worth noting that in spite of the cultural differences, we find support for the same inductive learning mechanism operating across cultures.

Looking at our best estimates for the inferred priors, we found a universal ordinal pattern in-line with a simplicity bias; however, the differences in simplicity varied across cultures suggesting that the same hypotheses might be more difficult to generate in one culture than another. A strong test of our finding for cultural differences in inductive bias would be to explicitly manipulate the probability of generating hypotheses by giving people a battery of mathematical tasks to increase the local utility of a particular underlying representation system (e.g., the approximate number system vs the count system). If the tasks that cultures value influence future learning, participants should carry a bias towards the previously useful system to learning a new task. Additionally, modeling people's behavior in a wide range of algorithmic tasks may also reveal these different biases.

Returning to our best estimates for rates, we find rates on the same order of magnitude. We cannot be sure that this order of magnitude is due to the model, the environment or the learner; however, intuition and corpus measurements would suggest that the WEIRD environments are similar; and the rates are on the same order of magnitude as those in a model-free analysis (Mollica & Piantadosi, 2017a). Therefore, while we acknowledge that there are complex interactions between environmental input, and the learner, we suggest that learning might be more constrained by the learner (e.g., attention, utility of task/data) than currently theorized. Without a clearer conceptualization of ELIs, we make fewer predictions for how to test for cultural differences in ELI rates.

There are several limitations of the current work. First, while there are attested cultural differences in the development of exact number word in the languages in our study, these languages are a small sample of the linguistic and numerical systems present throughout the world. Future research should examine how the learning process differs across more diverse numerical systems. Second, our model frames number word learning as inductively learning from the structures present in the world. However, it's highly unlikely that the data learners use is sampled randomly from the structure in the world as opposed to from an informed teacher. Given number's status as a cognitive artifact, learning numerical systems should be a fertile domain for examining how utility structures and pedagogical contexts differ across cultures and influence learning.

To conclude, even though our model comparison pointed toward culturally-specific processes in exact number word learning, there are universal components to learning. The rate of data is on the same order of magnitude suggesting fundamentally similar use of data and the priors all follow the same overall magnitude and relative proportion. Additionally, the shape of the priors is in line with an assumed model prior for simplicity suggesting that it's a driving force in learning number concepts.

4.5 Methods

We fit the models using adaptive Hamiltonian Monte-Carlo methods (Hoffman & Gelman, 2014; Betancourt, 2017) as implemented in Stan (Carpenter et al., 2017) and corresponding R package (Stan Development Team, 2018). For each model, we ran four chains for 2000 iterations steps discarding the first 1000 steps as warm-up. Visual inspection and R-hat metrics suggest the chains converged for all models. For model comparison, leave-one-out cross-validation with Pareto-smoothed importance sampling was implemented by the loo package (Vehtari, Gabry, Yao, & Gelman, 2018) in R (R Core Team, 2018).

4.A Appendix to Chapter 4

4.A.1 Mathematical Derivation of the model

The generative model in Figure 4.1a represents a joint probability distribution:

$$P(t,\lambda,n,\theta,d,h,k) = P(k|h)P(h|D)\prod_{d=0}^{n} P(d|\theta)P(\theta)P(n|t,\lambda)P(\lambda)P(t). \qquad \text{(A.1.1)}$$

To derive model predictions for knower level k as a function of age t, we want an analytical form for

$$P(k|t,\lambda,\theta) = \sum_{n}^{\infty}\sum_{h}^{\mathcal{H}}\sum_{d}^{n} P(k|h)P(h)P(n|t,\lambda)\prod_{d=0}^{n} P(d|\theta)P(d|h). \qquad \text{(A.1.2)}$$

Let's start with definitions. The probability of being a k knower give a hypothesis can be expressed as

$$P(k|h) = \delta_{kh}, \qquad \text{(A.1.3)}$$

where δ is Kroenecker's delta.

The probability of observing n ELIs given a rate of λ ELIs per month and an age of t months is given by

$$P(n|t,\lambda) \sim \text{Poisson}(\lambda \cdot t). \qquad \text{(A.1.4)}$$

There are two generative processes for the data that we need to consider: the probability of the world generating the data $P(d|\vec{\phi})$ and the probability of the learner's hypothesis generating the data $P(d|h)$. For convenience, we represent the processes of generating data from the world as the probability of the data generated from the world being consistent with a knower level $P(d|\theta_k)$, which is a binomial distribution. As a result, we rewrite the environmental distribution of data $\vec{\phi}$

$$D \sim \text{Multinomial}(\phi_w, n) \tag{A.1.5}$$

as

$$P(d|\theta_k) = \frac{n!}{d!(n-d)!}\theta_k^d \theta_k^{1-d}, \tag{A.1.6}$$

where

$$\theta_h = \sum_{w \in h} \phi_w. \tag{A.1.7}$$

The probability of generating data given a hypothesis is given by Equation 4.2. For convenience, we rewrite Equation 4.2 as a bernoulli distribution:

$$P(d|h) = w^d l^{1-d}, \tag{A.1.8}$$

where $l = \frac{(1-\alpha)}{10}$ and $w = \alpha + l$.

Turning back to A.1.2, we first push in the summations:

$$P(k|t,\lambda,\theta) = \sum_h P(k|h)P(h) \sum_n P(n|t,\lambda) \sum_d \prod_d P(d|\theta)P(d|h). \tag{A.1.9}$$

Now we make use of the binomial theorem and the identity of the exponential function to derive an analytical form for our marginalization over all possible amounts of data and all possible combinations of data type at every amount. Focusing first on the marginalization over all possible combinations of data, we can analytically define the sum as follows:

$$\begin{aligned}\sum_d \prod_{d=0}^{n} P(d|\theta)P(d|h) &= \sum_d \frac{n!}{d!(n-d)!}(\theta w)^d((1-\theta)l)^{n-d} \\ &= ((1-\theta)l)^n \sum_d \frac{n!}{d!(n-d)!}\left(\frac{\theta w}{(1-\theta)l}\right)^d \\ &= ((1-\theta)l)^n \left(1+\frac{\theta w}{(1-\theta)l}\right)^n \\ &\qquad \text{via the binomial theorem: } (1+x)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}x^k \\ &= ((1-\theta)l+\theta w)^n.\end{aligned} \tag{A.1.10}$$

Next we turn to the marginalization over all possible data amounts n and combinations of data types at each amount:

$$\begin{aligned}\sum_n P(n|t,\lambda)\sum_d \prod_d P(d|\theta)P(d|h) &= \sum_n \frac{(\lambda t)^n e^{-\lambda t}}{n!}((1-\theta)l+\theta w)^n \\ &= e^{-\lambda t}\sum_n \frac{(\lambda t)^n((1-\theta)l+\theta w)^n}{n!} \\ &= e^{-\lambda t}\sum_n \frac{(\lambda t[(1-\theta)l+\theta w])^n}{n!} \\ &= e^{-\lambda t}\cdot e^{\lambda t[(1-\theta)l+\theta w]} \text{ via the identity } e^z = \sum_{k=0}^{\infty}\frac{z^k}{k!} \\ &= e^{\lambda t\{[(1-\theta)l+\theta w]-1\}}.\end{aligned} \tag{A.1.11}$$

Rewriting Equation A.1.2 with our analytical solution for the marginalizations:

$$P(k|t,\lambda,\theta) = \sum_{h} \gamma_h \delta_{kh} e^{\lambda t\{[(1-\theta)l+\theta w]-1\}} \tag{A.1.12}$$

where $P(h) = \gamma_h$.

4.A.2 Converting raw Give-N data to Knower Level

To convert raw Give-N data to knower levels, we used the Bayesian data analysis model developed by Lee and Sarnecka (2010, 2011). Briefly, the model assumes that children have a base rate π at which they return a set of m objects regardless of instructions. When a child hears a prompt q, they will update the probability of returning a set of objects π'_m based on their knower level k as follows:

$$\pi'_m = \begin{cases} \pi_m & \text{if } q > k \\ v\pi_m & \text{if } q \leq k < CP \text{ and } m = q \\ \frac{1}{v}\pi_m & \text{if } q \leq k < CP \text{ and } m \neq q \\ v\pi_m & \text{if } k = CP \text{ and } m = q \\ \frac{1}{v}\pi_m & \text{if } k = CP \text{ and } m \neq q \end{cases} \tag{A.2.1}$$

where v is a free parameter reflecting the strength of this update. When faced with a query beyond their knower level, a child should respond according to the base rate (case one). When the query is within their subset range, they should up-weight the probability of a correct response (case two) and down-weight the probability of an incorrect response (case three). Similarly, if a child is a CP knower, they should up-weight the correct response (case four) and down-weight the incorrect responses (case five). A child's actual response is then a categorical draw from this updated distribution.

Using this model, we infer each child's knower level $k_{i=0}^{N}$, the population level strength of update v and the population level base rate π from the Give-N data. Before we analyse the inferred parameters, we provide a posterior predictive checks of the model in Figure 4.7. First, we can look to see how strongly the model captures the knower level pattern in the literature. The updated probability of responding with a set of objects π_m is plotted for each knower level (x-facet) and culture (y-facet) as shaded tiles. Darker shades reflect greater probability. Based on the shading, the model successfully predicts the knower level pattern reported in the literature. Now, we turn to how well the model assigns children to knower levels. The raw data points are over-layed with the maximum-a-posteriori assignment of children to knower levels. The size of the points reflects the number of trials. Consistent with Lee and Sarnecka (2010, 2011), we find the model's assignment of children to knower levels matches the pattern reported in the literature.

Having established the model provides a good account of the data, we can turn

	Evidence Strength	95% High Density Interval
English ($N = 211$)	29.13	24.03 - 35.44
Hungarian ($N = 151$)	30.57	25.12 - 41.87
Japanese ($N = 48$)	22.59	14.06 - 35.27
Russian ($N = 59$)	58.53	32.28 - 108.99
Tsimane ($N = 593$)	19.48	17.24 - 21.95

Table 4.2: Inferred evidence strength parameters

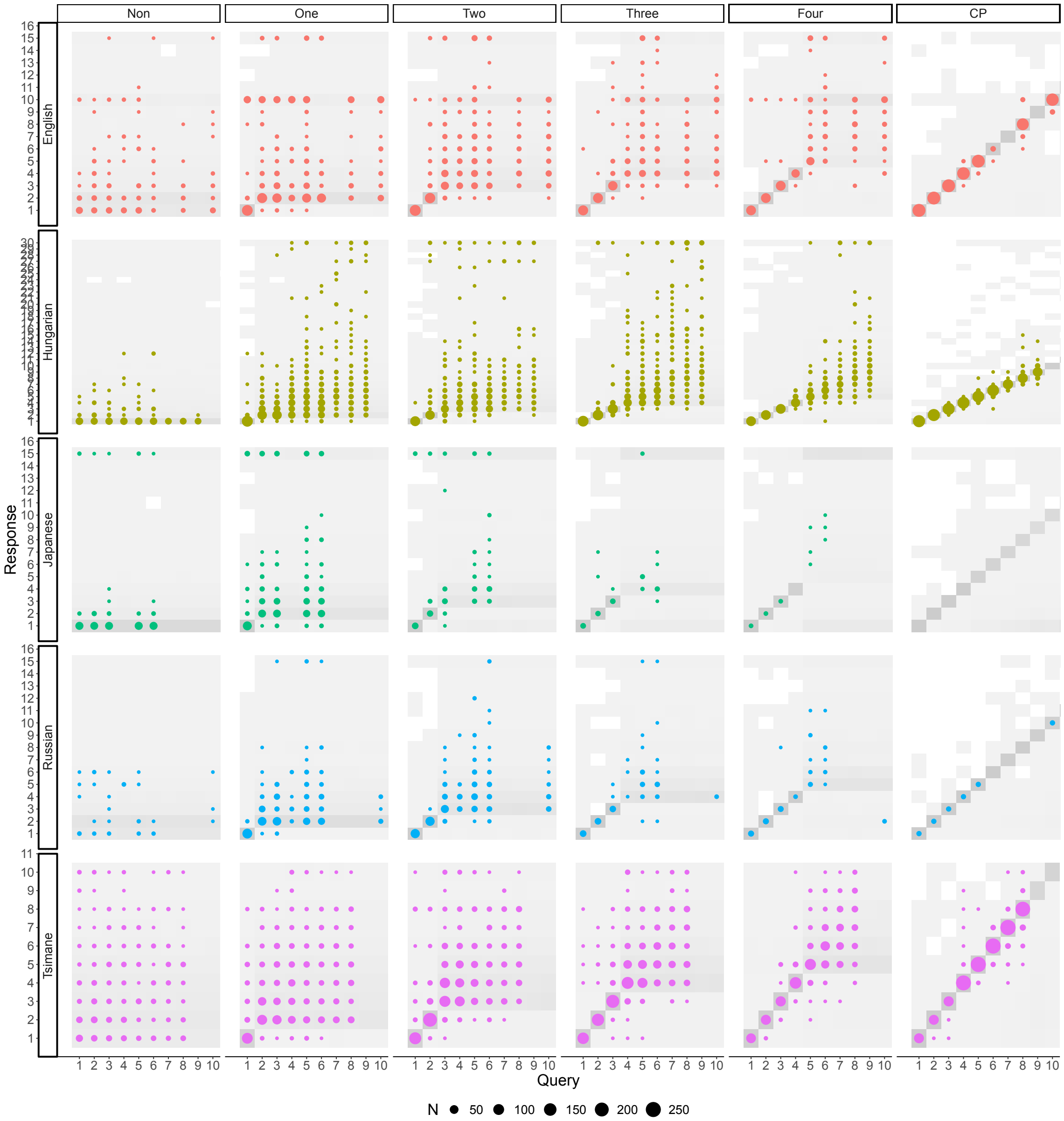


Figure 4.7: Posterior predictive check of knower level assignment. The shading in each facet reflects the posterior predictive distribution for the updated query-response matrix π_m for every knower level (x-facet) and culture (y-facet). Darker shades reflect greater probability. Points reflect the raw trial data under the MAP assignment of children to knower levels, with size reflecting the number of trials.

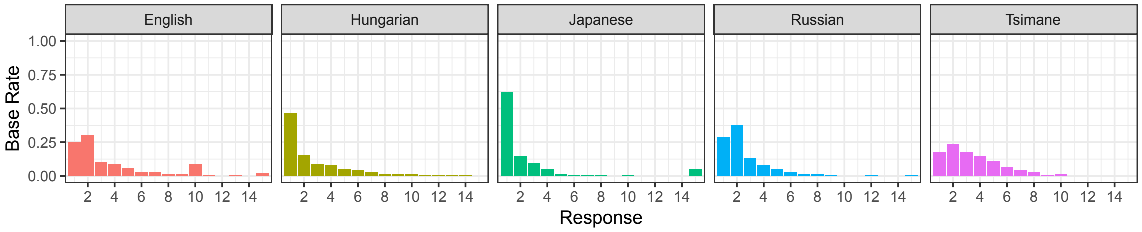


Figure 4.8: The inferred base rate response distribution π for children of different cultures in the Give-N task.

to the parameters of the model. The knower level distributions for each culture can be found in the main text Figure 4.3. Figure 4.8 shows the base rate parameter for each culture. Lee and Sarnecka (2010)'s study with English speaking children found that children preferred to respond with small numbers or with the maximum number available. We see a similar pattern across cultures; however, we only see children responding with the maximum number allowed for English and Japanese. For English, the bumps at 10 and 15 correspond to different studies that had 10 and 15 total objects, respectively. Perhaps there are different pragmatics to the task across cultures. With regard to the parameter for evidence strength, Lee and Sarnecka (2010) estimate for English children was approximately normal with a mean of 29.2 and standard deviation of 7.4, suggesting children's updating changes the probability of responses by a factor of 30. We find similar estimates for English and Hungarian, slightly lower estimates for Japanese and Tsimane, and much higher estimates for Russian (Table 4.2).

Chapter 5

Humans store about 1.5 megabytes of information during language acquisition

One of the foundational debates about human language centers on an issue of scale: is the amount of information about language that is learned substantial (empiricism) or minimal (nativism)? Despite theoretical debates on the how much of language is or can be learned, the general question of the amount of information that *must* be learned has yet to be quantified. Here, we provide an estimate of the amount of information learners must extract in the process of language acquisition. We provide *lower-bound*, *best guess*, and *upper-bound* estimates of this information, using a "back of the envelope" approach that is popular in physics. During the testing of the atomic bomb, physicist Enrico Fermi famously estimated the strength of the bomb by dropping scraps of paper as the blast passed. He noted that they were blown about 8 feet by the explosion and, after quickly computing in his head, announced that the blast was equivalent to about 10,000 tons of TNT (Schwartz, 2017). The true answer was 18,000 tons, meaning Fermi's crude experiment and quick calculation got the right answer to within a factor of two. Similar back-of-the-envelope *Fermi*-calculations are commonly used in physics as a sanity check on theories and computations.[1] However, such sanity checks are needed—although rarely applied—in fields that suffer from under-constrained theories, like psychology.

We apply this approach of rough estimation in order to quantify a lower-bound on the number of *bits per day* that language learners must extract and remember from their environments. While a substantial literature has focused on the differences between nativist and empiricist approaches, when they are translated into the domain of information theory, nativism and empiricism may blur together. Specifically, hard nativist constraints on sets of hypotheses may not necessarily provide much more information than the correct set of biases over an unconstrained space. Since we do not know much about which initial language learning biases children have, and how they interact with cognitive constraints, theories about how the hypotheses are constrained (or not) do not unambiguously determine the number of bits learners must store.

So, instead of debating nativism versus empiricism, we take up the challenge of quantifying how much information of language must be learned in order to inform the underlying acquisition theories. To avoid dependence on a particular representation scheme, we focus on the possible *outcomes* of learning a language, i.e., we compute the number of bits required to specify the target outcome out of a plausible space of logically possible alternatives. To avoid dependence on a particular learning algorithm, we study the problem abstractly without reference to *how* learning works,

[1] These computations are also used as a training exercise that allows surprising quantities to be approximated. An example is to compute the thickness of a car tire that is worn off with each rotation. Here's a hint: you can use your knowledge of how many miles car tires last for and how much thickness they lose over their lifetime.

but rather based on how much a relatively unbiased (e.g. maximum entropy) learner would have to store.

Our study is inspired by prior work which has aimed to characterize the capacity of human memory. Early literature approached the question of memory capacity from a neuroanatomical perspective. Upper bounds on memory capacity have been estimated via the number of synapses in cortex (10^{13} bits) or the number of impulses conducted in the brain across a lifetime (10^{20} bits) (Von Neumann, 1958). More recently, bounds for information transfer in neural coding have been estimated using information theoretic techniques (Borst & Theunissen, 1999; Bartol Jr et al., 2015). Working from behavioral performance, Landauer (1986) used a variety of techniques to estimate the number of bits of information humans must have in memory in order to show a given level of task performance. In one example, he converted accuracy in a recognition memory task to bits by imagining that each image was stored using a random code. This technique has been used recently by Brady, Konkle, Alvarez, and Oliva (2008) in a large-scale empirical study, which estimated that human memory could encode $2^{13.8}$ unique items. Even more recently, Ferrara, Furlong, Park, and Landau (2017) estimated 4 and 6 year old children's memory capacity to be $2^{10.43}$ unique items. Strikingly, both of these estimates lie within Landauer's estimated range of 10-14 bits per item. Landauer also used a dictionary study to estimate the number of words that Stanford students knew, and converted the estimates for a phonetic code into bits, requiring about $30 - 50$ bits per word (Landauer, 1986). All of his estimates converged on the same order of magnitude, suggesting that the "functional capacity" for human memory is on the order of 10^9 bits. A detailed critique of Landauer can be found in Hunter (1988), with a response given by Landauer (1988).

Our focus here is on estimating *linguistic* knowledge across multiple levels of structure and function: phonemes, wordforms, lexical semantics, word frequency and syntax. At each level, there is a large space of logically possible linguistic representations (e.g., acoustic cue values, syntactic parses). The challenge for learners is to discover and store which representations are used in their language. Tools in information theory allow us to estimate the relevant quantities. First, we assume that before learning, children begin with a certain amount of uncertainty over the required representation, R, denoted $\mathbf{H}[R]$. Shannon entropy quantifies the number of bits that must be received on average to remove uncertainty about what R is the true one (Shannon, 1948). After observing some data D, learners will have a new amount of uncertainty (perhaps zero) over R, denoted $\mathbf{H}[R \mid D]$. Note that here, D is not a random variable, but rather a specific value of data in learning a given language.

We can formalize the amount of information that D provides about R, here denoted $\mathbf{\Delta H}[R \mid D]$ as the difference between $\mathbf{H}[R]$ and $\mathbf{H}[R \mid D]$:

$$\mathbf{\Delta H}[R \mid D] = \mathbf{H}[R] - \mathbf{H}[R \mid D] = -\sum_{r \in R} P(r) \log P(r) + \sum_{r \in R} P(r \mid D) \log P(r \mid D). \tag{5.1}$$

This quantity, i.e. the reduction in entropy, gives the amount of information that D (e.g. data from learning) provides about a representation R.[2] Thus, in order to esti-

[2] The average of $\mathbf{\Delta H}[R \mid D]$ over D is the *mutual information* between R and D (Cover & Thomas, 2012).

Table 5.1: Summary of estimated bounds across levels of linguistic analysis

Section	Domain	Lower bound	Best Guess	Upper bound
5.1.1	Phonemes	375	750	1500
5.1.2	Phonemic Wordforms	200,000	400,000	640,000
5.1.3	Lexical Semantics	553,809	12,000,000	40,000,000
5.1.4	Word Frequency	40,000	80,000	120,000
5.1.5	Syntax	134	697	1394
	Total (bits)	794,318	12,481,447	40,762,894
	Total per day (bits)[4]	121	1,900	6204

mate the amount of information learners must have acquired, it suffices to estimate their uncertainty before learning, $\mathbf{H}[R]$, and subtract from it their uncertainty after learning $\mathbf{H}[R \mid D]$. The resulting value will tell us the number of bits of information that the learning data D has provided.[3]

5.1 Results

We will build up our estimates separately for each linguistic domain. The results of each section are summarized in Table 5.1. Table 5.2 summarizes the key assumptions behind each of our estimations.

5.1.1 Information about Phonemes

Our phonemic knowledge enables us to perceive discrete linguistically-relevant sounds, or phonemes, from rich high-dimensional but noisy speech signals. Before a child

[3] In the case of continuous distributions, Equation 5.1 has continuous analogs where the sums turn into integrals.

[4] For this value, we assume language is learned in 18 years of 365 days.

knows the sounds of their language, they have uncertainty over the acoustic features of speech sounds. After learning their language, children have much less uncertainty over the acoustic features of speech sounds as they now have several acoustic cues to help them identify phonemes. Following the above logic, the decrease in the amount of uncertainty children have about where their speech sounds lie in acoustic space is the amount of information they must now store about phoneme cues.

Identifying linguistically relevant acoustic cues has proven challenging for scientists, as there is no obvious invariance, or uniquely identifying component, in speech sounds. For our estimation, we analyze the information stored for three well studied acoustic cues: voice onset time (VOT) in ms—a cue to voiced-voiceless distinctions (e.g., the difference between /p/ and /b/), central frication frequency in barks—a cue to the place of articulation for fricatives (e.g., the difference between /f/ and /s/), and the first two formant frequencies of vowels in mels—a cue to vowel identity. We assume that initially learners have maximum uncertainty along each cue R, following uniform distributions bounded by the limits of perception. In this case, each $r \in R$ has an equal probability of $P(r) = 1/(B - A)$, giving

$$\mathbf{H}[R] = -\int P(r) \log P(r) \; dr = \log(B - A), \tag{5.2}$$

where B and A are respectively the upper and lower bounds of perception. For VOT, we assume the range is -200 to 200 ms. For frequencies, we assume bounds on human hearing of 20 to $20,000$ Hz, which translate to $0.2 - 24.6$ in barks and $32 - 3817$ in mels. As a measure of the uncertainty over the cue dimension after learning, we will assume that speaker's knowledge is captured by a normal prior distribution, giving $\mathbf{H}[R \mid D]$ as

$$\mathbf{H}[R \mid D] = \int N(x \mid \mu, \sigma) \log N(x \mid \mu, \sigma)\, dx = \frac{1}{2} \log(2\pi e \sigma^2) \tag{5.3}$$

where N is a normal probability density function,[5] and μ and σ are the participants' inferred mean and standard deviation. To find σ for real humans, we use the values inferred by (Table 7; Kronrod, Coppess, & Feldman, 2016) to account for the perceptual magnet effect.[6]

We find that language users store 3 bits of information for voiceless VOT, 5 bits for voiced VOT, 3 bits for central frication frequency and 15 bits for formant frequencies. As these acoustic cues are only a subset of the cues required to identify consonant phonemes, we assume that consonants require three cues; each cue requiring 5 bits of information. For vowels, we do not adjust the 15 bits of information conveyed by formant frequencies. As a best guess, again paying attention to primarily the order of magnitude, we assume there are 50 phonemes each requiring 15 bits, totaling 750 bits of information. For lower and upper estimates, we introduce a factor of two error [375-1500 bits].

5.1.2 Information about Wordforms

When dealing with wordforms, the first challenge is to define a "word," a term which could be used to refer to lemmas, phonological forms, families of words, senses, etc.

[5] Using a normal distribution with the domain truncated to our perceptual bounds does not change our estimate.

[6] For vowels, we extend these distributions to their multidimensional counterparts as formant space is (at least) two dimensional.

Entire dissertations could be (and have been) written on these distinctions. These difficulties are in part why the Fermi-approach is so useful: we don't need to make strong theoretical commitments in order to study the problem if we focus on rough estimation of orders of magnitude. Estimates of the number of words children acquire range on the order of $20,000$-$80,000$ total wordforms (Anglin, Miller, & Wakefield, 1993). However, when words are grouped into families (e.g. "dog" and "dogs" are not counted separately) the number known by a typical college student is more in the range of $12,000 - 17,000$ (Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995; D'Anna, Zechmeister, & Hall, 1991)—although see Brysbaert et al. (2016) for an estimate over twice that size. Lexical knowledge extends beyond words too. Jackendoff (1997) estimates that the average adults understands $25,000$ idioms, items out of the view of most vocabulary studies. Our estimates of *capacity* could of course be based on upper-bounds on what people *could* learn, which, to our knowledge, have not been found. Looking generally at these varied numbers, we'll use an estimate of $40,000$ as the number of essentially unique words/idioms in a typical lexicon.

The most basic thing each learner must acquire about a word is its phonemic wordform, meaning the sequence of phonemes that make up its phonetic realization. If we assume that word forms are essentially memorized, then the entropy $\mathbf{H}[R \mid D]$ is zero after learning—e.g. for all or almost all words, learners have no uncertainty of the form of the word once it has been learned. The challenge then is to estimate what $\mathbf{H}[R]$ is: before learning anything, what uncertainty should learners have? To answer this, we can note that $\mathbf{H}[R]$ in Equation 5.1 can be viewed as an *average* of the negative log probability of a wordform, or $-\log P(R)$. Here, we use a language model to estimate the average negative log probability of the letter sequences that make up words and view this as an estimate of the amount of entropy that has been removed for *each* word. In other words, the average surprisal of a word under a language model provides one way to estimate the amount of uncertainty that learners who know a given word must have removed.[7][8]

To estimate these surprisals, we used the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), we computed the surprisal of each word under 1-phone, 2-phone, 3-phone and 4-phone models (see Manning & Schütze, 1999) trained on the lexicon. This analysis revealed that 43 bits per word on average are required under the 1-phone, 33 bits per word under the 2-phone, 24 under the 3-phone and 16 under the 4-phone model. Noting the sharply decreasing trend, we will assume a lower bound of about 5 bits per word to store the phonetic sequence, a "best guess" of 10 bits per word and an upper bound of 16 as in the 4-phone.[9] When our best guess is multiplied by the size of the lexicon ($40,000$ words), that gives an estimate of $400,000$ $[200,000 - 640,000]$ bits of lexical knowledge about the phonetic sequences in words.

[7] In this view, we neglect the complexity of the language model, which should be a much smaller order of magnitude than the number of bits required for the entire lexicon.

[8] Analogously, we can view the surprisal as the number of bits that must be remembered or encoded for a particular outcome—e.g. to learn a specific wordform.

[9] As the average word length in this database is ~ 7.5 phonemes, this gives slightly over one bit per phoneme.

5.1.3 Information about Lexical Semantics

The information contained in lexical semantics is difficult to evaluate because there are no accepted theories of semantic content, or conceptual content more generally (Laurence & Margolis, 1999). However, following Fermi, we can make very simplified assumptions and try to estimate the general magnitude of semantic content. One way to do this is to imagine that the set of word meanings are distributions in an N-dimensional semantic space. If we assume that the entire space is a Gaussian with standard deviation R and the standard deviation of an individual word meaning is r, then we can compute the information contained in a word meaning as the difference in uncertainty between a N-dimensional Gaussian with radius R as compared to one with radius r. The general logic is shown in Figure 5.1. The "space" shown here represents the space of semantic meanings, and words are viewed as small distributions in this space covering the set of things in the extension of the word's meaning. Thus, when a leaner acquires a word like *accordion*, they know that the word refers to some relatively small subset (of size r) of possible objects, but they may not be certain on the details (Does the extension cover harmoniums? Concertinas? Bayans?). The reduction in entropy from a total semantic space of size R—no idea what a word means—to one of size r is what we use to approximate the amount of information that has been learned.

Equation 5.3 above gives the change in entropy for a one-dimensional Gaussian. However, the dimensionality of semantic space is considerably larger. In the case of an N-dimensional Gaussian, with independent dimensions and constant, or homoge-

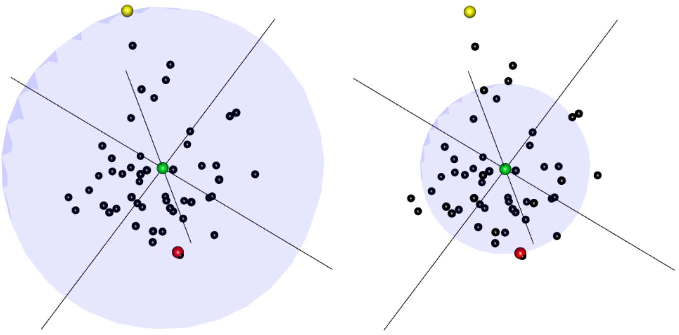


Figure 5.1: The shaded spheres represent uncertainty in semantic space centered around a word (in green). Left: The uncertainty is given with respect to the word's farthest connection in semantic space (in yellow), representing R. Right: The uncertainty is given with respect to the N^{th} nearest neighbor of the word (in red), representing r. The reduction in uncertainty from R to r reflects the amount of semantic information conveyed by the green word.

neous standard deviation σ in each dimension, the entropy is given by:

$$\mathbf{H}[R] = \frac{N}{2}(1 + \log 2\pi + \log \sigma). \tag{5.4}$$

This means that if we go from an R standard deviation Gaussian to an r standard deviation one, the amount of information we have learned is the difference between these entropies,

$$\mathbf{\Delta H}[R \mid D] = \frac{N}{2}(1 + \log 2\pi + \log R) - \frac{N}{2}(1 + \log 2\pi + \log r) = \frac{N}{2} \log \frac{R}{r} \tag{5.5}$$

We estimate R and r in several different ways by looking at WordNet (Fellbaum, 1998) to determine the closeness of each word to its neighbors in semantic space. In particular, we take r to be a characteristic distance to nearby neighbors (e.g. the closest neighbors), and R to be a characteristic distance to far away ones (e.g. the

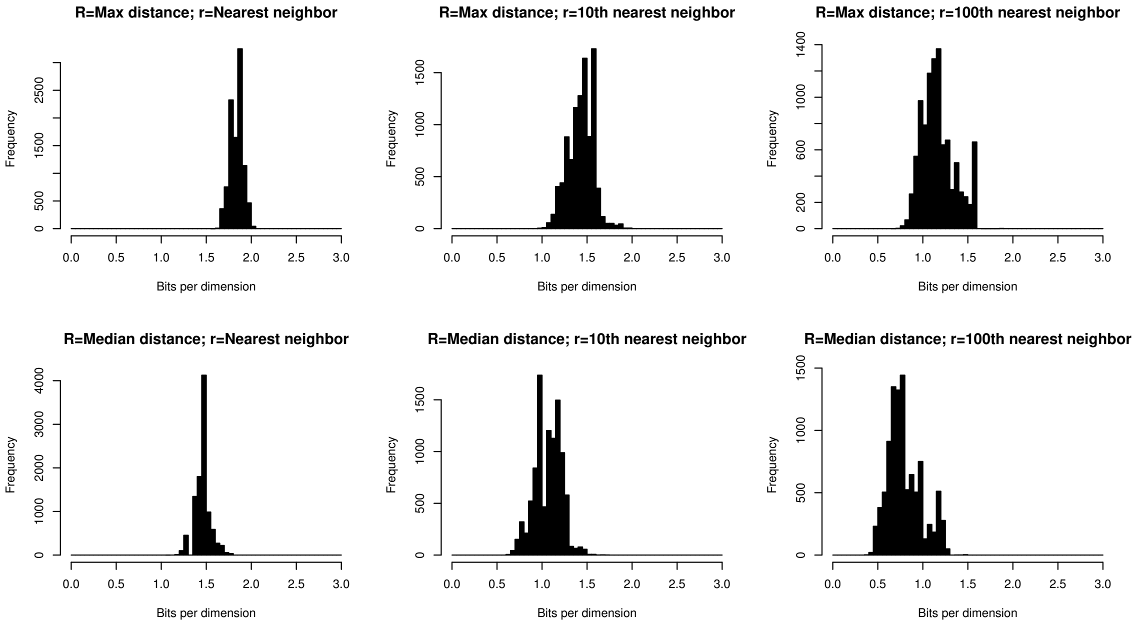


Figure 5.2: Histograms showing the number of bits-per-dimension ($\frac{1}{2}\log\frac{R}{r}$) for various estimates of R and r. These robustly show that $0.5 - 2.0$ bits are required to capture semantic distances.

max distance). Note, that this assumes that the size of a Gaussian for a word is about the same size as its distance to a neighbor, and in reality this may *under*-estimate the information a word meaning contains because words could be much more precise than their closest semantic neighbor.

Figure 5.2 shows $\frac{1}{2}\log\frac{R}{r}$ for several estimates of R and r for $10,000$ random nouns in WordNet. The likely values fall within the range of $0.5 - 2.0$ bits. Because we are plotting $\frac{1}{2}\log\frac{R}{r}$ and not $\frac{N}{2}\log\frac{R}{r}$, these values may be interpreted as the number of *bits per dimension* that lexical semantics requires. For instance, if semantic space was one-dimensional then it would require $0.5 - 2.0$ bits per word; if semantic space were 100-dimensional, lexical semantics would require $50 - 200$ bits per word. The nearness of these values to 1 means that even continuous semantic dimensions can be viewed as *approximately* binary in terms of the amount of information they provide about meaning.

The dimensionality of semantic space has been studied by Landauer and Dumais (1997) and Mandera, Keuleers, and Brysbaert (2017), with numbers ranging from $100-500$ dimensions. Our best guess will use 1 bit per dimension and 300 dimensions following Landauer and Dumais (1997) for $12,000,000$ bits. Our upper bound uses 2 bits-per-dimension and 500 dimensions for a total of $40,000,000$ bits.

For our lower-bound in this domain, we may pursue a completely alternative technique, which, surprisingly, gives a similar order of magnitude as our best guess. If there are $40,000$ lexical items that must be learned, we can assume that they correspond to $40,000$ distinct concepts (a la *principle of contrast* (E. V. Clark, 1987)). In the "most nativist" case, favored by Fodor (Fodor, 1975), we could assume that there are a corresponding $40,000$ meanings for these words that learners innately have. In this case, the problem of learning is figuring out which of the 40000! pairings of words and concepts is the correct one. It will take $\log_2(40000!) \approx 553,809$ bits of information to specify which of these is correct. We will use this as our lower-bound. While this seems like an unmanageable task for the child, it's useful to imagine how much information is conveyed by a single pedagogical learning instance. Our estimate is derived by a combinatorial argument: to choose the first word's meaning, there are N choices, the second there $N-1$, and so on. The total number of choices is therefore $N...(N-1)...(N-2)...(N-40000) = N!/(N-40000)!$. So if initially $N = 400000$ $(553, 809$ bits), there will be $N = 39999$ $(553, 794$ bits) after one correct mapping is learned, meaning that a single pedagogical instance rules out $39, 999$ possible pairings or, equivalently conveys 15.29 bits.

5.1.4 Information about Word Frequency

Word frequencies are commonly studied in psychology as factors influencing language processing and acquisition (e.g., Forster & Chambers, 1973; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; Zorzi, Houghton, & Butterworth, 1998; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Murray & Forster, 2004) as well as for their peculiar Zipfian distribution (Piantadosi, 2014b). However, relatively little work has examined the fidelity of people's representation of word frequency, which is what is required in order to estimate how much people know about them. In one extreme, language users might store perhaps only a single bit about word frequency, essentially allowing them to categorize high vs. low frequency words along a median split. On the other extreme, language users may store information about word frequency with higher fidelity—for instance, 10 bits would allow them to distinguish 2^{10} distinct levels of word frequency as a kind of psychological floating point number. Or, perhaps language learners store a full ranking of all $40, 000$ words in terms of frequency, requiring $\log(40000!) = 553, 809$ bits of information.

In an experimental study, we asked participants from Mechanical Turk ($N = 251$) to make a two-alternative forced choice to decide which of two words had higher

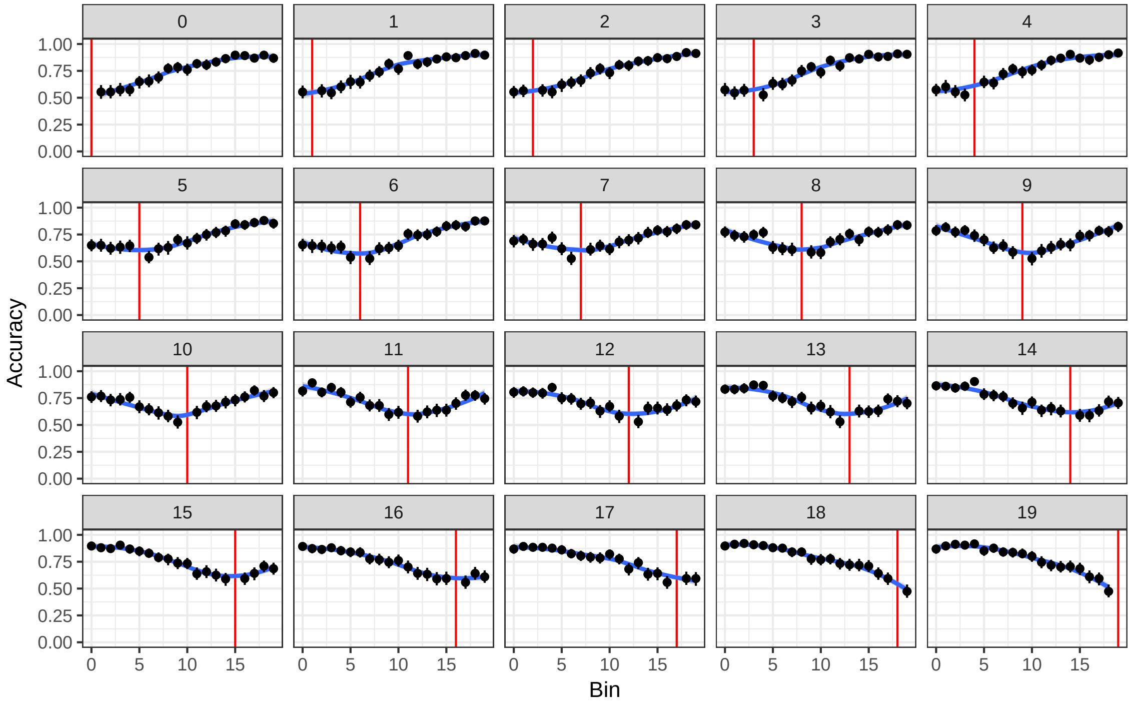


Figure 5.3: Accuracy in frequency discrimination accuracy as a function of log word frequency bin faceted by log reference word frequency bin. Vertical red lines denote within bin comparison. Line ranges reflect 95% bootstrapped confidence intervals.

frequency[10]. Words were sampled from the lexical database SUBTLEX (Brysbaert & New, 2009) in 20 bins of varying log frequency. We removed words below the bottom 30'th percentile (frequency count of 1) and words above the upper 99'th percentile in word frequency in order to study the intermediate-frequency majority of the lexicon. Each participant completed 190 trials.

Participants' accuracy in answering is shown in Figure 5.3. The i'th subplot shows participants' accuracy (y-axis) in distinguishing the i'th bin from each other j'th bin, with the red line corresponding to $i = j$. This shows, for instance, that

[10] Participants were instructed that we were interested in their first impression and that there was no need to look up word frequencies.

people are poor at distinguishing very close i and j (near the red line), as should be expected.

Participants' overall accuracy in this task was 76.6%. Neglecting the relatively small difference in accuracy (and thus fidelity) with a words' absolute frequency, this accuracy can be modeled by imagining that participants store M levels of word frequencies. Their error rate on this task will then be given by the probability that two words fall into the same bin, or $1/M$. Setting $1/M = 1 - 0.766$ gives $M \approx 4$, meaning that participants appear to track approximately four categories of frequencies (e.g. high, medium-high, medium-low, low). This trend can also be observed in Figure 5.3, where the flat bottom of the trough in each plot is approximately 5 bins wide, meaning that each bin cannot be well distinguished from its 5 nearest neighbors, giving a total effective number of bins for participants as $20/5 = 4$.

If $M = 4$, then participants would only need to learn $\log 4 = 2$ bits of information about a word's frequency, as a best guess. This would yield a total of $2 \cdot 40,000 = 80,000$ bits total across the entire lexicon. We construct our lower- and upper-bounds by introducing a factor of two error on the computation (e.g. per word lower bound is 1 bit and upper is 3 bits). It is important to note that by assuming objective frequency rankings, our estimate is conservative. If we could analyze participants' responses with regard to their subjective frequency rankings, we would expect to see greater accuracy reflecting higher resolution representations of frequency.

5.1.5 Information about Syntax

Syntax has traditionally been the battleground for debates about how much information is built in versus learned. Indeed, syntactic theories span the gamut from those that formalize a few dozen binary parameters (K. Wexler & Manzini, 1987; Kohl, 1999) to ones that consider alternative spaces of infinite models (e.g., Perfors et al., 2011; Chater & Vitányi, 2007) or data-driven discovery from the set of all parse trees (Bod, Scha, Sima'an, et al., 2003). In the face of massively incompatible and experimentally under-determined syntactic theories, we aim here to study the question in a way that is as independent as possible from the specific syntactic formalism.

We do this by noting that every ordinary English speaker's knowledge of syntax provides enough information to correctly parse every sentence of English. In many cases, the sentences of English will share syntactic structure. However, we can imagine a set $s_1, s_2, \ldots, s_n$ of sentences which share as little syntactic structure as possible between each s_i and s_j. For instance,

Bill [met John]. [Jill's sister] cried. (5.6)

both have three words but have largely non-overlapping syntactic structures due to the use of a transitive verb in the first and a possessive and intransitive verb in the second. We will call theses "essentially independent" sentences when they share almost no syntactic structure. In this case, the bits specifying these parses can be added together to estimate the total information learners know. If the sentences were not essentially independent in terms of syntactic structure, information from one sentence would tell us how to parse information from another, and so adding together the information for each would be an over-estimate of learners' knowledge.

We assume that learners who do not know anything about a parse of a sentence s_i start with a maximum entropy distribution over each parse, assigning each an equal probability of one over the number of logically possible parses of s_i, so that

$$\mathbf{H}[R] = -\sum_{r \in R} \frac{1}{\#parses} \log \frac{1}{\#parses} = \log(\#parses). \tag{5.7}$$

We assume the knowledge of an adult leaves zero uncertainty in general, yielding $\mathbf{H}[R \mid D] = 0$ so that

$$\boldsymbol{\Delta}\mathbf{H}[R \mid D] = \mathbf{H}[R] - \mathbf{H}[R \mid D] = \log(\#parses) \tag{5.8}$$

for a single sentence s_i. In general, the number of logically possible parses can be computed as the number of binary trees over s_i, which is determined only by the length of s_i. The $(l-1)$'th Catalan number gives the number of possible binary parses for a sentence of length l. Then, the number of bits required to specify *which* of these parses is correct is given by $\log C_{l-1}$. The Catalan numbers are defined by

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{5.9}$$

As an example, to determine each of (5.6), knowledge of syntax would have to specify $\log C_2 = 1$ bit, essentially specifying whether the structure is $((\circ\,\circ)\,\circ)$ or $(\circ\,(\circ\,\circ))$.

But C_n grows exponentially—for instance, $C_{10} = 16796$, requiring 14 bits to specify which parse is correct for an 11-word sentence.

Looking at a collection of sentences, if s_i has length $l(s_i)$, then the total amount of information provided by syntax will be given by

$$\sum_{i=1}^{n} \log C_{(l(s_i)-1)}. \tag{5.10}$$

Again, Equation 5.10 assumes that there is no syntactic structure shared between the s_i—otherwise Equation 5.10 over-estimates the information by failing to take into account the fact that some bits of information about syntax will inform the parses of distinct sentences. Our upper and lower bounds will take into account uncertainty about the number of distinct sentences s_i that can be found.

To estimate the number of such sentences, we use the textbook linguistic examples studied by Sprouse and Almeida (2012). They present 111 sentences that are meant to span the range of interesting linguistic phenomena and were presented independently in Adger (2003). Our best estimate is therefore Equation 5.10 taking s_i to be the lengths of these sentences. We take the lower-bound to be Equation 5.10 where $l(s_i)$ is *half* the sentence length of s_i, meaning that we assume that only half of the words in the sentence participate in a structure that is independent from other sentences. For an upper bound, we consider the possibility that the sentences in Sprouse and Almeida (2012) may not cover the majority of syntactic structures, particularly when compared to more exhaustive grammars like Huddleston, Pullum, et al. (2002). The upper bound is constructed by imagining that linguists could perhaps construct two times as many sentences with unique structures, meaning that we should double our best guess estimate. Notably, these tactics to bound the estimate do not qualitatively change its size: human language requires very little information about syntax, 697 [134 – 1394] bits. In either case, the number is much smaller than most other domains.

5.2 Discussion

Summing across our estimates for the amount of information language users store about phonemes, wordforms, lexical semantics, word frequency and syntax, our best guess and upper bound are on the order of 10 million bits of information, the same order as Landauer (1986)'s estimate for language knowledge. It may seem surprising but, in terms of digital media storage, our knowledge of language almost fits compactly on a floppy disk. The best-guess estimate implies that learners must be remembering 1000-2000 bits *per day* about their native language, which is a remarkable feat of cognition. Our lower bound is around a million bits, which implies that learners would remember around 120 bits each day from birth to 18 years. To put our lower estimate in perspective, each day for 18 years a child must wake up and remember, perfectly and for the rest of their life, an amount of information equivalent to the information in this sequence,

0110100001101001011001000110010001100101011011100110000101100011
01100011011011110111001001100100011010010110111101101110

Naturally, the information will be encoded in a different format—presumably one which is more amenable to the working of human memory. But in our view, both the lower and best-guess bounds are explainable only under the assumption that language is grounded in remarkably sophisticated mechanisms for learning, memory, and inference.

There are several limitations to our methods, which is part of the reason we focus on orders of magnitude rather than precise estimates. First, our estimates are rough and require simplifying assumptions (listed in Table 5.2). Second, there are several domains of linguistic knowledge whose information content we do not estimate here including word predictability, pragmatic knowledge, knowledge of discourse relations, prosodic knowledge, models of individual speakers and accents, among others. Many of these domains are difficult because the spaces of underlying representations are not sufficiently well formalized to compute information gains (e.g. in pragmatics or discourse relations). In other areas like people's knowledge of probable sequences of words, the information required is difficult to estimate because the same content can be shared between constructions or domains of knowledge (e.g. the knowledge that "Mary walks" and "John walks" are high probability may not be independent from each other, or from knowledge about the lexical semantics of the verb). We leave the estimation of the amount of information language users store about these domains of language to further research.

Importantly, our estimates vary on orders of magnitude across levels of representation. These differences could suggest fundamental differences in the learning mechanism for specific language learning problems. As these analyses show, the majority of information humans store about language is linked to words, specifically lexical semantics, as opposed to other systems of language knowledge, such as phonemes and syntax. In fact, the estimate for syntax is of a similar order of magnitude proposed by some nativist accounts, in that the number of bits required for syntax is in the hundreds, not tens of thousands or millions. To illustrate, if syntax learning is principally completed in the first 5 years, children would have to learn a single bit about syntax every 2-3 days on average. Despite this, the possible outcomes for learners in our best guess for syntax consists of $2^{697} \approx 10^{210}$ different systems. In other words, learners still would need the ability to navigate an immense space of possibilities, far greater than the number of atoms in the universe ($\sim 10^{80}$). In the other areas of language, even more enormous hypothesis spaces are faced as well, pointing to the existence of powerful inferential mechanisms.

Turning back to nativism and empiricism, it is unfortunate that the majority of the learnability debates have centered on syntactic development, which requires far less information in total than even just a few word meanings. Despite the remarkable mechanisms that must be deployed to learn hundreds of thousands or millions of bits about lexical semantics, there are *no viable accounts* of lexical semantics representation and learning, either from empiricists or nativists—despite some claims by Fodor (1975). Our results suggest that if any language-specific knowledge is innate, it is most likely for helping tackle the immense challenge of learning lexical semantics, rather than other domains with learnability problems that require orders of magnitude less information.

Table 5.2: The assumptions we make in our estimates.

Section	Domain	Assumptions
5.1.1	Phonemes	1. The language system must contain information about acoustic cues to phoneme identity. 2. The maximum entropy over the frequency dimension can be represented as a uniform distribution over audible frequency ranges. 3. The maximum entropy over the VOT dimension can be represented as a uniform distribution ranging from -200 to 200 ms. 4. The variance in language users' representations of acoustic cues for phonemes can be well approximated by normal distributions following (Kronrod et al., 2016).
5.1.2	Phonemic Wordforms	1. The language system favors compression of statistical co-occurrences. 3. The cost of specifying a language model over phonemes is negligible. 4. Adult language users have a lexicon of $40,000$ lexical entries. 5. The sample of words we used to induce our estimate is an adequate approximation to the adult lexicon.
5.1.3	Lexical Semantics	1. Semantic space can be represented as a multivariate normal distribution with independent dimensions. 2. The maximum entropy over the space can be approximated by a normal distribution whose standard deviation is the maximum distance between words. 3. What learners come to know about the semantics of words narrows the distribution over semantic space based on distance to the nearest semantic neighbor. 4. Adult language users have a lexicon of $40,000$ lexical entries. 5. Our sample of words is a decent approximation to the distances of the average word.
5.1.4	Word Frequency	1. Errors in word frequency discrimination are a result of insufficient representational resolution. 2. Subjective frequency rankings are well approximated by objective frequency rankings (via corpus statistics). 3. Adult language users have a lexicon of $40,000$ lexical entries. 4. The sample of words we used in our experiment are representative of the words in the adult lexicon.
5.1.5	Syntax	1. The language system must contain information to uniquely identify one binary parse tree from all possible binary parse trees. 2. The maximum entropy over syntactic parses is given by the number of binary parse trees. 3. Sentences from (Sprouse & Almeida, 2012) are a good approximation/coverage of the essentially independent syntactic components of English grammar.

Chapter 6

Discussion

Children's early word use has the potential to inform our understanding of language acquisition and conceptual development. Nonetheless, early word use data remains significantly under-utilized in research endeavors for at least two important reasons. The first is that measuring rapid acquisition is difficult. Even weekly visits to the lab may miss milestones in the life of a word. The second is that there are multiple accounts of word learning that have the potential to predict the observed data and can only be adjudicated by addressing the questions of how the data are used, the degree to which maturational constraints affect learning, the nature of abstraction and generalization from data, the inductive biases of the learner and the environmental constraints on learning. We argued that this requires a first-principles computational approach, which involves specifying each component of learning and observing how these components interact to explain and predict behavior, to identify boundary conditions and efficiency trade-offs and to determine fruitful research directions. We introduce such an approach and apply it in two protracted developmental domains: *kinship*, which shows similar patterns as children's earliest concrete nouns (Haviland & Clark, 1974; Keil, 1989; Benson & Anglin, 1987; Greenfield & Childs, 1977; Price-Williams et al., 1977), and *number*, where it is generally agreed that conceptual development constrains adult-like lexical acquisition (Carey, 2009; Wynn, 1990, 1992).

The goal of this thesis was to develop the hypothesis that the systematic patterns of children's word use over the course of development are the natural consequence of a sophisticated inductive learning mechanism operating with insufficient data—i.e., rational construction (Xu, 2007, 2016, in press). The goal was accomplished by i) demonstrating that this learning mechanism matches common patterns of early word use (Chapter 2), ii) developing a linking hypothesis to make model predictions as a function of time instead of data (Chapter 3), iii) evaluating an implementation of the learning mechanism against a large, empirical dataset (Chapter 4) and iv) further justifying the sophistication of the learning mechanism with information-theoretic estimates of the size of the learning problem (Chapter 5).

In Chapter 2, we applied our model framework to the task of children acquiring kinship terms. In addition to demonstrating that our model has the capacity to learn any kinship terms system with cross-cultural simulations, we illustrate that under-extension (E. V. Clark, 1973; Brown, 1973), over-extension (Rescorla, 1980; Anglin et al., 1993), the characteristic-to-defining shift (Landau, 1982; Keil, 1989; Keil & Batterman, 1984) and the order of acquisition (Haviland & Clark, 1974; Benson & Anglin, 1987) are all explained by our model. In our model, under-extension is explained by the local data distribution of kinship terms favoring unique, initial word-referent mappings. Over-extension is the consequence of balancing abstraction and the complexity of the abstracted function. The characteristic-to-defining shift is the natural consequence of over-extension in environments where characteristic over-extensions capture most of the possible data but cannot completely and exclusively capture all the possible data. The order of acquisition of kinship terms can be explained either by the environmental availability of data (as suggested by Benson & Anglin, 1987) or the construction of inter-related systems, but not through a natural bias for simplicity alone (as suggested by Haviland & Clark, 1974).

In Chapter 3, we undertook the challenge of relating the amount of data utilized by a learner (e.g., as specified by a model) to empirical measurements of children's age. In the process, we tackled a longstanding question in developmental science: Is the acquisition of the lexicon primarily delayed by maturation (Markman, 1990; Borer & Wexler, 1987) or driven by learning (Hoff, 2003; J. C. Goodman, Dale, & Li, 2008; Huttenlocher et al., 1991; Shneidman et al., 2013; Weisleder & Fernald, 2013; Hidaka, 2013)? Across 14 different languages in Wordbank (Frank et al., 2015), we determined that early lexical acquisition is primarily driven by data. Further, our model inferred the profile of children's data use: children utilize on the order of 10 effective learning instances when acquiring an early word, not a single instance (e.g., Carey & Bartlett, 1978) nor on the order of hundreds or thousands of instances (e.g., Siskind, 1996; Yu & Ballard, 2007; Yu & Smith, 2007); the instances that children made use of occurred relatively infrequently—once every other month; and children start paying attention to data for word learning relatively early—around two months old for comprehension. Further, the model evaluations in Chapter 3 are a significant methodological contribution, providing a broadly applicable, generative, linking hypothesis between data amounts and time and a model-free baseline for learning. Future work should utilize these techniques to assess the temporal viability of learning mechanism under different models of the environment.

In Chapter 4, we utilized the waiting time models constructed in Chapter 3 to evaluate an exact number word learning implementation of the model framework sketched in Chapter 2 (Piantadosi et al., 2012). We demonstrate that our first-principles learning model predicts the data better than the standard off-the-shelf multinomial regression analysis model. Additionally, we utilize our model as a descriptive Bayesian data analysis (Tauber et al., 2015) to learn about learning. Specifically, how does culture influence the timing of exact number word learning? Culture has the potential to influence learning through the availability of effective learning instances and through the inductive biases children have learned for approaching new tasks. We demonstrate that the cross-cultural differences in timing are better explained by differences in both the rate of effective learning instances and the biases guiding children's hypothesis generation.

Finally, in Chapter 5, we provide convergent evidence for the complexity of our rational constructivist model framework from an information-theoretic description of the memory requirements for language. Regardless of how a learner acquires linguistic representations, they must store enough information about their language to distinguish representations in their language from representations not in their language. Chapter 5 estimated this amount of information to be surprisingly small (about 1.5 MB) compared to the size of current artificial language models. Importantly, our *conservative* estimate allocates 69% of the stored linguistic information to lexical semantics (96% in our *best guess* estimate). In light of these estimates, we argue that if the acquisition of any linguistic representation required a complex learning mechanisms, the scale of the problem suggests lexical acquisition should be a prime candidate.

6.1 The human learning machine

The ultimate task of cognitive science is to identify the *representations* and *processes* that support our rich mental lives. The biggest challenge in uncovering the mental models that support everyday cognition is not comparing models, but *generating* the models to compare. Model selection techniques require specifying the space of possible models and the evaluation metrics for these models. In practice, we never have the space of all possible models for a single task. Further, current techniques have little applicability when theories are *non-identifiable*—i.e., when they make exactly the same empirical predictions. Without formalization of the problem, it is often unclear when and for what measures theories are non-identifiable. Therefore, the exploration of the space of possibilities is just as important and valuable a contribution as empirical work or model comparisons focused on falsification.

A major contribution of this thesis is illustrating the power of first-principles computational models as explanatory vehicles. A first principles account of a behavior begins with established[1] postulates and describes the consequences for behavior. When the problem is formalized, then in addition to making clear predictions, and

[1] Of course, first principle accounts will make assumptions, but they should test the robustness of their assumptions.

evaluation metrics for a particular task, constraints and predictions about alternative evaluation metrics and measures are often made explicit. A thorough exploration of a cognitive model involves identifying the first-principles behind the model, the evaluation metrics for the model, the linking hypothesis between models and empirical measures, and the conditions under which the model is tractable. Consider the human learning machine.

Proposition 1 *Every model of learning requires the specification of a space of possible representations and a rule for updating representations based on observing data.*

The models in Chapters 2 and 4 build a representational space of computer programs. This space is motivated by the first-principles that learners possess rich compositional machinery and a set of "core knowledge" primitive computations (Carey, 2009). Our update rule is Bayes rule: consisting of both a prior inductive bias over hypotheses and a likelihood function. Our prior is grounded by the observation that people learn simpler rules faster than complex ones (J. Feldman, 2000). Our likelihood is motivated by sampling assumptions (Tenenbaum, 1999). It's important to note that any representation space and data update rule are implicit statements about inductive biases for how a model should generalize (Mitchell, 1980). Chapter 2 demonstrated how our choices result in the behavioral patterns that we observe in children for kinship and Piantadosi et al. (2012) does so for number.

Proposition 2 *Every model of learning will need data-analysis assumptions or a linking hypothesis (Tanenhaus, 2004; Teller, 1984).*

Data analysis assumptions arise out of epistemic concerns—i.e., concerns about the adequacy of measures. On the other hand, linking hypotheses are motivated by first-principles. The uncertainty given by a linking hypothesis is aleatoric in nature—i.e., reflecting the true unknowable uncertainty of the world[2]. Both data analysis assumptions and linking hypotheses are essential for scientific progress but acknowledging which are being used and allocating blame accordingly when models do not fit data is essential for meaningful progress moving forward. In Chapter 3, we develop the linking hypothesis between data amounts in models to empirical times of acquisition. We show that compared to logistic regression—the standard data analysis assumption for this data—waiting time models provide the best explanation for the mean and the variance of the observed data.

Proposition 3 *If humans learn with the intent of successfully achieve goals in the world, they needs to model the environment/context (Conant & Ross Ashby, 1970; Scholten, 2010).*

The importance of the environment/context has not been lost on cognitive science (e.g., Anderson & Schooler, 1991; H. H. Clark, 1992, 1996). Modeling a cognitive agent requires both describing their environment and delineating their conceptualization of the environment/context, which need not be accurate. In addition to efficiently generating novel predictions across contexts, an analysis of the environments in which a first-principle model must operate has the potential to alleviate the computational intractability of working with first-principles cognitive models directly

[2] This is not to say there cannot be a completely deterministic linking hypothesis or that we should not be concerned with measurement error!

(Van Rooij, 2008; Van Rooij & Wareham, 2008). In Chapter 2, our explicit models of the environment allowed us to demonstrate that the order of acquisition effects in kinship are only predicted under certain environments, but not by the learner internalizing the data-generating process of the environment. In the number domain, the differences in timing across cultures could not be reduced to differences in the input distribution because ostensibly there are none (Dehaene & Mehler, 1992). Similarly, without our measurement of informants' kinship learning contexts in Chapter 2, we would not have found our novel account of the characteristic-to-defining shift and might have proposed fundamental changes in learning mechanisms or representations as was common in the literature (e.g., Kemler, 1983). With regard to a cognitive agent's internal conceptualization of a context, environmental/contextual constraints are important for understanding the redundancy and sufficiency of a representation and the tractability of a process. In our models, we explicitly represent the context as a mental construct (family tree) and, further, incorporate the context into our representation space: programs take as input contexts, which allows them to work in any context (see also Katz et al., 2008). Future work should explore the extent to which mental notions of the contexts 1) are constructed, 2) reflect the environment and goals of the cognitive agent, 3) are incorporated in conceptual representations, and 4) influence their deployment.

6.2 The frontier of the concept-language interface

This dissertation has focused primarily on developmental constraints for the concept-language interface, including conceptual primitives and compositional machinery, likelihoods, inductive biases, and reflections of the environment (e.g., via internal structures). Looking to the future, the concept-language interface will be an exciting frontier for progress. Language for communication is predicated on the goals of updating shared mental representations between interlocutors and inspiring illocutionary force, i.e., action in the world. Neither of these goals exist meaningfully without a rich notion of context—be it the state of the world or the common ground between individuals. The properties of the world and the goal structures of agents, provide promising sources of contextual constraint that will make first-principle model exploration more tractable. Further, the role of contextualized conceptual representations will be vital to capturing the multiple affordances of concepts (e.g., generalization, reasoning, simulation) prevalent in our everyday cognitive lives. In my own work, the interface between language and concepts in context has informed our understanding of how mental reflections of the context are constructed (Mollica & Piantadosi, 2015; Yan, Mollica, & Tanenhaus, 2018; Rubio-Fernández, Mollica, Oraa Ali, & Gibson, 2019) and how conceptual representations are flexibly deployed in language processing (Rubio-Fernández, Mollica, & Jara-Ettinger, under review; Register, Mollica, & Piantadosi, under review). Nevertheless, we have only scratched the surface of how cognition reflects the environment/contexts of use and the consequences of this reflection for mental representations and processes.

References

Adger, D. (2003). *Core syntax: A minimalist approach* (Vol. 33). Oxford University Press Oxford.

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., ODonnell, T., Barner, D., et al. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, 201313652.

Ambridge, B. (2018, Jul). *Against stored abstractions: A radical exemplar model of language acquisition.* PsyArXiv. Retrieved from `psyarxiv.com/gy3ah` doi: doi: 10.31234/osf.io/gy3ah

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of child language*, *42*(02), 239–273.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, i–186.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). Celex2 ldc96l14. *Web Download. Philadelphia: Linguistic Data Consortium.*

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology*, *60*(1), 40–62.

Barner, D., Chow, K., & Yang, S.-J. (2009). Finding ones meaning: A test of the relation between quantifiers and integers in language development. *Cognitive psychology*, *58*(2), 195–219.

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from japanese. *Journal of experimental child psychology*, *103*(4), 421–440.

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

Bartol Jr, T. M., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., & Sejnowski, T. J. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *Elife*, *4*, e10778.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, *21*(1), 85–123.

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

Beller, S., Bender, A., Chrisomalis, S., Jordan, F. M., Overmann, K. A., Saxe, G. B., & Schlimm, D. (2018). The cultural challenge in mathematical cognition. *Journal of Numerical Cognition*, *4*(2), 448–463.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of child language*, *6*(2), 183–200.

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

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bergelson, E., & Swingley, D. (2015). Early word comprehension in infants: Replication and extension. *Language Learning and Development*, *11*(4), 369–380.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

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.

Bod, R., Scha, R., Sima'an, K., et al. (2003). *Data-oriented parsing.* Citeseer.

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

Borer, H., & Wexler, K. (1987). *The maturation of syntax.* Springer.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature neuroscience*, *2*(11), 947–957.

Bowerman, M. (1974). Early development of concepts underlying language. In *Language perspectives: Acquisition, retardation, and intervention* (pp. 191–209). University Park Press.

Bowerman, M. (2007). Constructing topological spatial categories in first language acquisition. *The categorization of spatial entities in language and cognition*, *20*, 177.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

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

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, *41*(4), 977–990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participants age. *Frontiers in psychology*, *7*.

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

Carey, S. (1978). *The child as word learner.* na.

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

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

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition*, *80*(1-2), 179–213.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

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

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

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 learningof natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

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

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

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

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

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, *1*, 33.

Clark, H. H. (1992). *Arenas of language use.* University of Chicago Press.

Clark, H. H. (1996). *Using language.* Cambridge university press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204.

Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*(2), 89–97.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Literacy Research*, *23*(1), 109–122.

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.

Dehaene, S. (1997). *The number sense.* Oxford University Press.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review*, *115*(1), 1.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

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

Epps, P., Bowern, C., Hansen, C. A., Hill, J. H., & Zentz, J. (2012). On numeral complexity in hunter-gatherer languages. *Linguistic Typology*, *16*(1), 41–109.

Everett, D., Berlin, B., Gonalves, M., Kay, P., Levinson, S., Pawley, A., . . . Everett, D. (2005). Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current anthropology*, *46*(4), 621–646.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: evidence from infants manual search. *Developmental Science*, *6*(5), 568–584.

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*(3), 295–313.

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.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Fellbaum, C. (1998). *Wordnet.* Wiley Online Library.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *Macarthur-bates communicative development inventories.*

Ferrara, K., Furlong, S., Park, S., & Landau, B. (2017). Detailed visual memory capacity is present early in childhood. *Open Mind*, *1*(3), 136–147.

Fodor, J. A. (1975). *The language of thought.* Harvard University Press.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627–635.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2015). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language.*

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, *108*(3), 819–824.

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.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

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

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

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

Gerstenberg, T., & Goodman, N. (2012). Ping pong in church: Productive use of concepts in human probabilistic inference. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development*, *1*(1), 23–64.

Gleitman, L. R., & Trueswell, J. C. (2018). Easy words: Reference resolution in a malevolent referent world. *Topics in cognitive science.*

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of child language*, *17*(1), 171–183.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

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

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language*, *35*(3), 515–531.

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.

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

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

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

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*(4), 725–743.

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.

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child development*, 1021–1034.

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

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PloS one*, *8*(11), e76242.

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

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

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

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*(5), 1368–1378.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

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

Huddleston, R., Pullum, G. K., et al. (2002). The cambridge grammar of english.

Language. Cambridge: Cambridge University Press, 1–23.

Hunter, L. (1988). Estimating human cognitive capacities: A response to Landauer. *Cognitive Science*, *12*(2), 287–291.

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

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

Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.

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

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

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 115.

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

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

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

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

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

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

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

Kemp, C., & Jern, A. (2014). A taxonomy of inductive problems. *Psychonomic bulletin & review*, *21*(1), 23–46.

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.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The goldilocks effect in infant auditory attention. *Child development*, *85*(5), 1795–1804.

Kohl, K. T. (1999). *An analysis of finite parameter learning in linguistic spaces* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Kushnir, T., & Xu, F. (2012). *Rational constructivism in cognitive development* (Vol. 43). Academic Press.

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.

Landauer, T. K. (1986). How much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477–493.

Landauer, T. K. (1988). An estimate of how much people remember, not of underlying cognitive capacities. *Cognitive Science*, *12*(2), 293–297.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. *Concepts: core readings*, 3–81.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438.

Le Corre, M., Li, P., Huang, B. H., Jia, G., & Carey, S. (2016). Numerical morphology supports early number word learning: Evidence from a comparison of young mandarin and english learners. *Cognitive psychology*, *88*, 162–186.

Lee, M. D., & Sarnecka, B. W. (2010). A model of knower-level behavior in number concept development. *Cognitive science*, *34*(1), 51–67.

Lee, M. D., & Sarnecka, B. W. (2011). Number-knower levels in young children: Insights from bayesian modeling. *Cognition*, *120*(3), 391–402.

Leslie, A. M., Xu, F., Tremoulet, P. D., & Scholl, B. J. (1998). Indexing and the object concept: developingwhat'andwhere'systems. *Trends in cognitive sciences*, *2*(1), 10–18.

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

Lewis, M., & 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).

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

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review*, *124*(6), 762.

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

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

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

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

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

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

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

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*(6619), 813–815.

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

Martí, L., Mollica, F., Piantadosi, S. T., & Kidd, C. (2016). What determines human certainty? In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 698–703).

Martí, L., Mollica, F., Piantadosi, S. T., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind*, *2*(2), 47-60. doi: doi: 10.1162/opmi_a_00017

Marušič, F., Plesničar, V., Razboršek, T., Sullivan, J., Barner, D., et al. (2016). Does grammatical structure accelerate number word learning? evidence from learners of dual and non-dual dialects of slovenian. *PloS one*, *11*(8), e0159208.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631.

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

Mitchell, T. M. (1980). *The need for biases in learning generalizations.* Department of Computer Science, Laboratory for Computer Science Research .

Modyanova, N., & Wexler, K. (2007). Semantic and pragmatic language development: Children know thatbetter. In *Proceedings of the 2nd conference on generative approaches to language acquisition–north america (galana 2)* (pp. 297–308).

Mollica, F., & Piantadosi, S. (in prep). *Universal and cultural-specific processes in exact number word acquisition.*

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. (2017a). How data drive early word learning: A cross-linguistic waiting time analysis. *Open Mind*.

Mollica, F., & Piantadosi, S. T. (2017b). An incremental information theoretic buffer supports sentence processing. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 805–810).

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

Mollica, F., & Piantadosi, S. T. (in revision). *Logical word learning: The case of kinship.*

Mollica, F., Piantadosi, S. T., & Tanenhaus, M. K. (2015). The perceptual foundation of linguistic context. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1613–1618).

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

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

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

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*(3), 721.

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

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.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, *14*(1), 11–28.

Núñez, R. E. (2017). Is there really an evolved capacity for number? *Trends in cognitive sciences*, *21*(6), 409–424.

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

Oey, L., Mollica, F., & Piantadosi, S. T. (2018). Adults use gradient similarity information in compositional rules. In *Proceedings of the 40th annual meeting of the cognitive science society.*

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

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

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

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

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

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

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

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

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

Piantadosi, S. T. (2014b). Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.

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., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, *17*(4), 553–563.

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.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, *103*(1), 56.

Plummer, M., et al. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop* *on distributed statistical computing* (Vol. 124, p. 125).

Posner, J. K., & Baroody, A. J. (1979). Number conservation in two west african societies. *Journal of Cross-Cultural Psychology*, *10*(4), 479–496.

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

Quine, W. V. O. (1960). Word and object (studies in communication). *New York and London: Tech-nology Press of MIT*.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

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

Register, Y., Mollica, F., & Piantadosi, S. T. (under review). *Semantic verification is flexible and sensitive to context.*

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

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* MIT press.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., … others (2006). The human speechome project. In *Symbol grounding and beyond* (pp. 192–196). Springer.

Rubio-Fernández, P., Mollica, F., & Jara-Ettinger, J. (under review). *Why searching for a blue triangle is different in english than in spanish.*

Rubio-Fernández, P., Mollica, F., Oraa Ali, M., & Gibson, E. (2019). How do you know that? automatic belief inferences in passing conversation. *Cognition*, *193*, 104011.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of one,two, and threein english, russian, and japanese. *Cognitive psychology*, *55*(2), 136–168.

Saxe, G. B. (1988a). Candy selling and math learning. *Educational researcher*, *17*(6), 14–21.

Saxe, G. B. (1988b). The mathematics of child street vendors. *Child Development*, 1415–1425.

Saxe, G. B., & Posner, J. (1983). The development of numerical cognition: Cross-cultural perspectives. *The development of mathematical thinking*, 291–317.

Schneider, R. M., Daniel, Y., & Frank, M. C. (2015). Large-scale investigations of variability in childrens first words. In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2110–2115).

Scholten, D. (2010). A primer for conant and ashbys good-regulator theorem. *Unpublished*.

Schwartz, D. N. (2017). *The last man who knew everything: The life and times of enrico fermi, father of the nuclear age*. Hachette UK.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 623–656.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of child language*, *40*(03), 672–686.

Shore, C. (1986). Combinatorial play, conceptual development, and early multiword speech. *Developmental Psychology*, *22*(2), 184.

Shultz, T., Thivierge, J.-P., & Laurin, K. (2008). Acquisition of concepts with characteristic and defining features. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

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

Smiley, P., & Huttenlocher, J. (1995). Conceptual development and the child's early words for events, objects, and persons. *Beyond names for things: Young children's acquisition of verbs*, 21–61.

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

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. *Psychological science*, *18*(1), 79–87.

Snedeker, J., Geren, J., & Shafto, C. L. (2012). Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children. *Cognitive Psychology*, *65*(1), 39–76.

Spiegel, C., & Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. *Journal of experimental child psychology*, *109*(1), 132–140.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, *48*(03), 609–652.

Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.17.3)

Storkel, H. L. (2001). Learning new wordsphonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1321–1337.

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

Tanenhaus, M. K. (2004). On-line sentence processing: past, present and, future. *On-line sentence processing: ERPS, eye movements and beyond (eds M. Carreiras & C. Clifton)*, 371–392.

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.*

Teller, D. Y. (1984). Linking propositions. *Vision research*, *24*(10), 1233–1246.

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.

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Childrens acquisition of duration words. *Cognitive psychology*, *78*, 57–77.

Tillman, K. A., Marghetis, T., Barner, D., & Srinivasan, M. (2017). Today is tomorrows yesterday: Childrens acquisition of deictic time words. *Cognitive psychology*, *92*, 87–100.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

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.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games:

Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

Van Rooij, I., & Wareham, T. (2008). Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *The Computer Journal*, *51*(3), 385–404.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.* Retrieved from `https://CRAN.R-project.org/package=loo` (R package version 2.0.0)

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Von Neumann, J. (1958). *Tbe computer and the brain.* Yale University Press.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer.

Wagner, K., Chu, J., & Barner, D. (2018). Do children's number words begin noisy? *Developmental science*, e12752.

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.

Waxman, S. R. (1999). Specifying the scope of 13-month-olds' expectations for novel words. *Cognition*, *70*(3), B35–B50.

Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive psychology*, *43*(3), 217–242.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

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

Wexler, K. (1999). Maturation and growth of grammar. *Handbook of child language* *acquisition*, 55–109.

Wexler, K., & Manzini, M. R. (1987). Parameters and learnability in binding theory. In *Parameter setting* (pp. 41–76). Springer.

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

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220–251.

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

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

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

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

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

Yan, S., Mollica, F., & Tanenhaus, M. K. (2018). A context constructivist account of contextual diversity. In *Proceedings of the 40th annual meeting of the cognitive science society.*

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414–420.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Literacy Research*, *27*(2), 201–212.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* New York: Addison-Wesley.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? a connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1131.