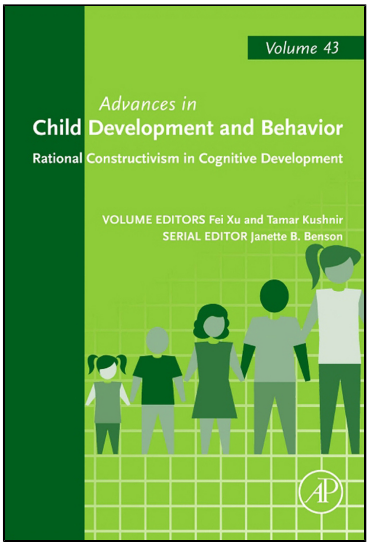

From Dawson, C. R., & Gerken, L., (2012). Can Rational Models Be Good Accounts of Developmental Change? The Case of Language Development at Two Time Scales. In J. B. Benson (Serial Ed.) & F. Xu & T. Kushnir (Vol. Eds.), *Rational Constructivism in Cognitive Development* (pp. 95–124). Elsevier Inc.: Academic Press.
ISBN: 9780123979193

Academic Press

CHAPTER FOUR

Can Rational Models Be Good Accounts of Developmental Change? The Case of Language Development at Two Time Scales

Colin R. Dawson* and LouAnn Gerken†,1

*School of information: Science, Technology, and Arts 1040 E. 4th Street University of Arizona, Tucson, AZ 85721-0077, USA
†Department of Psychology 1503 E University Blvd. University of Arizona, Tucson, AZ 85721-0068, USA
[1]Corresponding author: E-mail: gerken@email.arizona.edu

Contents

Abstract

Rational models of human perception and cognition have allowed researchers new ways to look at learning and the ability to make inferences from data. But how good are

Advances in Child Development and Behavior, Volume 43
ISSN 0065-2407,
http://dx.doi.org/10.1016/B978-0-12-397919-3.00004-6

Rational Constructivism in Cognitive Development, First Edition, 2012, 95–124

such models at accounting for developmental change? In this chapter, we address this question in the domain of language development, focusing on the speed with which developmental change takes place, and classifying different types of language development as either *fast* or *slow*. From the pattern of fast and slow development observed, we hypothesize that rational learning processes are generally well suited for handling fast processes over small amounts of input data. In contrast, we suggest that associative learning processes are generally better suited to slow development, in which learners accumulate information about what is typical of their language over time. Finally, although one system may be dominant for a particular component of language learning, we speculate that both systems frequently interact, with the associative system providing a source of emergent hypotheses to be evaluated by the rational system and the rational system serving to highlight which aspects of the learner's input need to be processed in greater depth by the associative system.

1. INTRODUCTION

All theories of language development, indeed all theories of cognitive development more generally, seek a balance between what knowledge about the likely structure of the world needs to come with the learner (i.e. must be innate) and the computational power of the learning mechanism needed to encode and analyze the learner's experiences. Until the last decade, theories of language development were essentially of two sorts, which we might view as having extreme values on the innate knowledge and on the mechanistic complexity scales.

On the one hand, *triggering* accounts posit that linguistic structure is innate, with some aspects of structure shared by all languages and other aspects varying parametrically across languages (Chomsky, 1957; Chomsky & Lasnik, 1993). This view assumes a very simple learning mechanism in which the child can determine which of a set of parameterized linguistic structures is valid for her language by encountering a single, specific input example or *trigger*. A similar mechanism is thought to be a work in ducklings and goslings that follow the first moving object they see after hatching (Lorenz, 1935).

On the other hand, *associative* accounts, often instantiated in connectionist network models (e.g. Rumelhart & McClelland, 1987) posit little in the way of innate knowledge, except for separate encoding of information arising from the different sensory systems. However, these accounts assume that learners can store large amounts of information over which a variety of statistical trends, relating any number of input dimensions, can be induced. An example of a statistical trend might be that the majority of

words in English begin with a stressed syllable (e.g. Cutler & Carter, 1987; Jusczyk, Cutler, & Redanz, 1993).

In addition to differing in the amount of innate knowledge and computational capacity they assume of learners, triggering and associationist accounts differ as to the basis on which learners generalize from previously encountered stimuli to new stimuli. Triggering accounts assume that generalization occurs in an all-or-none fashion, via a model of the language system, or *grammar*, whereas associationist accounts assume that generalization occurs in a gradient fashion via a measure of similarity which is induced from statistical patterns in the data. The two accounts are at opposite extremes in terms of the explicitness of the representations they posit. While triggering accounts take the mental representation of grammar to be a set of discrete rules, associationist accounts reject the notion of an identifiable representation of grammar, supposing that what appears to the linguist to be a grammar is really just a collection of statistical relationships.

Both triggering and associationist accounts fail to comport with certain aspects of language development data. We explore these failures in more detail in the subsequent sections; however, we can briefly identify two types of problems that undermine both approaches. The first concerns the nature of generalization. Triggering accounts predict that once the learner encounters the relevant data (the trigger) that indicates the correct setting for a particular aspect of the child's grammar, that aspect of the grammar should have been learned. Furthermore, once a generalization is made, it should not be changed. These predictions are contradicted by the fact that mature language abilities develop over time with errors gradually decreasing (e.g. Elman, 2003; Freudenthal, Pine, Aguado-Orea, & Gobet, 2007) and by data showing that children flexibly change the generalizations they make (e.g. Gerken, 2010). At the other extreme, associationist accounts predict that large quantities of data should be required for generalization. In several cases, however, children learn from just a few examples (one of the observations that motivated triggering-style accounts in the first place, e.g. Gerken & Bollt, 2008).

The second class of problem concerns the nature of the input that is required for generalization to take place. Several recent studies have found that infants can learn patterns that are linguistically unnatural, which is at odds with triggering (e.g. Cristià, Seidl, & Gerken, 2011; Gerken & Bollt, 2008). On the other hand, children appear to generalize very little from input that contains many tokens from a single type, but given an equal number of tokens overall distributed over several types, they generalize well (Xu & Tenenbaum, 2007a). Connectionist accounts often make the

incorrect prediction that repeated exposure to a single type will result in overlearning of the properties of that one category and will swamp learning about other categories that had taken place previously.

In recent years, a middle way, which at least partially addresses all these problems, has emerged between the two extremes (e.g. Frank & Tenenbaum, 2011; Gerken & Dawson, in press; Perfors, Tenenbaum, & Regier, 2006; Xu, 2007). This middle way is rational statistical inference, which, like triggering accounts, assumes that generalization from old experiences to new ones occurs via a grammar. Unlike triggering, however, rational inference does not assume that a highly constrained set of possible grammars needs to be innate. Rather, a learner can select the most probable grammar, given a set of data, by asking for each grammar under consideration: If the real grammar is G_n how likely is the set of data that I have observed so far? Thus, like associationist approaches, rational inference assumes that learners keep track of statistical patterns in their input. However, once an appropriate hypothesis space is specified, the amount of input data needed to converge on a probable grammar is considerably smaller than in associationist accounts (e.g. Ng & Jordan, 2002).

The goal of this chapter is to ask how well rational statistical inference can explain a set of language abilities for which we have some knowledge of the developmental time course and of the nature of the input that is required for development. The set we have selected appears in Table 4.1. Our plan is to describe what is known about each ability, analyze how well it is explained by rational inference, and where relevant, discuss what triggering and associationist accounts have to say about the ability. To foreshadow, our analysis will reveal that rational statistical inference performs well for most developing language abilities. However, fast rational inference performs less well for abilities that entail knowledge of the statistical distribution of forms at various levels, with the distribution being largely governed by diachronic forces on the language (e.g. which syllable onsets are most frequent, which stress patterns are most frequent, etc.).

We will conclude that what is needed to account for a full theory of language development is a model that involves both rational inference and associationist elements, as well as two important kinds of interaction between them. On the one hand, we suggest that the rational system might use the representations generated by the associative system to structure and constrain its space of hypotheses. At the same time, the data encountered by the rational system may be judged as unlikely under *any* hypothesis currently under consideration, which could serve as a signal that new hypotheses, perhaps depending on new representations, are needed. This "surprise

Table 4.1 Types of linguistic abilities reviewed, speed of acquisition (see text), and sample references

What is learned?	Fast or slow?	Sample studies
Which phonetic features are distinctive in the native language	Slow	Werker and Tees (1984); Polka and Werker (1994)
Typical sound patterns of native language words	Slow	Jusczyk et al. (1993); Jusczyk et al. (1994)
Phonological rules	Fast	Gerken and Bollt (2008); Cristià et al. (2011); Chambers et al. (2003)
Ordering of adjacent words	Fast	Gervain et al. (2008; Gómez and Gerken (1999); Marcus et al. (1999)
Ordering of nonadjacent words	Slow	Gómez and Maye (2005); Santelmann and Jusczyk (1998)
Word meanings	Fast	Carey and Bartlett (1978); Xu and Tenenbaum (2007a); Medina et al. (2011)
Likely referent properties involved in word meaning	Slow	Smith et al. (2002)

signal" could serve to induce greater activity in the associative system, leading it to more readily form new connections and representations. This interplay between surprise and the search for new explanations with the potential to reduce surprise was discussed by the philosopher Charles Sanders Peirce (1935). Conversely, statistical patterns which are *explained away* by hypotheses currently entertained by the rational system, patterns which might otherwise spur new associations, can be safely ignored by the associative system, as they have little to offer in the way of new statistical information. The process of statistical explaining away is an important feature of rational inference (Dawson, 2011; Pearl, 1988).

1.1. A Few Words about the Set of Language Abilities We Have Chosen

All the language abilities that we have chosen have been documented in experiments with infants and young children to the age of approximately 4 years, with an emphasis on the earlier ages in the range. We have chosen these earlier developing abilities for two reasons. First we characterize

linguistic skills by how much time it takes to acquire them. Therefore, we are most interested in abilities for which there is reasonable agreement about the time course of learning, either because the studies involve learning in the laboratory or because infants of different ages reliably show different abilities with their native language.

Second, also because of our interest in establishing time course, we have chosen abilities that have been explored using experimental techniques in which learners do not need to follow instructions any more elaborate than "show me the X" (where X is an actual or nonce word) or "what is this?" (where "this" can be given a single word label). Most of the experiments do not entail giving learners any instructions at all but rather depend on behavioral measures of interest (mostly looking). Some of the experiments focus on knowledge of the learners' native language as measured by behavioral techniques in the laboratory. Others entail exposing infants to novel words or linguistic structures and testing what they were able to learn about these stimuli in a brief laboratory visit.

The abilities represent a range of linguistic components, including phonetics, phonology, syntax involving word order, and lexical semantics. Most obviously missing are studies in which children are asked to interpret or produce more complex syntax. The reason for this gap is largely that, in our view, this is an area where there is considerable disagreement about when children demonstrate knowledge of linguistic structure (e.g. Fisher, 2002; Tomasello, 2000; Tomasello & Abbot-Smith, 2002).

Finally, let us comment on the division of learning speed into the obviously too gross measure of "fast" versus "slow." We chose these categories to see if any pattern emerged if we used them. We will attempt to provide a somewhat more nuanced discussion of learning speed under each ability under consideration in turn. We have applied these labels using the following (admittedly rough) criteria: If a linguistic ability can be shown to be acquired in a laboratory visit, and there is evidence that learners of different ages perform similarly, we assign the label "fast." In contrast, if the ability is differentially present in learners of different ages, we conclude that there is a longer time course required for learning, and we assign the label "slow."

2. RATIONAL VERSUS ASSOCIATIVE INFERENCE

Before turning to the developing linguistic abilities shown in Table 4.1, let us provide some background on rational and associative learning

models and how they might interact. We use the term "rational statistical inference" to describe model-based probabilistic inference, wherein each member of a (possibly infinite) set of hypotheses about the structure of language specifies how likely any particular pattern of data should be. Linguistic input is used to determine how likely each hypothesis is a posteriori (we have Bayesian inference in mind here, though this is not the only possible form of model-based probabilistic inference). In this way, rational inference combines top-down and bottom-up information. In contrast, an associative learning mechanism does not rely on structured representations and instead tracks a wide variety of statistics, possibly allowing new structure to emerge, which can then be leveraged in rational learning.

We suggest that, once a sufficiently constrained set of hypotheses is formed, conclusions can be drawn rather quickly, without necessarily requiring huge quantities of data. On the other hand, in less structured, associative learning, associations and statistical trends may be present within and between a wide variety of environmental sources. We discuss some potential examples of each type of learning in the next section; first, we will discuss some key features of the manifestations of rational and associative learning that currently enjoy dominance in cognitive science: Bayesian inference and connectionism, respectively.

2.1. The Fruits of Knowledge and Vice Versa

Consider a simple nonlinguistic example. Suppose you are stranded on an uninhabited island, and you are looking for some tasty fruit to eat. After some wandering, you come across a tree with some bright orange fruit. You pick one and take a bite. It is sweet and juicy. What do you expect of the next bite? It is possible part of the fruit is rotten, and the next bite will taste terrible. With only one data point so far, you do not have much raw statistical evidence to make generalizations. What will happen if the next bite is delicious as well? Probably you will be more confident that the third bite will be delicious than you were prior to taking the second bite.

Suppose you have finished your piece of fruit, but you still feel hungry. Consider three options: (1) you could reach for another piece from the same tree, (2) you could take a piece from a tree a few yards away with similar-looking fruit, or (3) you could reach for the tree immediately next to the original tree that bears some deep purple berries. Which options are most likely to reproduce your previous delicious experiences? Likely your intuition is that the chances of deliciousness are greatest in (1) and lowest in (3).

Why is (1) better than (2)? An obvious answer is that the new experience would share more features with the old ones. But why is (2) better than (3)? After all, the purple berry tree is physically closer to the original tree, so if deliciousness is related to location (perhaps the soil is especially nutrient rich at that spot), you might expect that eating a purple berry from the nearby tree would be better than eating an orange fruit from a tree farther away.

Most of us would be fairly confident that the second bite from the original piece of fruit will taste like the first, even though we can entertain the possibility that only part of the fruit could be rotten. Similarly, almost no one will doubt that the orange fruit from the far away tree is a better bet than the purple berry from the nearby one, provided the former appears sufficiently similar to be judged a member of the same type as the one already eaten. As sophisticated, worldly intellectuals, we have biological knowledge that tells us that there is usually little variability among parts of an individual piece of fruit and that taste usually depends more on the type of fruit than the location of the tree it came from. When we bring the full force of this knowledge to bear, we can generalize confidently with very little data.

Imagine you did not have that fancy university education and thus were completely ignorant about the ontology of fruit and fruit trees. Now the second bite (as well as the third) would be more of an adventure. Later, while you might still prefer option (1) to option (2), you would have a more difficult time choosing between (2) and (3). You would need to gather more data. If you had tried both and found that, indeed, the other orange fruit was delicious, but that the purple berry was sour, then perhaps you would begin to believe that appearance matters more than location. Even more so if you tried another purple berry near the second orange-fruit tree, and it was also sour. You still only have two data points from each type, and two from each location, but if you come to the problem predisposed to attend to appearance and location (as opposed to, say, whether it was 4:03 PM vs. 4:11 PM), you do not need much data to begin to feel at least somewhat confident that the former is an important predictor of flavor and the latter much less so. You may even make an even more sophisticated leap and conclude that various pieces of the orange fruit might share properties in addition to tastiness as do various instances of the purple berries.

The learner who begins with a predisposition (whether from some innate bias or from other previous experience) to treat appearance and location as potentially informative, might entertain some vague notion that fruit is divided into categories, with a vague prior distribution on tastiness,

appearance, and location given category. The means and variances for each dimension, along with the correlations between them, could then be inferred from data. Data of the sort described above, in which two pieces of tasty orange fruit and two sour purple berries were eaten, one of each from each of two locations, would be likely if the orange fruit came from one category and the purple fruit from another and unlikely if the fruit were grouped by location. Moreover, the a posteriori correlation would be relatively high between tastiness and appearance but low between tastiness and location. In contrast, a classical connectionist network which incrementally updates its weights could learn very little from four data points. The structure of the rational learner's representation constrains the learning problem enough that (what turns out to be) the correct hypothesis (that the orange fruit belong together) is already considerably better supported than the alternatives.

2.2. Types and Tokens

A key aspect of the structure possessed by the rational model which differentiates it from the associative learner is the partitioning of variability into multiple hierarchical levels. Consider what the two systems would learn as they continued to gather data from that first piece of orange fruit. After one bite, neither system is very confident about what to expect on subsequent tastes. After the second bite, the rational system gets a big boost to its confidence as it now has evidence of low variability in tastiness among bites from the same piece of fruit. The associative system gets a boost as well, but it is small. Over the next several bites, the rational system confirms its impression that intra-fruit variability is low, but since it already expected this, the returns diminish quickly. Moreover, since it separates intra- and inter-fruit variability, it learns almost nothing beyond the first few bites that helps it predict what the next piece of fruit will taste like as the relevant measure of evidence for inferences about inter-fruit variability within a type is the number of distinct fruit tokens of that type observed and not the total number of observations. If the membership of the particular piece of fruit to a type is in question initially, then the effective number of tokens observed may be less than one. As such, additional observations can provide information about inter-token/intratype variability by increasing that number toward one; however, as membership becomes near-certain, no more information which is relevant for generalization can be gained from that piece.

Contrast this behavior with that of the connectionist learner. As this learner continues to take bites from that first orange fruit, it gets more and more confident that not only this piece of fruit but also other things like it (whether the similarity is in appearance, location, or any of a variety of other features) will taste good. Without an ontology to carve its experience into types and tokens, it will have an increasing tendency to predict that orange objects taste good. Its estimate of the correlation between tastiness and orangeness, as well as of the correlation between tastiness and location, keeps rising, as it keeps receiving evidence which is consistent on all three dimensions.

In the long run, as plentiful and diverse evidence is gathered, both systems will make the correct inferences, but the rational system learns a lot early (provided it represents the problem in a useful way) and then requires new varieties of experience to continue learning, whereas the associative system makes less commitment to the structure of the problem and learns gradually and steadily from even repeated experience.

2.3. The Bias–Variance Trade-off

The trade-off between representational commitment and learning speed is encountered in statistics and machine learning problems under the name of the "bias–variance trade-off." In a formal statistical problem, one looks for an appropriate *estimator* of some latent quantity. Naturally, with finite data, perfect estimation is impossible, and so every estimator comes with some degree of error. Error arises from two sources. The *bias* of an estimator is the extent to which it deviates *on average* from the true quantity (where the average is taken over the true distribution of the data). The *variance* of the estimator is the extent to which its value is sensitive to the particular data encountered. When the variance is large but the bias is small, the error associated with any given set of input tends to be large, but because errors occur in different directions, the average value is close to truth.

As the amount of input increases, the variance of an estimator decreases. Estimators for which the variance is low at a given sample size are called *efficient*. Estimators in another desirable class (called *consistent* estimators) may contain bias for any given amount of input, but the bias vanishes in the limit of infinite data.

In the context of the present discussion, rational and associative learners have opposing advantages: associative learning is *consistent*, but rational learning is *efficient*. Rational learning is consistent as well when it is able to

entertain the correct structure, though even here it may be biased in the short term (the short-term bias here comes from quantitative, as opposed to structural, prior information).

2.4. A Rational–Associative Synergy

We envision a learning system which employs both associative and rational components in interaction. The associative component mines statistical relationships from a wide variety of sources, slowly winnowing the number of interdomain connections that it considers, as many do not produce any stable associations. As subspaces become sufficiently "modular," rational learning proceeds to construct and test manageable sets of hypotheses. In the other direction, as certain high-level hypotheses are sufficiently well supported by rational inference, the predictions they make serve to constrain associative learning, at the lower levels, *explaining away* some statistical patterns, thereby rendering them relatively uninformative in subsequent associative learning. Conversely, patterns that are particularly poorly predicted by existing hypotheses are ripe targets for additional data mining by the associative system.

3. A SELECTIVE REVIEW OF EARLY LANGUAGE ABILITIES AND THEIR SPEED OF ACQUISITION

In this section, we review the early language abilities shown in Table 4.1, above. As in the table, we characterize each ability as having been acquired quickly or slowly. We suggest that, in general, abilities that are acquired slowly reflect the gradual accumulation of data by the associative system. In contrast, abilities that can be acquired quickly and that generally do not show a difference in the age of acquisition reflect the rational system.

3.1. Learning Which Phonetic Features Are Distinctive in the Native Language

A well-documented phenomenon in language development is that infants begin their lives with the ability to discriminate most of the sound contrasts used in the world's languages but lose this ability some time during the first year of life (e.g. Polka & Werker, 1994; Werker & Tees, 1984). For example, while nearly all the 6- to 8-month olds and about half of the 8- to 10-month olds tested by Werker and Tees (1984) could discriminate two nonnative consonant contrasts, only about 20% of the 10- to 12-month olds

could do so. One possible mechanism that has been suggested to explain infants' growing focus on native speech sounds and their decreasing focus on nonnative sounds requires learners to track the distribution of phonetic features in their input. Features that occur in a bimodal distribution (e.g. voice onset time in English) are treated as phonemic (distinctive for marking meaning differences in words), while features that occur in a unimodal distribution (e.g. aspiration in English) are treated as allophonic variants of a single phoneme (Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002). One might imagine that tracking the distributions for dozens of phonetic features (and, indeed, determining which phonetic features to process more deeply, perhaps using a rational model) might take several months, thereby explaining the developmental time course of this aspect of language development. This conjecture is further supported by the observation that there are fewer phonetic features involved in distinguishing vowels than consonants and that infants lose their ability to discriminate nonnative vowels sooner than nonnative consonants.

Although collecting enough input data to determine whether a particular phonetic feature is unimodally or bimodally distributed requires time, the inference from a stable bimodal distribution to two distinct sound categories appears to be a relatively fast process. In laboratory studies examining this process, infants are presented for a brief time with nonce words in which a single phonetic feature is manipulated to create either a unimodal or a bimodal distribution. Infants who are presented with a bimodal distribution are more likely to discriminate new word tokens that vary on the critical feature than infants who are presented with a unimodal distribution (Maye et al., 2002, 2008). By isolating for infants the relevant phonetic feature while keeping other features constant, these studies allow infants to rapidly change the way in which they perceive the feature in question.

It appears that what takes developmental time in the studies of Werker and others is accumulating enough data from the multidimensional acoustic space to identify dimensions on which stable clusters emerge. In a hypothesis space which is constrained only by basic innate biases (not least the limits of perceptual hardware and the physical connectivity of the sensory system), any perceivable dimension may be related to any other, provided only that the neural representations have the capacity to communicate. Hence, the probability of spurious clusters which are the products of mere coincidence is high, and the presence of any given correlation is insufficient for the rational learner to posit with confidence that there is any "there" there.

However, as the associative system gradually alters the learner's representations, reducing the number of dimensions under consideration and moving from low-level "primitive" dimensions to more abstract "functional" dimensions[1], a more constrained rational learner can find meaningful structure.

3.2. Learning the Typical Sound Properties of Native Language Words

Another aspect of language that appears to take several months to develop is the sensitivity to frequent sound properties of native language words. Two of these properties are typical stress patterns and typical phoneme sequences, which we will refer to as phonotactic patterns. With respect to typical stress patterns, the ground-breaking work of Peter Jusczyk demonstrated that while English-learning 6-month olds fail to show a listening preference for the typical strong–weak stress pattern of English words over a weak–strong pattern, 9-month olds show a robust preference for the typical pattern (Jusczyk, Cutler, et al., 1993). Subsequent research demonstrated that 7.5-month olds are able to use their expectation about the frequency of strong–weak lexical stress to segment words with this pattern from running speech, while it is not until 3 months later that they are able to segment weak–strong words (Jusczyk, Houston, & Newsome, 1999).

Other studies generally support these early findings concerning typical word stress patterns in both English and other languages in which stress is important (e.g. Morgan & Saffran, 1995; Skoruppa et al., 2009). However, one study has shown that German 6-month olds (but not 4-month olds) prefer strong–weak over weak–strong consonant–vowel–consonant–vowel (CVCV) nonce words (Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009). The authors offer two explanations for these findings. First, German has proportionally fewer monosyllabic words than English, which might give German infants more experience with bisyllabic, strong–weak words. A second explanation concerns the fact that infants in the study by

[1] The idea here is that relevant structure is often defined not in terms of raw perceptual primitives but in terms of the relationships between those primitives, as well as quantities that are derived by combining primitives. This process is analogous to dimensionality reduction techniques in machine learning such as principal components analysis and factor analysis. Finding relationships and combinations that in some sense maximize the signal-to-noise ratio is likely a result of the associative system operating alongside some innate biases.

Höhle et al. (2009) were presented with the same CVCV nonce words, just with different stress patterns (e.g. /gába/ vs. /gabá/).

The latter explanation is consistent with the finding by Maye and colleagues described in the previous section, in which infants were able to rapidly discern unimodal versus bimodal feature distributions when only a single phonetic feature was allowed to vary. In parallel fashion, the infants in the studies by Höhle et al. may have been better able to recognize the more frequent stress pattern of German when segmental (consonant and vowel) variation was minimized. Again, it appears that the statistical machinery required by the associative learner is in place quite early, but what takes time in real language learning is applying that machinery to a very large dimensional space that needs to be winnowed down to the relevant dimensions. During the winnowing process, the learner's ability to access the relevant dimensions is not very robust; however, access can be improved if the dimension space is reduced by the experimenter.

Turning to infants' learning of typical phonotactic patterns of the words in their language, early work by Jusczyk and colleagues demonstrated that, like for typical word stress patterns, 9-month-old English learners, but not their 6-month-old counterparts, prefer lists of nonce words that exhibit more frequent phonotactic patterns over less frequent patterns (Jusczyk, Luce, & Charles-Luce, 1994). Furthermore, as in the case of typical stress patterns, 9-month olds can use typical phonotactic patterns to segment words from fluent speech (Mattys, Jusczyk, Luce, & Morgan, 1999). And as in the case of typical stress patterns, the data from English learners is corroborated by studies of children learning other languages (e.g. Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Sebastián-Gallés & Bosch, 2002).

In addition, the work on phonotactic pattern learning further supports the view that accumulating data on what is statistically typical of one's language is a slow process that can be used robustly throughout development. For example, in one study (Archer & Curtin, 2011), both 6- and 9-month-old infants discriminated legal onset clusters (probability in English > 0, e.g. /bl/) from illegal clusters (probability in English $= 0$, e.g. /dl/). However, only the 9-month olds discriminated onset clusters according to their type frequency. For example, clusters such as /pr/, which occurs as the onset of many English words, were distinguished from clusters such as /bl/, which does not begin many English words. Interestingly, neither the 6- nor 9-month olds discriminated onset clusters based on token frequency (i.e. the overall frequency in English without regard to how many words the cluster occurred in).

The ability to distinguish these different sources of variability (among words vs. among tokens of a single word) is a defining characteristic of model-based probabilistic inference. The fact that even 9-month olds appear to keep track of how often a phonotactic pattern occurs in particular words is evidence that the ability to track types versus tokens is one that is present very early in life. Whereas the type-token distinction is characteristic of a rational inference system in general, employing this distinction in the course of the slow accumulation of input statistics may reflect the influence of the rational system on the associative system.

3.3. Learning Phonological Rules

In contrast to the apparently slow accumulation of data regarding the typical word stress and phonotactic patterns of the native language, learning rule-like generalizations about stress and phonotactics appears to occur very rapidly in the laboratory. Beginning with stress pattern learning, Gerken (2004) exposed 9-month olds to three- to five-syllable words in which the pattern of strong and weak syllables was governed by a set of ranked (optimality theory) principles. At test, infants were able to distinguish new words with new stress patterns that confirmed to the previously encountered principles from those that did not. One of the principles for stress assignment in the Gerken's (2004) study was that syllables ending in a consonant should be stressed. Gerken and Bollt (2008) demonstrated that 9-month olds could learn that principle if they encountered three different syllable types ending in a consonant, but not if they encountered multiple tokens of only a single type. This finding is consistent with a growing body of evidence that infants and young children learn to generalize over linguistic types and not tokens, an important component of rational, but not associative, accounts of language development (Archer & Curtin, 2011; Xu & Tenenbaum, 2007b).

One finding from the study by Gerken and Bollt (2008) described above illustrates how the slow accumulation of data about what is typical in the native language interacts with the faster generalization based on rule-like structure that is a hallmark of rational inference. In one experiment, Gerken and Bollt presented 7- and 9-month olds with words whose stress patterns reflected a principle that does not occur in human language: "stress syllables that begin in /t/." The younger infants learned this principle, distinguishing a new stress pattern in which the principle interacted predictably with other ranked principles from one in which it did not. However, 9-month olds, who were able to learn the principle that syllables ending in a consonant are

stressed, were not able to learn the unnatural rule that syllables starting with /t/ are stressed. A likely explanation for this developmental change is that although both groups of infants were able to make the types of rational inference required for rule learning, the older infants did not view syllable onsets as having a likely effect of word stress.

Why might this be? One possibility is related to the relation of syllable content and stress in English. At first glance, a learner might perceive a correlation between syllables starting with /t/ and stressed syllables since /t/ is the sixth most frequent onset of stressed syllables in one- and two-syllable words. In other words, based on the sheer frequency of occurrence and co-occurrence, a plausible generalization is that syllables starting in /t/ are stressed. However, a learner who was able to accumulate additional statistics of what is typical of English would find that /t/ is no more likely to be an onset of stressed than of unstressed syllables. Put another way, the frequent co-occurrence of /t/ onsets and stress can be *explained away* in English once the statistics of stressed and unstressed syllables are known. However, discovering this fact would require knowing enough words that start with an unstressed syllable to detect that proportionally no fewer of these start in /t/ than of words starting with a stressed syllable.

We have already noted that English-learning infants at 7 months have difficulty segmenting words with a weak–strong stress pattern from the speech stream, perhaps, because they have focused their word-form-learning efforts on the most frequent word forms in the language (Jusczyk et al., 1999). Quite possibly 7-month olds would not have sufficient data accumulated about the onsets of weak–strong words to view onsets as unlikely to affect word stress. In contrast, 9-month olds may have begun to accumulate sufficient data to weight syllable endings as more likely to affect stress assignment than syllable onsets. This explanation of the difference in learning between 7- and 9-month olds requires the accumulation of data about the language input over developmental time.

Although a greater knowledge of the statistics of English reveals that a relation of stress and syllable onsets is spurious, a relation between stress and syllable codas should continue to be viable with more data. Not only are final consonants very frequent on stressed syllables, conditional probabilities (Prob (coda|stress)) also support the relation in English. Therefore, there is no basis for 9-month olds to explain away the principle that syllables ending in codas are stressed in an artificial language, even though that principle is not absolutely upheld in English. In short, the data suggest that the developmental change seen in infants' ability to learn a principle about word stress

assignment involves an interaction of fast rational inference and slower accumulation about the statistics of English.

Turning to infants' ability to learn about phonotactics quickly in the lab, several of studies provide parallel results to those discussed for stress patterns. Chambers, Onishi, and Fisher (2003) familiarized 16.5-month-old infants with CVC syllables in which particular consonants were artificially restricted to either initial or final position (e.g. /bæp/ not /pæb/). During test, infants listened significantly longer to new syllables that violated the familiarized positional constraints than to new syllables that obeyed them. In this study, infants could have responded based on familiar segment-by-syllable position correlations (e.g. /b/ first, /p/ last).

A similar study by Saffran and Thiessen (2003) suggests that infants are rapidly able to consider patterns that embody more abstract featural relations. They familiarized 9-month olds with words with a consistent word-shape template. For example, in one condition of their second experiment, infants were familiarized with CVCCVC words which had the pattern +V, −V, +V, −V (in which +V = voiced and −V = voiceless) on the four consonants (e.g. /gutbap/). Infants were then tested to determine if they were able to segment from fluent speech new words that fit versus did not fit the familiarized pattern. The familiarization and test words were designed so that no particular sequence of consonants occurred in both familiarization and test (e.g. g_tb_p occurred in familiarization but not in test and g_kb_p occurred in test but not in familiarization). Therefore, the influence of the familiarization phase on infants' preference during test was presumably due to word templates specified in terms of features, not specific phonemes.

In an interesting parallel to the work of Gerken and Bollt (2008), Cristià and colleagues (Cristià & Seidl, 2008; Cristià, Seidl, & Gerken, 2011) tested both 7- and 4-month olds' ability to learn phonotactic patterns that involve natural and unnatural sound classes. Infants were exposed to CVC nonce words in which the onset position was either filled by stops and nasals (which form the natural sound class of minus-consintuant) or the unnatural class of stops and fricatives. During test, infants were exposed to new words with different onsets that were either consistent or inconsistent with the grouping the infant was familiarized with (stops and nasals or stops and fricatives). While 4-month olds showed evidence of learning both natural and unnatural groupings, 7-month olds only learned the natural groupings. In keeping with the discussion of developmental change in infants' willingness to entertain natural and unnatural stress assignment principles, we suggest that the slowly accumulating statistics of English phonotactics is responsible

for the 7-month olds studied by Cristià et al. (2011) rejecting the grouping of stops and fricatives as a possible generalization. One possible statistical pattern of English that might be responsible is that both stops and nasals can occur after /s/, while most fricatives do not (for further discussion, see Cristià & Seidl, 2008). However, not all stops can occur after /s/, and glides and liquids can also occur after /s/. Because glides and liquids are not part of the same putative natural class as stops and nasals, further research is needed to determine if the same developmental pattern seen for stops and nasals applies to these other sounds as well.

In summary, both stress principles and phonotactic restrictions can be learned rapidly in the laboratory by infants as young as 4 months. However, the rapid learning we see for such generalizations appears to be influenced by the slow accumulation of statistics about typical stress patterns and typical phonotactic patterns of the infant's native language.

3.4. Learning the Order of Adjacent Words

A number of studies have demonstrated that infants know about the word order or the general word-order properties of their native language. For example, Shady, Gerken, and Jusczyk (1995) presented 10.5-month olds with normal English sentences as well as sentences in which determiners and nouns were reversed, resulting in phrases like *kitten the*. The stimuli were recorded using a speech synthesizer to avoid disruptions in prosody that are likely to occur when a human talker produces ungrammatical sentences. Infants listened longer to the unmodified sentences, suggesting that they were able to tell the difference between the two types of stimuli. More recently, a group of researchers asked whether Italian and Japanese 8-month olds differently parsed a string of nonce syllables with an AXBY format as beginning or ending with more frequently produced A/B elements (Gervain, Nespor, Mazuka, Horie, & Mehler, 2008). Japanese is a language in which the most frequently occurring words (functors) occur sentence-finally, whereas the comparable elements in Italian occur sentence-initially. Consistent with the abstract word-order properties of their language, Japanese-learning infants listened longer to word strings that ended in frequent A and B syllables, whereas Italian-learning infants showed the opposite preference.

Other studies demonstrate that infants as young as 4 months can learn the order of word-like units in short syllable strings (Dawson & Gerken, 2012; Gómez & Gerken, 1999), as well as learning the more abstract patterns of

repeated or alternating syllables (Gerken, 2006, 2010; Gómez & Gerken, 1999; Marcus, Vijayan, Rao, & Vishton, 1999). For example, several studies have shown that 7- and 9-month olds can learn an AAB pattern (first two syllables are the same) or an ABA pattern (first and third syllables are the same) easily with minimal input (Gerken, 2006, 2010; Marcus et al., 1999). Dawson and Gerken (in preparation) found that even 4-month olds were able to learn such a pattern. Interestingly, although 7- and 9-month olds can learn the AAB versus ABA pattern instantiated in syllables, they cannot learn the same patterns instantiated in musical notes or chords. In contrast, 4-month olds can learn the pattern in both media (Dawson & Gerken, 2009). Dawson and Gerken explain this developmental difference by noting that repeated notes are very frequent and therefore highly predictable, once you know the structure of Western tonal music. Research suggests that only older infants know about this structure (e.g. Saffran, 2003), and therefore, only they can explain away musical repetition as the result of general properties of musical structure and not as a local "grammatical" feature. In contrast, repetition of words in English is very rare and requires a separate explanation at all the ages tested.

All the studies cited in this section suggest that learning the order of particular words in a string, as well as more abstract patterns of frequent or repeating words, occurs quickly and shows no consistent developmental change (i.e. the long-term changes that have been observed appear as both gains and losses in capacity, presumably reflecting changes in broader knowledge, and not the gradual acquisition of the specific linguistic skills being tested).

3.5. Learning the Order of Nonadjacent Words

Often in natural language, the presence of a particular word or morpheme is dependent not on the word immediately preceding, but to preceding nonadjacent word. For example, in the sentence "Granny is buttering your toast," the inflection "-ing" depends not on "butter" but on "is." Santelmann and Jusczyk (1998) found that 18-month olds, but not 15-month olds listened longer to sentences like "Granny is buttering your toast" than ungrammatical versions like "Granny can buttering your toast." Taken alone, this result might either suggest that younger infants either had not accumulated enough input data to reliably learn longer distance dependencies or that they do not have the computational inclination or ability to consider dependencies between nonadjacent elements.

The latter explanation is supported by work in which a similar developmental effect for nonadjacent dependencies was observed for learning of an artificial grammar in the laboratory (Gómez, 2002; Gómez & Maye, 2005). In these studies, infants of different ages were exposed during a 2-min familiarization period to three-element strings (e.g. *pel-kicey-jic*) in which the third word depended on the first word. The middle word was not relevant to word order, and there were 3, 12, or 24 middle words, depending on the condition in which the infant participated. Across several studies, infants were only able to learn the dependency between the first and third word when the set size of the middle element was 24. Gómez (2002) argued that it is only when the set size of the middle element is large enough (as it is in natural language) to force infants to abandon their preferred pattern-finding strategy of looking for correlations between adjacent elements.

Interestingly, 17- and 18-month olds indicated that they learned the dependency by demonstrating a novelty preference at test, that is, listening longer to strings that violated the pattern that they had heard during the preceding familiarization period. In contrast, 15-month olds demonstrated that they learned the dependency but demonstrated a familiarity preference, which Gómez and Maye (2005) take to indicate that they had learned the dependency less well than the older infants. In contrast, 12-month olds failed to learn the dependency at all. The set of findings described in this section suggests that younger infants are unlikely to even look for dependencies among nonadjacent elements, while older infants (and adults) will look for such dependencies, provided their normal strategy of looking for adjacent relations is made sufficiently difficult.

One possible explanation for the developmental change observed in these studies is that infants are developing a representation of the grammar of their language using the rational inference system. This grammar can include dependencies among elements contained within a syntactic constituent. The associative system then accumulates data about dependencies in the learner's native language, and in English, the data demonstrate that "is" but not "can" predicts "-ing." Although this proposal is clearly speculative at this point, it suggests a way in which the rational and associative systems might interact over the course of development.

3.6. Learning Word Meanings

A well-documented phenomenon in early childhood is children's ability to learn the meaning of a word in a single exposure and to remember the word

over time. This ability, often termed *fast mapping* was reported by Carey and Barlett (1978) and has been observed by numerous researchers since (e.g. Medina, Snedeker, Trueswell, & Gleitman, 2011). Recent research by Xu and Tenenbaum (2007a, 2007b) has explored fast mapping from a rational statistical inference perspective (Bayesian modeling). In particular, they examined the course of learning when a label was applied to more than a single referent. Xu and Tenenbaum (2007b) showed 3- to 4-year-olds either a single Dalmatian or three different Dalmatians and labeled each example *fep*. They then asked children to give them another *fep* from a set of toys that included Dalmatians, non-Dalmatian dogs, and other animals. Children always treated a Dalmatian as the most likely extension of *fep*. That is, even in when presented with a fast-mapping, one-referent one-label, situation, children behaved as expected. However, when the label was applied to three different Dalmatians, children (and adults) were less likely to select a dog that was not a Dalmatian than when the label was applied to a single Dalmatian. That is, word learners seem to increase their confidence in the appropriate label-referent pairing, but they achieve near-perfect performance very quickly. Importantly, Xu and Tenenbaum (2007b) compared a Bayesian model to an associative (Hebbian) learning model, which did not distinguish between types (different Dalmatians) and tokens (the same Dalmatian seen three times). The Bayesian model better matched the behavioral data.

Despite the general agreement that children are able to learn word-referent mappings relatively quickly, there is some debate about just how much exposure is needed. Xu and Tenenbaum (2007b) found that children picked the subordinate category (e.g. Dalmatian) significantly more when given three input types than when given a single input type. However, a study employing more naturalistic scenes and asking adults and preschoolers to guess the meaning of a word uttered in that scene suggested that if a particular scene was informative, no additional scenes in which the same word was used improved participants' performance (Medina et al., 2011). Medina and colleagues suggest that their results support a view in which a single hypothesis is entertained about the meaning of a word, although the hypothesis might be rejected wholesale if it is subsequently disconfirmed. A number of features differ between the study by Medina et al. (2011) and other studies, including the complexity of the scenes and importantly, whether the speaker intended to teach the participant a word (Xu and Tenenbaum—yes, Medina et al.—no), and whether a set of alternative referents was provided at test (Xu and Tenenbaum—yes, Medina

et al.—no). Although providing alternative referents may be less reflective of word learning "in the wild," expecting very young learners to hear a word form and guess its meaning in a free field may also be unusual. Therefore, until additional evidence comes to light that word learning is not a form of hypothesis testing, we will view this domain as generally consistent with rational inference.

3.7. Learning Likely Referent Properties Involved in Word Meaning

As noted, the findings in the previous section suggest that learning the meanings of words can occur quite quickly, which we take to be generally consistent with rational statistical inference in the form of Bayesian models (though see Yu & Smith, 2012). However, it is important to note that the children in the study by Xu and Tenenbaum (2007b) were relatively experienced word learners. Other work with younger learners suggests that determining which features of word referents are likely to be important in assigning word meaning is a slower process (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

If learning the appropriate semantic extension of category labels is the first level of word learning, then learning to prioritize some features over others when extending category labels to novel exemplars can be thought as a form of second-order learning as it requires the child to abstract across multiple object categories and extract similarities in their featural organization. Some authors argue for an associative approach to learning at this level, suggesting that learners must first master several first-order cases before moving up the abstraction ladder to form the higher order generalization. Samuelson (2002) as well as Colunga and Smith (2005) propose connectionist models of the "shape bias" documented by Smith et al. (2002), which leverage input statistics (e.g. that labels for solid objects tend to be preceded by "a" and "the" and appear in both singular and plural forms, whereas labels for materials have only one form and can appear without a determiner or with the determiner "some") to arrive at a taxonomy in which solid object categories and substance categories occupy different regions in semantic space, and hence, labels of the former should generalize along a shape dimension but not along a color dimension, whereas the reverse is true for labels of materials.

Due to the wide variety of potentially relevant input statistics, and due to its tendency to build abstractions from the bottom-up, the associative learner

requires a lot of input to acquire second-order generalizations like the (selective) shape bias. Along the way, they overextend the shape bias beyond the appropriate ontological kind, reflecting the empirical behavior of children.

There have been attempts to account for higher order learning of this sort using rational probabilistic models as well. Kemp, Perfors, and Tenenbaum (2007) frame knowledge about which feature dimensions to use in generalization as arising from the learner's representation of variability within categories along each such feature. Low variability for a particular feature reflects high consistency, and hence, novel exemplars are more likely to share this feature with those previously experienced. Kemp et al. present a hierarchical Bayesian model (HBM) which begins with the assumption that objects are divided into kinds (but does not know how many there are) and that kinds are divided into categories and learns from experience with labeled objects that solid and nonsolid categories are organized along different features.

An interesting feature about the HBM approach taken by Kemp et al. (2007) is that, under some conditions, lower order generalizations are learned before higher order ones, and under others, learning occurs in the opposite order. In the case of the shape bias, depending on the statistical distributions in the input, it is possible to learn the general tendency for labeled object categories to be organized by shape with very little data from any particular category. This is a result of the representational distinction made between types and tokens: If the learner encounters two tokens from each of several different types, and within types, the pairs always have the same shape, then the model will be very confident that categories are shape homogeneous and can confidently predict the properties of a new category from a single instance.

The case of reference may lie at the intersection of the associative and rational systems. A rational learner like the one exhibited by Kemp et al. (2007) is able to learn at multiple levels of abstraction simultaneously, provided it is looking for the right kind of ontology, namely one in which a certain class of linguistic constituent (somewhere between a noun and a noun phrase) is assumed to refer to an object and where nouns are organized into broad classes, each of which has different semantic organizing principles. With relatively few properties to focus on, learning proceeds quickly, and the model discovers the distinction between categories which are organized by shape and categories which are organized by material.

Before such a rational learner can proceed, however, the child would need sufficient statistical evidence that there is more than one type of noun

to begin with. In the model by Kemp et al., the fact that a different set of variability parameters should be inferred for each of a number of ontological kinds was given at the start. That is, the model's representation is structured in such a way that the color distribution of a particular object category is taken to be informative about only one ontological kind, even though it is unknown which one. This is analogous to the type–token representational distinction, at a higher level in the hierarchy. Unlike the Bayesian model, the children in Smith et al. (2002) overgeneralize their shape bias to mass-noun categories, suggesting that they do not yet have this clean representational distinction. It seems plausible that some slow, associative data mining is needed to reach the point where nouns can come in distinct ontological kinds, after which point a rational learner can take over.

4. DISCUSSION

In this chapter, we have reviewed empirical evidence pertaining to a variety of linguistic domains. For each domain, we have attempted to roughly classify it as "fast" if it can be learned in a short laboratory visit by learners of different ages or "slow" if the ability is differentially present in learners of different ages.

One way to characterize the pattern of fast- and slow-developing abilities that we have described is as follows: Fast learning appears to involve either domains in which the pattern observed in the input can be described as generated by a rule or in which a word-referent pairing is established (particularly by experienced word learners). Slow learning seems to share two properties. One is that it involves domains in which the learner needs to establish detailed distributions of features in the input. Examples of this type of slow learning from Table 4.1 are learning which phonetic features are distinctive in the native language, learning the typical sound patterns of native language words, and learning the likely referent properties involved in word meaning. The other example of slow learning shown in Table 4.1 is learning the ordering of nonadjacent words. Here we argue that what might take developmental time is not only accumulating data about what words and morphemes co-occur (e.g. "is" and "-ing") but knowing to look for co-occurrences among nonadjacent elements in the first place. The combinatorial explosion involved in looking for all potential co-occurrences without restricting oneself to a bounded domain is computationally prohibitive. Therefore, it appears that learners must first appropriately represent the

syntactic constituents in their language, such as sentences and phrases, before they can make significant progress in finding meaningful nonadjacent relationships. Once they have made such a determination, restricting the search for co-occurrences within constituents can proceed.

To summarize, we have characterized language learning as involving two distinct but interacting inference systems. The first is a rational system (of the sort that occurs in Bayesian probabilistic inference) that is able to learn quickly, provided it begins with the appropriate hypothesis space. The second is an associative system (of the sort modeled by Hebbian associative networks) that learns more slowly, but also more flexibly, than the rational system. We have suggested two principal ways in which each system takes advantage of the "output" of the other: First, the associative system alters and simplifies the representations employed by the learner, allowing the rational system to test a better constrained set of hypotheses. In turn, the rational system provides a grammatical framework, including prospective units of analysis (e.g. syntactic constituents, word types instead of tokens), that guides the data accumulation of associative system and also allows what would otherwise be "suspicious" coincidences to be explained away, preventing overlearning of spurious associations.

4.1. Why Is One System Insufficient?

Occam's razor dictates that one should only propose two entities when one cannot adequately account for the data. We have outlined some strengths and weaknesses of each of the two systems and described how the weaknesses of each are compensated for by the strengths of the other in a hybrid system. But it is certainly worth considering whether a single system could reasonably account for the empirical data, even if it does not have all the advantages of a dual system. We conclude that a "pure" learner of either stripe will encounter some major difficulties when faced with the complex challenge of acquiring all the linguistic abilities that adults seem to possess. We consider these difficulties in turn for associative and rational inference.

4.1.1. Logical and Empirical Challenges for a Purely Associative Theory

A purely associative account of learning faces problems both in principle and in its ability to account for observed experimental data. The most obvious logical challenge was pointed out by Chomsky and others: without representational constraints, the number of correlations and generalizations that

are possible from any finite data set is prohibitively enormous. In the specific context of a neurally inspired model, there is a problem of combinatorics: it is physically impossible for everything to connect to everything else. Admittedly, this is a straw argument: the most die-hard connectionist purist makes some representational assumptions, and biases are built into the way the network is arranged.

The principal empirical evidence against a purely associative account, as we see it, is twofold. First, associative accounts predict slow, gradual learning, which is at odds with data from many areas of language learning such as word learning (Xu & Tenenbaum, 2007a, 2007b), phonological categorization (Maye et al., 2002), and syntactic acquisition (Gerken, 2010). Second, and perhaps most directly in support of a need for a rational component, infants represent variability at multiple levels, treating types and tokens differently (Archer & Curtin, 2011; Xu and Tenenbaum, 2007b), which an associative account would not predict.

4.1.2. Challenges for a Purely Rational Theory

The chief logical problem with a pure "hypothesis-testing" theory of language learning is determining where the hypotheses come from. Triggering theories rely on a fairly detailed innate hypothesis space, but their proponents arrive at this conclusion indirectly, by arguing that language learning is impossible, and not by direct empirical evidence. It would be more satisfying, as a scientific matter, to assume as little as possible in the way of innate knowledge and develop an account of a learner that could acquire the right kinds of biases from input.

Empirically, purely rational accounts have a difficult time predicting the time course of linguistic development. While they tend to do well at accounting for patterns of behavior within the laboratory, Bayesian models rely on a precise characterization of the input in order to make specific predictions, which is not generally available for longer time courses. This is the opposite of the problem faced by connectionist models, which rely on "asymptotic" results.

4.1.3. Solutions Offered by a Hybrid Model

We have discussed some ways in which rational and associative inference have complementary strengths and weaknesses. One area we have focused on is learning speed. Some aspects of language development, especially those studies in brief laboratory visits, appear to occur very quickly, while others appear to proceed more slowly. The hybrid model that we are proposing

arose largely as a framework for understanding these different developmental time scales.

In addition, we believe that the hybrid model offers a solution to the problem of the explosion of units and statistics that either an unconstrained associative learner or an unconstrained rational learner would face on its own. Each time that the rational inference system adds something to the grammar, the associative learner is newly constrained in terms of the units over which it keeps statistics. Although we have not dealt specifically with the different possible statistics that an associative learner might track, all of our examples rely on tracking only frequency distributions, forward conditional probabilities, adjacent dependencies, and nonadjacent dependencies within the bounds of syntactic constituents. Conversely, the associative learner may limit the set of hypotheses considered by the rational learner.

Finally, the hybrid model has the potential to harness the hypothesis-testing power of the rational system, while leveraging the "creative" power of the associative system to generate hypotheses in the first place. Peirce (1935) described "inference to the best explanation," also known as abductive inference, as follows:

(1) The surprising fact, C, is observed.
(2) But if A were true, C would be a matter of course.
(3) Hence, there is reason to suspect that A is true.

—*Charles Sanders Peirce (1935)*

Once the set of possible explanations is determined, a rational inference system can proceed in this manner, settling on the explanation that makes the data the least surprising. However, the associative system is needed to construct a pool of potential representations out of the sound and fury, some of which the rational system can explain away as truly signifying nothing.

We close this chapter by remarking that the hybrid framework we have outlined here is clearly not yet a fully formed theory of language learning. We have roughly divided linguistic capacities into two categories and attempted to fit these categories into the mold of either rational Bayesian inference or associative Hebbian learning, and we have attempted to describe ways in which these two systems might interact. It will likely be possible to expand upon our conception of either rational or associative so as to expand its territory beyond the blurry boundary lines we have drawn, but while the precise limits are flexible, we are hopeful that the conceptual distinctions we have made here will prove fruitful in future discussions of the nature of language learning.

ACKNOWLEDGMENTS

This research was supported by NICHD #R01 HD042170 and NSF 0950601 to LAG.

REFERENCES

Archer, S. L., & Curtin, S. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development, 34*(4), 534–540.

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

Chambers, K. E., Onishi, K. H., & Fisher, C. L. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87*, B69–B77.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N., & Lasnik, H. (1993). *Principles and parameters theory in syntax*. Berlin: de Gruyter.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review, 112*(2), 347–382.

Cristià, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development, 4*(3), 203–227.

Cristià, A., Seidl, A., & Gerken, L. A. (2011). Young infants learn sound patterns involving unnatural sound classes. *University of Pennsylvania Working Papers in Linguistics, 17*(1). Article 9.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language, 2*, 133–142.

Dawson, C. (2011). *"Explaining-away" effects in rule-learning: evidence for generative probabilistic inference in infants and adults*. Doctoral thesis. Tucson, AZ: The University of Arizona.

Dawson, C., & Gerken, L. A. (2009). Learning to learn differently: the emergence of domain-sensitive generalization in the second six months of life. *Cognition, 111*, 378–382.

Dawson, C., & Gerken, L. A. (2012). *Stimulus complexity affects generalization: A developmental finding*. Manuscript in preparation.

Elman, J. (2003). Development: it's about time. *Developmental Science, 6*, 430–443.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello (2000). *Cognition, 82*(3), 259–278.

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

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science: A Multidisciplinary Journal, 31*(2), 311–341.

Gerken, L. A. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition, 93*, B89–B96.

Gerken, L. A. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition, 98*, B67–B74.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115*(2), 362–366.

Gerken, L. A., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: implications for generalization mechanisms and constraints. *Language Learning and Development, 4*(3), 228–248.

Gerken, L. A., & Dawson, C. (in press). Grammar learning as model building. In T. Mintz (Ed.), *Statistical approaches to language*. The Hague: Taylor Francis.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: a Japanese Italian cross-linguistic study. *Cognitive Psychology, 57*(1), 56–74.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Gómez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*(2), 109–135.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency. *Infancy, 7*(2), 183–206.

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: evidence from German and French infants. *Infant Behavior & Development, 32*(3), 262–274.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64*(3), 675–687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory & Language, 32*(3), 402–420.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*(3–4), 159–207.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory & Language, 33*(5), 630–645.

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

Lorenz, K. (1935). Der Kumpan in der Umwelt des Vogels. *Journal of Ornithology, 83*, 137–413.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77–80.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*(4), 465–494.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science, 11*, 122–134.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.

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 of the United States of America, 108*(22), 9014–9019.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66*(4), 911–936.

Ng, A. Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* (pp. 841–848). Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Peirce, C. S. (1935). Pragmatism and abduction. In C. Hartshorne (Ed.), *Collected papers of Charles Sanders Peirce* (pp. 112–135). Cambridge, MA: Harvard University Press.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). *Poverty of the stimulus? A rational approach*. Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, British Columbia, Canada.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 421–435.

Rumelhart, D., & McClelland, J. (1987). Learning the past tenses of English verbs: implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 195–248). Mahwah, NJ: Lawrence Erlbaum Associates.

Saffran, J. R. (2003). Absolute pitch in infancy and adulthood: the role of tonal structure. *Developmental Science, 6*(1), 35–43.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484–494.

Samuelson, L. K. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20-month-olds. *Developmental Psychology, 38*(6), 1016–1037.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition, 69*(2), 105–134.

Sebastián-Gallés, N., & Bosch, L. (2002). Building phonotactic knowledge in bilinguals: role of early exposure. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 974–989.

Shady, M. E., Gerken, L. A., & Jusczyk, P. W. (1995). Some evidence of sensitivity to prosody and word order in ten-month-olds. In MacLaughlin, D, & McEwan, S. (Eds.), *Proceedings of the 19th Boston University Conference on Language Development*, Vol. 2. Sommerville, MA: Cascadilla Press.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Núria Sebastián-Gallés, N., et al. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science, 12*(6), 914–919.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*, 13–19.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*(3), 209–253.

Tomasello, M., & Abbot-Smith, K. (2002). A tale of two theories: response to Fisher. *Cognition, 83*(2), 207–214.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Xu, F. (2007). Rational statistical inference and cognitive development. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Foundations and the future* (pp. 199–215). Oxford, UK: Oxford University Press.

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

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: prior questions. *Psychological Review, 119*(1), 21–39.