Modeling syntactic acquisition

For the Oxford Handbook of Experimental Syntax

Lisa S. Pearl
University of California, Irvine
lpearl@uci.edu

February 1, 2021

1 Why model?

Just like other scientific tools of investigation, modeling is something we use to answer certain kinds of questions. For syntactic acquisition, these questions tend to concern the process of acquisition that yields adult syntactic knowledge – that is, *how* exactly syntactic acquisition proceeds, using particular learning strategies (Pearl and Sprouse, 2015). In essence, an informative model of syntactic acquisition is the embodiment of a specific theory about syntactic acquisition. So, to build an informative syntactic acquisition model, you need to first have a theory about how syntactic acquisition works. Then, the model can be used to (1) make all the components of that acquisition theory explicit, (2) evaluate whether it actually works, and (3) determine precisely what makes it work (or not work).

1.1 Making the components explicit

It might seem surprising at first to claim that one benefit of modeling is that it makes theory components explicit – but this truly *is* important. It often turns out that the acquisition theories that seem explicit to humans don't actually specify all the details necessary to implement the strategies these theories describe. For example, suppose a proposed learning strategy is that children use triggers in their input to signal certain parametric values, and the triggers are explicitly defined ahead of time for children (perhaps through Universal Grammar (UG)). As a concrete example, let's say that the trigger for *wh*-movement is seeing a *wh*-word in a position different from where it's understood (e.g., *what* in the question *What did the penguin do* $\_\__{what}$*?*). So, the learning strategy for acquiring the right *wh*-movement structure in your language involves identifying these *wh*-movement triggers.

Are we finished? Not quite. What do children need in order to recognize the appropriate *wh*-movement trigger in their input? They probably at least need to know that a certain word is one of these special *wh*-words (in English, this would include *who*, *what*, *how*, and so on). They probably need to be able to reliably segment the words in the utterance and recognize that the *wh*-word is

not appearing where it's understood (which means they need to understand enough of what the utterance means). They probably need to be able to remember the existence of the fronted *wh*-word in the utterance long enough and reliably enough so they can update their internal parameter value. They probably need to ignore the utterances in English where the *wh*-word doesn't move (e.g., echo questions like *The penguin did* ***what***?).

And then, what about the *wh*-in-situ option (for languages like Mandarin Chinese and Japanese) – is there a trigger for that as well? If so, what is it? More specifically, suppose we only have a trigger defined for *wh*-movement. In this case, we're effectively implying that the lack of that trigger indicates *wh*-in-situ. This is fine as a supposition, but how exactly does that work in practice? For instance, is *wh*-in-situ a default setting that's overridden by the *wh*-movement trigger? This isn't unreasonable as a default assumption and might result from a general tendency to disprefer movement unless the language indicates otherwise (e.g., based on something like the Minimal Chain Principle: De Vincenzi 1991, Sakas and Fodor 2012, Fodor and Sakas 2017). Or, if there are no defaults, does a child use indirect negative evidence to decide that her language uses *wh*-in-situ? In particular, she would observe that *wh*-movement keeps not happening when she expects it to. If this is the learning strategy, then how long does she have to wait before she decides that her language doesn't have *wh*-movement?

These are just some of the aspects of the triggering theory that need to be specified in order to implement it in a model. More generally, by trying to build an implementable model for a particular syntactic acquisition theory, we can see where the gaps are. This is because a computer program can only implement an acquisition theory where every relevant detail is specified (Kol, Nir and Wintner, 2014, Pearl, 2014, Pearl and Sprouse, 2015). So, even if an acquisition theory has already been developed, a computational model provides a way to flesh out the necessary components of that theory.

1.2 Evaluating the theory and explaining what happened

Once an acquisition theory is specified enough to implement in a computational model, we can then evaluate it by comparing the predictions it generates against the empirical data available from children. I should note that we have to be somewhat careful about interpreting model results. There are two basic outcomes: (1) the model predictions match children's data, or (2) they don't.

If the predictions match, this is an existence proof that the acquisition theory, as implemented by the computational model, is a way that acquisition *could* proceed. That is, the computational model demonstrates exactly how successful acquisition could work. So, this model is support for that acquisition theory. Notably, however, this modeling evidence doesn't rule out alternative acquisition theories. This is why I emphasized *could* above: Just because the model demonstrates one way acquisition could work doesn't mean that other ways couldn't also work. So, modeling evidence is interpreted with respect to the implemented acquisition theory only.

Still, of course, sometimes the model predictions don't match children's data. What then? This is then evidence against that acquisition theory, *as implemented by the model*. That is, we can't immediately rule out all versions of the acquisition theory unless we explicitly implement them (or rule them out for principled reasons). Remember: A model often specifies components of a theory that the original theory didn't. So, if this particular theory implementation doesn't work, maybe

it's a problem with those components, and not the theory more broadly. The only way to know is to test each of the possible implementations, or rule them out in principle somehow. This of course becomes hard to do in practice – there may be quite a number of implementations for any given acquisition theory. (Just think about all the implementational details for the triggering theory described in the previous section.) The other option, ruling implementations out on principle, depends greatly on the principles available and how agreed upon they are. (Are these principles of human computational capacity, theoretical economy, something else?) This practical consideration about model interpretation is why we may often seen published results involving models that succeed (even if the results also include models that fail), rather than only results about models that fail.

Related to that, if you have an implemented model (whether it succeeds or fails), a very useful benefit is that you can look inside it to determine what exactly makes it work or not work. This is something that's much more difficult to do with children's minds. That is, we can sift through the components of the implemented acquisition theory to see which ones are important for acquisition success. For example, suppose we have a successful implementation of the triggering theory for learning about *wh*-movement. We can see if it's important for English children to ignore *wh*-echo questions where there's no *wh*-movement, or how necessary a Mandarin Chinese default *wh*-in-situ value is. Suppose we find that these components are vital for filtering children's input appropriately (e.g., ignoring echo questions) or navigating the hypothesis spaced based on the input (e.g., a default *wh*-in-situ value) – that is, without them, the model's predictions don't match children's behavior. Then, we can say that these are necessary components of the successful trigger-based acquisition theory that explains how children learn where *wh*-words appear in their language. Moreover, we can also explain *why* (e.g., one component filters the input and the other helps children navigate the hypothesis space).

This highlights how modeling can be used as a tool for both developing and refining acquisition theories. Notably, an acquisition theory actually includes two types of theories: theories of the learning process and theories of the representations to be learned. An informative model requires us to be explicit about both. So, when we build a model that incorporates both theory types and see the results, we get feedback about both. To understand why an informative model incorporates theories of the learning process and theories of representation, it's helpful to consider all the pieces that go into characterizing the language acquisition task.

1.3 Characterizing the acquisition task

There have been several recent discussions of the acquisition process (Pearl and Mis, 2011, Pearl and Sprouse, 2013b, Pearl, 2014, Lidz and Gagliardi, 2015, Omaki and Lidz, 2015, Pearl and Sprouse, 2015, Pearl and Mis, 2016), and I find the model articulated by Omaki and Lidz (2015) and more fully by Lidz and Gagliardi (2015) to be especially helpful (an adapted version is shown in Figure 1). This model specifies components external and internal to the child during the acquisition process, and is meant to capture the iterative process of acquisition unfolding over time. I should note that this is one particular way of specifying different important components of the acquisition process – it very well may be that future versions alter how the specified components relate to each other or even what the specified components are. For now, however, I think it sets us

up quite well to think concretely about components of the acquisition task that are important for computational models of acquisition.

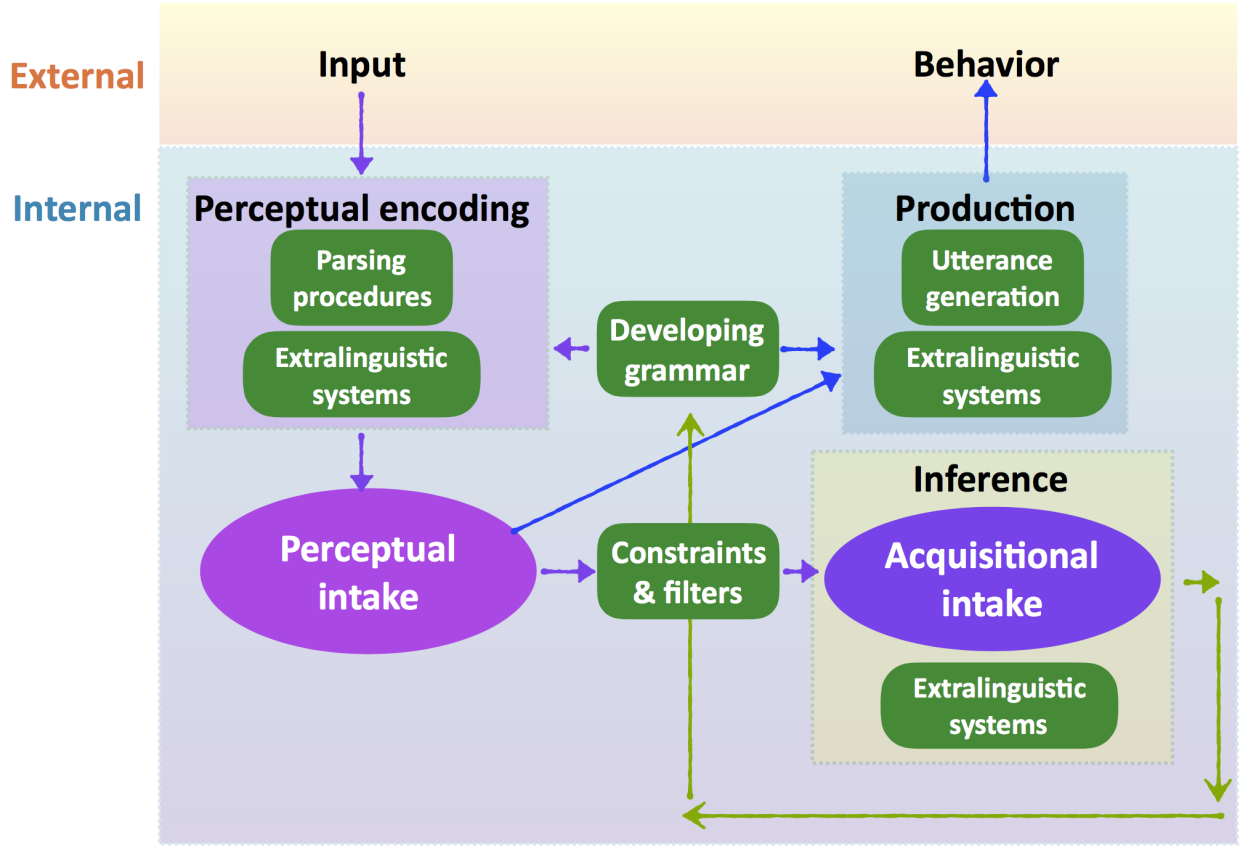


Figure 1: Model of the acquisition process adapted from Lidz & Gagliardi (2015), highlighting the contributions of several key components. Observable components are external to the child (input signal and the child's behavior). Internal components include the pieces used to perceptually encode information from the input signal (developing grammar, perceptual encoding), the pieces used to produce the observable behavior (perceptual intake, developing grammar, and production systems), and the pieces used for inference over the perceptually encoded intake (inference). These yield the next stage of the developing grammar, which itself is used in subsequent perceptual encoding and production.

External components are observable. We can observe the input signal available to children during acquisition, and we can also observe children's behavior at any stage of development, either through naturalistic productions or clever experimental designs.

The internal components involve several clusters. The first cluster centers around perceptual encoding – that is, it concerns the information in the input signal that the child is able to perceive at this stage of development. The *perceptual intake* that results from this depends on the child's developing grammar (i.e., the linguistic knowledge currently available), the parsing procedures that are able to apply that linguistic knowledge in real time to the incoming input signal, and a variety

of extralinguistic systems (e.g., memory, pattern recognition) that are also necessary for extracting information from the input signal. This serves to highlight the intrinsic link between developing representations and developing processing abilities during acquisition (Lidz and Gagliardi, 2015, Omaki and Lidz, 2015, Phillips and Ehrenhofer, 2015, Perkins, Feldman and Lidz, 2017, Perkins, 2019).

When the child generates observable behavior, she relies on the current representations she's been able to perceptually encode (the perceptual intake) and her developing grammar. She then applies her production systems to those representations to generate behavior like speaking (which relies on utterance generation processes) or responding nonverbally (e.g., looking at a picture representing a scene described by an utterance she just heard, which relies on extralinguistic systems like motor control, attention, and decision-making).

The last cluster involves the inference process – that is, the process of updating internal hypotheses about the developing grammar, given the input data perceived as relevant. The data perceived as relevant has been referred to as the *acquisitional intake*, and is based on the perceptual intake. Notably, the acquisitional intake is typically *not* all of the perceptual intake. That is, it's not everything the child is able to encode. Instead, depending on what the child is trying to learn, what's relevant is likely some subset of the perceptual intake. This is one place where UG can have an impact: it can filter the perceptual intake down to the relevant pieces by providing both contraints on possible hypotheses and attentional filters. Inference then operates over the acquisitional intake, and typically involves extralinguistic abilities like probabilistic inference, statistical learning, or hypothesis testing.

For example, consider the *wh*-movement triggers we discussed before. Suppose a child hears *What did the penguin do?* and is at the stage of development when she can perceptually encode several aspects of the utterance: the individual words, the utterance's phrase structure, and the dependency between *what* and the place where it's understood after *do*. This is the child's perceptual intake. If the child is still learning whether her language has *wh*-movement, UG will highlight for her the salient information in her representation: *what* has moved, as indicated by the dependency. This is an example attentional filter provided by UG; the fact that the child is only considering +*wh*-movemement and -*wh*-movement as her hypotheses (rather than various partial *wh*-movement options that only apply in certain contexts) would be an example hypothesis space constraint provided by UG. The *wh*-movement part of her perceptual intake is then relevant for learning about *wh*-movement, and so the *wh*-movement acquisitional intake consists of that information alone (i.e., *what* moved) – the rest of the information is irrelevant for learning about *wh*-movement. Inference occurs over this piece of acquisitional intake, presumably strengthening the child's belief that her language has *wh*-movement.

This inference process thus updates the child's developing grammar. Because the developing grammar is involved in the perceptual encoding process and the behavior production process, every update to the developing grammar can impact both what the child is able to perceive from the input signal and what the child is able to produce. In our example of *wh*-movement, suppose the child has just decided that her language uses *wh*-movement, based on the acquisitional intake accrued. This may allow her developing parser to extract *wh*-dependencies more reliably in future utterances. So, for instance, she might now be able to correctly extract the *wh*-dependency in a longer utterance

like *What did Jack think that Lily said the penguin did?*, where before this information might not have been available to her. Similarly, she may be able to produce longer *wh*-dependencies in her output and respond correctly to utterances involving longer *wh*-dependencies.

Inspired by this characterization of the acquisition process, I want to suggest several components for precisely defining the acquisition task that an acquisition theory is meant to solve.

1. **Initial state**: What knowledge, abilities, and learning biases does the modeled child already have? These include both language-specific and domain-general components. In Figure 1, the initial state encompasses the current status of the developing grammar, parsing procedures, utterance generation, constraints and filters, and extralinguistic systems in perceptual encoding, production, and inference. This is the starting point for the modeled child, given the acquisition task being considered.

2. **Data intake**: What data is the modeled child learning from? This captures the pipeline from the external input signal to perceptual encoding to the acquisitional intake. The external input is what's available for the child to learn from. Her perceptual encoding filters down that input signal to what she's capable of perceiving at this stage of development (the perceptual intake). The acquisitional intake is the result of further winnowing down, based on the constraints and filters in the modeled child's initial state (e.g., provided by UG). So, this is where we can see the immediate impact of the modeled child's initial state – those initial state components influence what's in the modeled child's acquisitional intake.

3. **Inference**: How are updates to the modeled child's internal representations made? This is the inference from Figure 1, which operates over the acquisitional intake to yield the latest developing grammar representations.

4. **Learning period**: How long does the modeled child have to learn? This is represented by the entirety of Figure 1, and primarily involves the developing grammar. That is, this is when the iterative process of updating the developing grammar occurs.

5. **Target state**: What does it mean for the modeled child to succeed at learning? We can assess this by matching model output against observable data from children, i.e., their behavior in Figure 1. So, it's useful to have a model that can generate behavioral output (e.g., Perfors, Tenenbaum and Regier 2011, Pearl and Sprouse 2013a, 2015, Pearl and Mis 2016, Savinelli, Scontras and Pearl 2017, Bar-Sever, Lee, Scontras and Pearl 2018, Bates, Pearl and Braunwald 2018, Savinelli, Scontras and Pearl 2018, Bates and Pearl 2019, Forsythe and Pearl 2019, Nguyen and Pearl 2019, Pearl and Sprouse 2019, Bates and Pearl 2020, Scontras and Pearl 2021). However, if we're confident about which internal representation a particular observable behavior corresponds to (i.e., a specific developing grammar component, such as +*wh*-movement), we may match model output directly against that grammatical knowledge (e.g., Yang 2004, Pearl and Sprouse 2021).

These components correspond to implementable computational model components of an acquisition theory. More specifically, the acquisition task can be characterized by these components,

and an acquisition theory consists of specifying each component according to a theory of developing representations and a theory of developing processing abilities.

In the rest of this chapter, I'll first highlight some modeling framework basics, including considerations of cognitive plausibility, different explanatory levels of modeling, and some commonly used inference mechanisms. I'll then discuss several modeling case studies in syntactic acquisition, including both parametric and non-parametric approaches to modeling syntactic acquisition. I'll conclude with some thoughts on exciting new directions for syntactic acquisition modeling.

2 Some modeling framework basics

2.1 Designing informative models: Cognitive plausibility

In order for our modeling results to tell us something about how children develop syntactic knowledge (i.e., for a model to be an *informative* model and not just an interesting programming exercise), we need to believe that the model reasonably approximates aspects of a child's acquisition process. The way to do this is to make sure the model components are *cognitively plausible*. That is, we make reasonable assumptions for what's actually going on during the acquisition process in children when implementing each model component. But how do we know what's cognitively plausible for any given component? This is where prior theoretical, corpus, and experimental research can help.

Theoretical research can help define parts of the initial state (the developing grammar, the contents of UG), which then impacts the data in the modeled child's acquisitional intake. The inference process may also be defined by theoretical work (e.g., language-specific inference mechanisms, such as the Structural Triggers Learner of Sakas & Fodor: Sakas and Fodor 2001, 2012, Sakas 2016). The target knowledge state (the developed grammar) is also typically defined by theoretical proposals for knowledge representations.

Corpus analysis can help define aspects of the data intake, in particular the input that children encounter as child-directed speech (which they subsequently perceptually encode). Corpus analysis can also provide quantitative descriptions of children's linguistic productions, which is one type of behavior that can indicate the underlying grammar (either the developing representations during the learning period or the target representations after the learning period is completed).

Experimental results can help define parts of the initial state (the parsing and extralinguistic abilities available at a particular age), which impact how the input is perceived (perceptual intake). Experimental results can also define how children's developing production systems generate observable behavior from underlying linguistic representations. Experimental results additionally help define what inference abilities are available for the modeled child, how long the modeled child's learning period is, and what behavior the modeled child should display both during the learning period and afterwards, when the target knowledge has been acquired.

Practically speaking, it can still be non-trivial to make each model component cognitively plausible, despite all this information. For example, let's consider the input component, something external to the child that we might try to estimate using corpus analysis. The key question is analysis of *what*: Do we think the relevant information is contained only in child-directed utterances

(which we have large samples of from electronic resources like the CHILDES database (MacWhinney, 2000)), or do we also need detailed information about the visual scene, accompanying actions, or other discourse context (some of which is available in CHILDES, but far less often)? If what we believe to be relevant input information is not easily available in existing corpora, corpus analysis won't help. We'll have to make an educated guess about what reasonable input looks like for those dimensions that we don't have precise estimates for.

As another example, consider the learning period. For some acquisition problems, we have empirical data about exactly what children know when; for others, we may only have the typically developing adult knowledge state or an atypically developing learner's final knowledge state. In these latter cases, it may be difficult to determine what should count as a plausible learning period. Again, because modeling requires us to implement all the model components explicitly, we may simply have to make an educated guess.

The target state can also present cognitive plausibility challenges, since we have to consider exactly what representations the modeled child ought to learn and what behavior that modeled child ought to produce in order to demonstrate that the target representation has been learned. If we have detailed empirical data available about the stages of learning (e.g., through experimental results or corpus analysis of children's productions), this can be a reasonable comparison for the model child's output – we can try to capture the appropriate learning trajectory (e.g., Alishahi and Stevenson 2008). However, if we don't, we may need to rely on other measures of what counts as acquisition success (Pearl, 2014). Perhaps the modeled child should attain adult-like knowledge. If so, we can use behavior correlated with adult knowledge as the desired target behavior (e.g., Pearl and Sprouse 2013a,b, 2015, Bates and Pearl 2019, 2020, Pearl and Sprouse 2021). Perhaps we have a measure of behavior at one particular age (when the child may not yet have the adult knowledge). If so, we can use that behavior as a metric of what the learner should have learned by that age (e.g., Pearl and Mis 2016, Savinelli et al. 2017, Forsythe and Pearl 2019, Nguyen and Pearl 2019, Pearl and Sprouse 2019). Perhaps we know that the target knowledge will be used to bootstrap future acquisition processes – e.g., the developing grammar impacts future perceptual encoding of linguistic representations. If so, we can measure how useful the modeled child's developing representations are, regardless of whether they match adult representations (e.g., Phillips and Pearl 2014a,b, 2015a, Bar-Sever and Pearl 2016).

The larger point about implementing a model of syntactic acquisition is that we should strive as much as possible to make sure each component is psychologically grounded. If empirical data are available to guide our implementation, then this is straightforward. However, sometimes the data we need aren't available yet, and so we need to make principled decisions about how to implement a given model component. When we make choices that are not derived from empirical data, we have to be prepared to explain why they're reasonable choices and what impact they have on the acquisition process. For more detailed discussion of how to computationally model language acquisition more generally, I recommend the acquisition modeling overviews by Alishahi (2010), Pearl (2010), Räsänen (2012), Freudenthal and Alishahi (2014), Pearl and Sprouse (2015), and Pearl and Goldwater (2016). The key is that when we make model implementation decisions in cognitively plausible ways, our resulting computational model is more likely to tell us something informative about the syntactic acquisition process in children.

2.2 Levels of explanation

Another important consideration is the level of explanation a model is seeking to provide, as this impacts decisions about the modeling components. I find it useful to think about this in terms of Marr's levels of explanation (Marr, 1982): *computational*[1], *algorithmic*, and *implementational*.

A computational-level explanation is about capturing the goal of the acquisition process – that is, what is the cognitive computation the child is trying to carry out? Another way to think about this is that we're trying to correctly conceptualize the acquisition task happening. In the terms we used previously, is it *possible* to reach a specific target state, given a specific initial state and data intake? A computational-level model cares about the form of the underlying representations and biases (initial state), realistic input (which yields the data intake when combined with the initial state), and the checkpoint provided by the target state – all of these components should be implemented in a cognitively plausible way.

Notice, however, that we're abstracting away from the details of the inference mechanism and learning period – we're really just concerned with seeing if we've got the right task description (this is sometimes implemented as an *ideal learner* model, such as those in Foraker, Regier, Khetarpal, Perfors and Tenenbaum 2009, Perfors, Tenenbaum and Wonnacott 2010, Perfors et al. 2011, Pearl 2011, Pearl, Ho and Detrano 2017, Nguyen and Pearl 2019, Pearl and Sprouse 2019, and Pearl and Sprouse 2021). Why do we do this? If we find that it's in fact *impossible* to reach the target state, given the initial state and data intake, this is a signal that we may not be describing the acquisition task correctly. So, if we try to implement specific learning strategies to solve that acquisition task, we'll probably find that none of them work (see Pearl (2011) for an example of this in metrical stress acquisition). We can save ourselves from being doomed to failure by making sure we first have a reasonable computational-level model of the acquisition task. A successful computational-level model will validate the underlying representations and biases in the initial state that went into it, because this model demonstrates how those initial state assumptions can lead to the target state using realistic input. In short, a computational-level model can demonstrate that it's possible *in principle* to solve that acquisition task using a specific set of assumptions about the initial state.

But what about being possible for humans? This is where an algorithmic-level model helps us (sometimes implemented as a *process* model, such as those in Regier and Gahl 2004, Yang 2004, Pearl and Lidz 2009, Pearl and Sprouse 2013a, and Pearl and Mis 2016). Algorithmic-level models are concerned with the steps humans would carry out to solve the acquisition task, and often include considerations of incremental processing and limited memory, as well as a realistic learning period for the steps to be carried out in. An algorithmic-level model can demonstrate that the acquisition task can be solved with the cognitive and time limitations children have. That is, it's possible *for children* to solve the specified acquisition task using a particular set of assumptions about the initial state.

What typically differs between computational-level and algorithmic-level models is the in-

[1]Note that this is a different use of "computational" than what we mean when we talk about "computational modeling". This is an unfortunate overloading of the term "computational", with two distinct meanings. Computational as an explanation level is what I talk about in this section. Computational as a modeling technique means implementing a model concretely with a computer program. I'll try to distinguish these meanings by using "computational-level models" for the explanation level and "computational modeling" for the technique.

ference process implementation. A computational-level inference process is focused on simply completing the computation accurately – there is no claim that the way inference is being carried out in the model is the way humans carry out inference. As a concrete example, Gibbs sampling is a statistical inference algorithm currently used in some computational-level models (e.g., Perfors et al. 2010, 2011, Phillips and Pearl 2015b, Pearl and Sprouse 2019), and it has the happy property of being guaranteed to converge on the best answer if given enough time to search the (potentially infinite) hypothesis space. While this inference algorithm involves a particular (clever) process of iteratively searching the hypothesis space, it's unlikely that humans perform the same process when they're doing inference. So, models using this inference process typically aren't concerned with learning period considerations, which encode how long the child has to perform the steps of the inference process. Put simply, because the steps of computational-level inference aren't the same as the steps of human inference, paying attention to how long it takes a modeled child to accomplish computational-level inference isn't really related to how long children have to accomplish their inference.

This contrasts with algorithmic-level models, whose inference process is meant to represent the steps children would go through to perform their inference (e.g., Regier and Gahl 2004, Yang 2004, Pearl and Lidz 2009, Pearl and Sprouse 2013a, and Pearl and Mis 2016). This is why algorithmic-level models are sensitive to the cognitive limitations children have when performing inference (e.g., limited memory) and the time limitations imposed by the learning period. In some cases, the algorithmic-level inference is known to be a heuristic approximation of computational-level inference (e.g., Regier and Gahl 2004, Pearl and Lidz 2009, Bonawitz, Denison, Chen, Gopnik and Griffiths 2011, Pearl and Mis 2016); in other cases, it's not obviously so (e.g., Yang 2004, Pearl and Sprouse 2013a).

The last explanation level is implementational, and this is concerned with how the cognitive computation of the acquisition task is implemented in the brain. This has direct links to the algorithmic-level explanation: If we think we have the right steps to carry out the acquisition computation, how are they actually carried out in the available neural medium? This involves a deep understanding of neural architecture, as the specific relevant properties of the brain need to be simulated (e.g., how information is represented in a distributed manner, the way neurons spread information to each other, the structural subdivisions of the brain). This consideration has significant impact on how the initial state assumptions, data intake, inference, and target state are encoded in the model. As an example focusing on the initial state, if we think there's a bias to have a default *wh*-movement value of no movement, what does that actually look like neurally? That's what we would need to encode in an implementational level model. Because the field is currently developing the linking theories between linguistic theory and cognitive neuroscience, this represents an exciting area of future collaborative research for syntacticians, computational modelers, and cognitive neuroscientists.

2.3 Inference

As noted above, inference is how the modeled child updates the developing grammar representations. Below I discuss some common inference components and inference implementations used in the syntactic acquisition modeling literature, though this is by no means an exhaustive list.

2.3.1 Counting things

A very common component in the modeled inference process is the ability to count things, which is a domain-general ability. In syntactic acquisition models, the more important thing for inference is *what's being counted* – this is where the learning theory and representational theory come in handy. These theories, through the initial state and its effect on the data intake, determine what actually makes it to the inference process – i.e., what things are counted. Depending on the theories involved, the inference mechanism could be operating over counts of lexical items (Yang, 2005, Freudenthal, Pine, Aguado-Orea and Gobet, 2007, Freudenthal, Pine and Gobet, 2009, 2010, Freudenthal, Pine, Jones and Gobet, 2015, Yang, 2016), syntactic category sequences (Perfors et al., 2010, 2011, Pearl and Sprouse, 2019), syntactic signals realized in certain overt structures (Sakas and Fodor, 2001, Yang, 2004, Legate and Yang, 2007, Mitchener and Becker, 2010, Pearl and Sprouse, 2013a, Becker, 2014, Sakas, 2016, Pearl and Sprouse, 2021), or something else entirely. Importantly, for our purposes as modelers, the inference mechanism itself doesn't change. Once we define the units over which inference is operating, inference then simply operates.

Counts are typically translated into probabilities (e.g. 700 instances of an element appearing in 1000 data points translated to $\approx$[2] 0.7), and this provides a sense of how relatively frequent the element is. Probabilities also have an intuitive interpretation as beliefs about categorical options. For example, let's consider the two *wh*-movement options: +*wh*-movement and -*wh*-movement. A probability of 0.7 for +*wh*-movement (and a probability of 0.3 for -*wh*-movement) might reasonably be interpreted as the modeled child believing +*wh*-movement is more likely to be true. This can also be thought of as the modeled child picking +*wh*-movement 70% of the time if asked to generate an utterance involving a *wh*-word. In this way, probabilities can make it easier to link categorical representations to observable behavior (e.g., a child using +*wh*-movement in her utterances about 70% of the time).

As a practical note, probabilities derived from counts typically involve *smoothing*, where things never observed still have a very small amount of probability assigned to them. This is because zero probability is very final – something with zero probability can never, ever occur. In contrast, something with a very small probability may occur only very occasionally, but it's not impossible. This is useful when we consider the acquisition task: Children get a finite data sample and have to make generalizations from it. It could be that some element they never observe is actually ungrammatical – but it could also be the case that it's just very rare. So, assigning the element a very small probability gives some wiggle room for generalization purposes. Perhaps later on, evidence will come in that indicates this element is just fine, such as actually hearing someone use it in naturalistic speech. If the element has zero probability, no amount of evidence can change that belief. But, if the element has a very small probability, belief in the element can still be adjusted.

Smoothing can be intuitively implemented via pseudocounts, where the modeled child imagines she's seen each element a certain number of times (typically <1) before ever encountering any data. In our *wh*-movement example, we might implement this as the modeled child using a pseudocount of 0.5 for each option. So, initially, the probabilities for each option would look like this:

[2] See the discussion about smoothing below for why the $\approx$ is used.

$$\text{p}(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{0+0+2*0.5} = \frac{0.5}{2*0.5} = 0.5$$
$$\text{p}(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{0+0+2*0.5} = \frac{0.5}{2*0.5} = 0.5$$

This means that as soon as the model's seen a single data point (for either option), the probability is very much swayed in that direction. Let's say we see 1 +*wh*-movement data point.

$$\text{p}(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{1+0.5}{1+0+2*0.5} = \frac{1.5}{2} = 0.75$$
$$\text{p}(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{1+0+2*0.5} = \frac{0.5}{2} = 0.25$$

Lots of data will yield a probability pretty close to the unsmoothed probability (e.g., the case where we see 700 +*wh*-movement data points and 300 -*wh*-movement data points).

$$\text{p}(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{700+0.5}{700+300+2*0.5} = \frac{700.5}{1001} = 0.6998$$
$$\text{p}(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{300+0.5}{700+300+2*0.5} = \frac{300.5}{1001} = 0.3002$$

I should note that translating counts directly to probabilities like this is just one way to generate probabilities for different aspects of the developing grammar (e.g., used by Pearl and Sprouse 2013a). I discuss three other common inference approaches below that generate probabilities for competing hypotheses by using more sophisticated reasoning over the counts available.

2.3.2 Reinforcement learning

Reinforcement learning is a principled way to update the probability of a categorical option which is in competition with other categorical options, e.g., +*wh*-movement vs. -*wh*-movement. Yang (2002, 2004) uses an implementation by Bush and Mosteller (1951) called the *linear reward-penalty scheme*. As the name suggests, there are two choices when a data point is processed – either the categorical option under consideration is rewarded or it's penalized. This translates to the option's current probability being increased (rewarded) or decreased (penalized).

The inference process's update function involves a parameter γ that determines how quickly the modeled child updates her beliefs on the basis of a single data point (i.e., how much she shifts the probability between the different options). The larger γ is, the more probability the modeled child shifts after processing a single data point. The other consideration is what the previous probability of the option was – this impacts how much shifting is possible. So, the important variables are these:

(1) a. γ: This is the learning rate (and is typically a number smaller than 1).

b. $p_{o_{1_t}}$: This is the previous probability of the option under consideration (option 1) o_1 at time t. For example, this might be p_{+wh_t}.

c. $p_{o_{i_t}}$: This is the previous probability of the opposing option(s) o_i for each of the i remaining options at time t. For example, this might be p_{-wh_t} when we only have the two options, +*wh*-movement and -*wh*-movement, competing with each other.

When the modeled child processes a data point, these probabilities are updated by focusing on one option (e.g., +*wh*-movement) and updating the collective option probabilities; the update itself is based on whether the option under consideration is rewarded for being compatible with

the data point or penalized because it's not compatible with the data point. For example, if the data point *What did the penguin do?* is processed considering the +*wh*-movement option, the data point can be accounted for: The fronting of *what* can be due to movement. So, the +*wh*-movement option is rewarded – its probability increases and the probability of all other options (here, simply -*wh*-movement) decreases. This is shown in (2b), with the learning rate γ=0.1, and the previous probabilities of both options = 0.5. In contrast, suppose the data point is an echo question like *The penguin did what?*. If this is processed with the +*wh*-movement option, the data point can't be accounted for because movement of *what* is expected, but not seen here. So, this data point would penalize +*wh*-movement – its probability decreases, as shown in (2c).

(2) Linear reward-penalty

a. p_{+wh_t} = 0.5, p_{-wh_t} = 0.5, γ=0.1, 2 options (N=2)

b. reward +*wh*-movement:

i. $p_{+wh_{t+1}} = p_{+wh_t} + \gamma*(1-p_{+wh_t}) = 0.5 + 0.1*(1-0.5) = 0.55$

ii. $p_{-wh_{t+1}} = (1-\gamma)* p_{-wh_t} = (1-0.1)*0.5 = 0.45$

c. penalize +*wh*-movement:

i. $p_{+wh_{t+1}} = (1-\gamma)*p_{+wh_t} = (1-0.1)*0.5 = 0.45$

ii. $p_{-wh_{t+1}} = \frac{\gamma}{N-1} + (1-\gamma)*p_{-wh_t} = \frac{0.1}{2-1} + (1-0.1)*(0.5) = 0.55$

Interestingly, while this is typically an algorithmic-level model inference procedure simulating the steps a child would use to accomplish inference, it has predictable convergence behavior when unambiguous data are available in the data intake. More specifically, because unambiguous data are compatible only with a single option (e.g., *wh*-fronting data compatible with +*wh*-movement, but not -*wh*-movement), these data will always reward that option – presumably, the correct one for the language – and penalize all other options. This will cause the correct option to receive a probability near 1, given enough unambiguous data in the child's intake. So, models implementing this kind of inference procedure will often do analyses of the child's intake with respect to unambiguous data, rather than implementing this inference mechanism explicitly (e.g., Legate and Yang 2007, Yang 2012, Legate and Yang 2013) – see section 3.1.2 for more detailed discussion of studies of this kind. For example, a data intake analysis might demonstrate that +*wh*-movement has a 25% advantage over the -*wh*-movement option. This could occur because 30% of the data are unambiguous for +*wh*-movement and 5% of the data are unambiguous for -*wh*-movement, yielding the 30-5=25% +*wh*-movement advantage. On the basis of this alone, we would conclude that a modeled child using a linear reward-penalty inference mechanism would converge on the +*wh*-movement value for her language. Again, this is *without* needing to explicitly implement the inference algorithm. This property can make this inference mechanism very attractive for testing different acquisition proposals. It also highlights the simulation effort that can be saved by understanding the known properties of the inference mechanism being modeled.

2.3.3 The Tolerance Principle

The Tolerance Principle (Yang, 2005, 2016) is another inference mechanism that has predictable behavior, given algorithmic-level considerations. As its name suggests, there's a certain aspect of

the inference process captured by the Tolerance Principle that's completely predictable (i.e., principled) on the basis of the counts of relevant data intake items; this aspect is how many exceptions a potential generalization can "tolerate" and still remain a useful generalization for the child to make.

More specifically, the Tolerance Principle is a formal approach for determining when a child would choose to adopt a generalization or "rule" to account for a set of items Yang (2005, 2016). This principle is based on algorithmic-level considerations of knowledge storage and retrieval in real time, incorporating how frequently individual items occur, the absolute ranking of items by frequency, and serial memory access. It's a sophisticated way of counting things that's been applied to the acquisition of both morphosyntactic and syntactic knowledge, including English *a*-adjective morphosyntax (Yang, 2015, 2016), English dative alternations (Yang, 2016), and linking patterns between thematic roles and syntactic positions associated with verbs (Pearl and Sprouse, 2021).

The intuition behind the Tolerance Principle is that the child is optimizing retrieval time. In particular, suppose a child is considering a rule that connects an item to some other information, such as whether verb arguments with thematic roles (e.g., *Penguins* as AGENT and *fish* as PATIENT in *Penguins eat fish*) link to specific verbal syntactic positions, like *subject* or *object* (Pearl and Sprouse, 2021). The potential rule compactly encodes some regularity – this is the pattern that several items in the dataset under consideration follow (e.g., a default linking pattern that connects AGENTs to *subject* position, and PATIENTs to *object* position). When does it become useful to have a rule? One answer is that it's useful when having a rule makes the average retrieval time for any item in the dataset faster. For example, suppose having a default linking rule in English makes it easier on average to pull up the correct interpretation of any verb and its arguments when they're used in a specific syntactic context. In this case, it's useful for the child to have the default linking rule. In contrast, if having a rule like that doesn't on average speed up retrieving an interpretation of a verb and its arguments, then it's not useful to have the default linking rule.

Yang (2005, 2016) specifies this utility by considering how long it would take to access an item's target information with vs. without the rule. The threshold for adopting the rule – its "tipping point" (Yang, 2016) – is determined by a fairly complex equation (see Yang 2005, 2016) that takes into account the cognitive aspects I mentioned above: item frequency, item rank by frequency, and serial memory access. Happily, this complex equation turns out to be well approximated by a much simpler equation: $\frac{N}{ln(N)}$, where N is the number of items the rule could potentially apply to. That is, if there are $\frac{N}{ln(N)}$ or fewer exceptions in the set of items the rule could apply to, it's useful in terms of retrieval time to adopt the rule.

So, the Tolerance Principle requires a certain number of items that match a rule in order for that rule to be adopted as useful: that number is $N - \frac{N}{ln(N)}$. If there aren't that many items that match the rule, the rule isn't useful because adopting the rule slows down the average retrieval time. The number of items a rule could vs. does apply to from the child's perspective is based on an estimate of that child's knowledge of the relevant items. This is information that can usually be assessed from corpus or behavioral data. In particular, corpus data of realistic child-directed speech can be used to estimate the relevant items the child has encountered so far in her input, which would then comprise her knowledge of the relevant items. Behavioral data, such as estimates of words children do or don't know at specific ages (e.g., from Stanford's Wordbank: Frank, Braginsky, Yurovsky

and Marchman 2017 [http://wordbank.stanford.edu]), can also be used as a direct estimate of a child's knowledge of relevant items.

The simplicity of applying the Tolerance Principle once the relevant items are known makes it a very attractive inference mechanism to use within an acquisition model when appropriate (i.e., when the question is whether a child would adopt a certain generalization or not). That is, the Tolerance Principle provides a snapshot of the tipping point for a generalization – to generalize or not to generalize – on the basis of the child's internal knowledge of relevant items at that point in time. Importantly, the Tolerance Principle doesn't aim to capture a process of updating over time, unlike the other inference mechanisms discussed in this section (reinforcement learning in the previous subsection, Bayesian updating in the next subsection). Also, the Tolerance Principle doesn't describe the steps (i.e., the algorithm) a child would use to achieve the calculation of the tipping point for generalizations. This makes the Tolerance Principle more a computational-level inference mechanism, even though it's derived from algorithmic-level cognitive aspects such as serial memory access.

2.3.4 Bayesian updating

Bayesian updating is another kind of probabilistic inference mechanism, and it involves both prior assumptions about the probability of different options (typically referred to as hypotheses) and an estimation of how well a given hypothesis fits the data. Pearl and Goldwater (2016) provide a detailed overview of Bayesian updating for language acquisition modeling, and I'll summarize the key points here.

A Bayesian model assumes the learner has some space of hypotheses H, each of which represents a possible explanation for how the data D in the data intake were generated. Given D, the modeled child's goal is to determine the probability of each possible hypothesis $h \in H$, written as $P(h|D)$, which is termed the *posterior* for that hypothesis. This is calculated via Bayes' Theorem as shown in (3).

(3) $$P(h|D) = \frac{P(D|h)*P(h)}{P(D)} = \frac{P(D|h)*P(h)}{\sum_{h' \in H} P(D|h')*P(h')} \propto P(D|h) * P(h)$$

Let's walk through each of the terms in (3). In the numerator, $P(D|h)$ represents the *likelihood* of the data D given hypothesis h, and describes how compatible that hypothesis is with the data. Hypotheses with a poor fit to the data (e.g., the -*wh*-movement hypothesis for a dataset with 30% unambiguous +*wh*-movement data) have a low likelihood; hypotheses with a good fit to the data have a high likelihood. How the likelihood itself is calculated is often a modeling choice. A simple way is basic compatibility, similar to the linear reward-penalty scheme: If the hypothesis can't account for the data, its likelihood is (near) 0. In contrast, if the hypothesis can account for the data, its likelihood is (near) 1.

$P(h)$ represents the *prior* probability of the hypothesis. Intuitively, this corresponds to how plausible the hypothesis is, irrespective of any data. This is often where considerations about the complexity of the hypothesis will be implemented by the modeler (e.g., considerations of simplicity or economy, such as those included in the grammar evaluation metrics of Chomsky 1965, and those explicitly implemented in Perfors et al. 2011). So, for example, more complex hypotheses may have lower prior probabilities.

The likelihood and prior make up the numerator of the posterior calculation, while the denominator consists of the normalizing factor $P(D)$, which is the probability of the data under any hypothesis. Mathematically, this is the summation of the likelihood * prior for all possible hypotheses in H, and ensures that all the hypothesis posteriors sum to 1. Notably, because we often only care about how one hypothesis compares to another, calculating $P(D)$ can be skipped over and the numerator alone used (hence, the $\propto$ in (3)). As a concrete example, suppose we have two hypotheses in H, h_1 and h_2. Suppose we calculate the likelihood * prior (lp) for each, and get the following: lp_{h_1} = 0.2, lp_{h_2} = 0.1. If we only care how h_1 compares to h_2, we can calculate this directly (called the *posterior odds ratio*) as $\frac{0.2}{0.1} = 2$, which indicates h_1 is twice as probable as h_2. If we take the time to normalize the probabilities and calculate the posterior for each hypothesis, we would get $p(h_1|D) = \frac{0.2}{0.2+0.1} = 0.666$, $p(h_2|D) = \frac{0.1}{0.2+0.1} = 0.333$. Doing the same comparison between h_1 and h_2 as before, we get $\frac{0.666}{0.333} = 2$, which again shows the h_1 is twice as probable as h_2, given these data. So, because this relative hypothesis comparison is the information the modeled learner is typically interested in, modelers will often use the likelihood * prior probabilities for each hypothesis, rather than normalizing the hypothesis probabilities by $p(D)$.

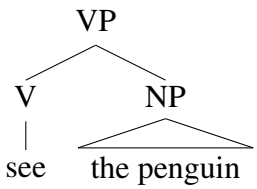
Something to note is that Bayesian updating can be done as a computational-level inference, with the data seen all at once and inference conducted over that batch of data (e.g., Foraker et al. 2009, Perfors et al. 2010, 2011, Pearl and Sprouse 2019). Bayesian updating can also be done as algorithmic-level inference, with the data seen incrementally and inference conducted over each data point as it's encountered (e.g., Regier and Gahl 2004, Pearl and Lidz 2009, Pearl and Mis 2016).

Another interesting aspect of Bayesian updating is that it can be implemented hierarchically, such that there are different levels of abstraction where the hypothesis space is concerned (Goodman, 1955, Kemp, Perfors and Tenenbaum, 2007, Kemp and Tenenbaum, 2008). More specifically, there may be hypotheses about specific observable properties, and more abstract hypotheses (called *overhypotheses*) about the nature of those more specific hypotheses. A concrete example taken from Pearl and Lidz (2013) can help demonstrate the intuition of this, specifically linking it to the concept of linguistic parameters used in the generative linguistics tradition (Chomsky, 1981).

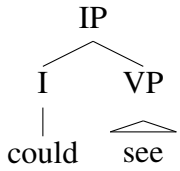
Suppose an English-learning child has a collection of utterances in her data intake. Some utterances contain verbs and objects, and whenever there's an object, suppose it appears after the verb, e.g., *see the penguin* rather than *the penguin see*. Other utterances contain modal verbs and nonfinite main verbs, and whenever both occur, the modal verb precedes the main verb, e.g., *could see* rather than *see could*. A child could be aware of the shared structure of these sentences – specifically that these observable forms can be characterized as the head of a phrase appearing before its complements, as shown in (4).

(4) Shared structure in observable forms: Phrasal heads before complements

a. *see the penguin*



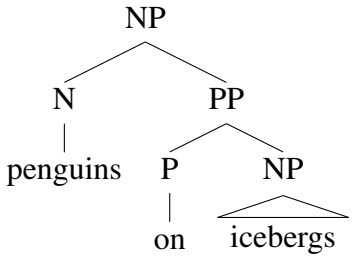
b. *could see*



This "head-first" idea can be encoded at a level that describes utterances in general, and would represent an overhypothesis about structure which is similar to the classical headedness linguistic parameter: namely, phrases have their heads before their complements. Each of the examples from (4) instantiates this overhypothesis for particular phrases, VPs and IPs, and so predicts the specific hypotheses these data support: VPs have their heads before their complements and IPs have their heads before their complements. Each of these specific hypotheses strengthens the more abstract head-first overhypothesis, and allows the child to make predictions about phrases not yet encountered. For example, if this child encountered an utterance with a preposition surrounded by NPs, such as *penguins on icebergs*, there are (at least) two structural hypotheses possible, shown in (5).

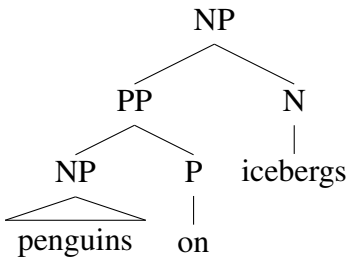
(5) Possible structures for *penguins on icebergs*

a. head-first:



Meaning: penguins who are on icebergs

b. head-last:



Meaning: icebergs that are on penguins

Using the head-first overhypothesis for both the NP and the PP, the child would prefer the first structure, whose meaning indicates this is about penguins who are on icebergs (rather than icebergs that are on penguins). Importantly, the child could have this head-first preference about NP and PP structure, even without having encountered a single NP or PP data point. This is the power of overhypotheses, and is the same powerful intuition that made the classical notion of linguistic parameters attractive to developmental linguists. Children can learn how to make generalizations for data they've never seen before on the basis of the data they do see precisely because of the abstract system underlying the generation of the observable data.[3]

[3]This way of learning is sometimes referred to as children using indirect positive evidence, and complements

In short, the classical notion of a linguistic parameter is an abstract structural property that constrains the hypothesis space of the child. So, a classical linguistic parameter is an overhypothesis about the specific structural hypotheses available. Because of this, there may be a very intuitive connection between hierarchical Bayesian inference and syntactic acquisition with parameters, and this particular application of Bayesian updating is an exciting area of future research in syntactic acquisition.

3 Some actual models

As with the brief review of some of the common syntactic acquisition modeling frameworks, this is an illustrative overview of several syntactic acquisition models, rather than an exhaustive one. Due to space considerations, I'll be focusing on models of syntactic acquisition at a developmental stage when syntactic categories are already known. However, I should note that modeling the fundamental process of syntactic category acquisition is an energetic research area as well (e.g., Mintz 2003, Xiao, Cai and Lee 2006, Wang and Mintz 2008, Chemla, Mintz, Bernal and Christophe 2009, Erkelens 2009, Frank, Goldwater and Keller 2009, Weisleder and Waxman 2010, Wang, Höhle, Ketrez, Küntay, Mintz, Danis, Mesh and Sung 2011, Yang 2011, Frank, Goldwater and Keller 2013c, Gutman, Dautriche, Crabbé and Christophe 2015, Bar-Sever and Pearl 2016, Bates et al. 2018, among many others). Most of the approaches reviewed below assume that syntactic categories – and often phrase structure – are (near) adult-like before the syntactic acquisition processed being modeled would initiate. For each one, I'll briefly describe the general acquisition task each model is attempting to capture, using the modeling components already discussed: initial state, data intake, inference, learning period, and target state. I'll then summarize the main findings for each model, highlighting the connections to theories of linguistic representation and theories of the learning process.

3.1 Parametric approaches

As mentioned above, a linguistic parameter – just like a statistical parameter – is meant to be an abstraction that can account for multiple observations (e.g., having a head-first value accounts for multiple observations of phrase structure across different phrase types). Once the child knows the value of the parameter, she can use it to make predictions about other data she expects to see. This is an excellent property from an acquisition standpoint, because a child can learn about the linguistic parameter's value from a variety of data she encounters, and then use it to infer the form of other data she may rarely (or never) encounter (Hyams, 1987, Pearl and Lidz, 2013). That is, by using parameters, the range of observations a child has to make to learn her syntactic grammar is greatly reduced. So, linguistic parameters are a way to solve the induction problems associated with syntactic acquisition (sometimes called the *poverty of the stimulus*: see Pearl (2020) for a recent overview) – the child constrains her syntactic generalizations on the basis of her data intake coupled with the linguistic parameters she's already aware of.

learning with direct positive evidence and indirect negative evidence. See Pearl and Mis (2016) for more discussion on this point, and its impact on acquisition.

For this reason, syntactic acquisition using parameters has been a core area of acquisition modeling in generative syntax (Clark, 1992, Gibson and Wexler, 1994, Niyogi and Berwick, 1996, Fodor, 1998b,a, Sakas and Fodor, 2001, Sakas and Nishimoto, 2002, Yang, 2002, Sakas, 2003, Yang, 2004, Fodor and Sakas, 2005, Fodor, Sakas and Hoskey, 2007, Sakas and Fodor, 2012, Sakas, 2016, Fodor, 2017, Fodor and Sakas, 2017). Notably, this is true despite disagreement over the specific linguistic parameters needed to account for the world's languages and the non-trivial relationships between different hypothesized parameters (e.g., see discussion in Boeckx and Leivada 2014, who note that once you assume a realistic number of parameters, you end up with a "complex, subway-map-like network"). This complexity is one reason why many parametric learning approaches focus on how to navigate the acquisition of syntax using linguistic parameters. Put simply, just because a child already knows these parameters exist doesn't mean it's obvious how to connect them to the observable data. The acquisition theory work is often about how the child can recognize the parametric signatures in her data intake and use them to acquire the correct syntactic parameter values for her language. The modeling work then implements different proposals for doing this.

3.1.1 Learning by parsing: The structural triggers learner

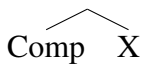
The *structural triggers learner* (STLearner) is an acquisition proposal investigated primarily by Fodor and Sakas (Fodor, 1998b,a, Sakas and Fodor, 2001, Sakas and Nishimoto, 2002, Fodor and Sakas, 2005, Fodor et al., 2007, Sakas and Fodor, 2012, Sakas, 2016, Fodor, 2017, Fodor and Sakas, 2017). This approach considers the problem of learning the syntactic parameters for a language from observable data that contain *structural triggers* recognized by a child during the process of parsing the input. For this reason, it's sometimes referred to as the *learning-by-parsing* approach.

The key intuition of the STLearner is that the child wants to comprehend the utterances around her, and uses her ability to parse an utterance – successfully or not – as a way to identify the structural properties of her language's grammar. Her *a priori* knowledge of linguistic parameters is implemented as a set of "treelet" options she can use to decode the structure of an utterance. For example, the treelets corresponding to a headedness parameter would include a "head-first" option and a "head-last" option, and could be implemented as in (6).

(6) Treelets for the headedness parameter

a. head-first:

b. head-last:



As the child incrementally encounters an utterance, she uses the treelets available to parse the structure of the utterance. In the initial stages of learning, she hasn't settled on any parametric options yet, so all the treelets are available (e.g, both the head-first and head-last treelet). As she identifies the language-specific treelets corresponding to her language's grammar (e.g., head-first

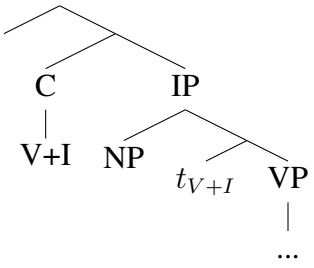
for English), only the language-specific treelet is available for parsing. So, the developing grammar consists of the language-specific treelets that are available to be deployed during parsing. In the model terms from before, the STLearner uses parsing with the developing grammar options to transform the input into the perceptual intake.

To transform the perceptual intake into acquisitional intake, the STLearner leverages parsing ambiguity to identify reliable parametric cues to learn from. More specifically, the observable signatures of different parametric options can be difficult to detect, especially in the early stages of acquisition when few parameters are set. This is because an observable utterance can be compatible (here: parseable) with combinations of different parametric options (here: treelets) – that is, an utterance is ambiguous for which parametric options it signals.

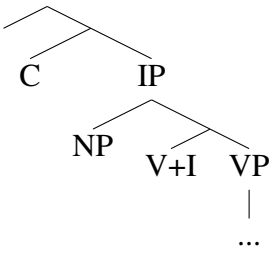
As a simple example of the ambiguity problem, let's consider some example parameters from the CoLAG dataset that Fodor and Sakas use to investigate the STLearner. This dataset[4] consists of 3072 natural-language-like artificial languages constructed using 13 binary syntactic parameters that relate to grammatical phenomena such as head direction, null subjects, *wh*-movement, and topicalization. Three parameters are (i) a headedness parameter that connects to various phrases including VP, (ii) ItoC movement, and (iii) Optional Topic. The treelets for headedness are above in 6, and the treelets for ItoC movement and Optional Topic are in (7-8). For +ItoC, the inflection *I* (typically joined with the verb, such as in *see+ed* = *saw*: *V+I*) moves to the C position; -ItoC doesn't allow this movement. For Optional Topic (?Top), some phrase (indicated with *XP*) can move to the specifier position of CP, but doesn't have to; for Obligatory Topic (+Top), some phrase *must* move to that position.

(7) Treelets for the ItoC parameter

a. +ItoC:



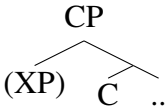
b. -ItoC:



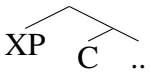
[4]Available at http://www.colag.cs.hunter.cuny.edu/downloadables.html

(8) Treelets for the Optional Topic parameter

a. Optional Topic (?Top):



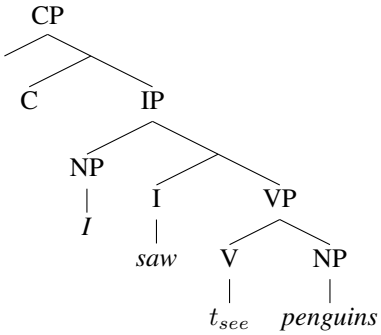
b. Obligatory Topic (+Top):



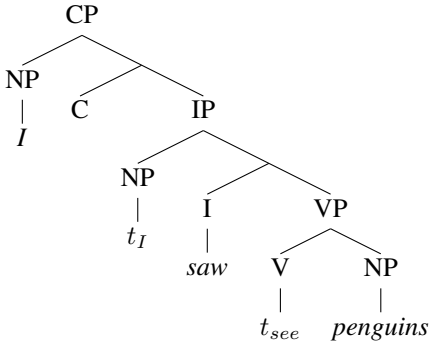
Now, with these parametric treelets in mind, suppose a child encounters the following utterance: *I saw penguins.* With only these three parameters, we have seven options for how to parse this utterance, as shown in (9).

(9) Seven parses for *I like penguins*, assuming V-to-I movement (t_{see} in *V*, *saw* in I)

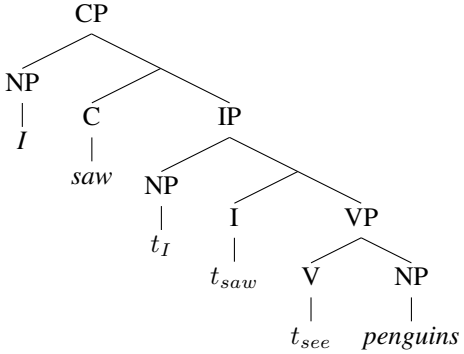
a. Parse 1: head-first, -ItoC, ?Top (no movement option)



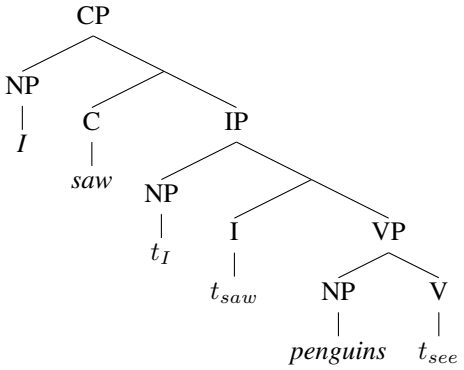
b. Parse 2: head-first, -ItoC, ?Top (movement option) **or**
Parse 3: head-first, -ItoC, +Top



c. Parse 4: head-first, +ItoC, ?Top (movement option) **or**
Parse 5: head-first, +ItoC, +Top



d. Parse 6: head-last, +ItoC, ?Top (movement option) **or**
Parse 7: head-last, +ItoC, +Top



Notably, every single parametric option is used in some possible parse (head-first vs. head-last, +ItoC vs. -ItoC, ?Top vs. +Top). So, at first blush, it's unclear exactly what to take away from this utterance. What is this data point telling the child about winnowing down her grammatical choices (if anything)? This highlights the learning issues surrounding ambiguous data. As this data point shows, early in learning before the child has set (m)any parameters for her language, data are more likely to be ambiguous. So, what's a child to do?

This is where the process of parsing is helpful for an STLearner. Fodor and Sakas suggest that an STLearner can pay attention to where in the parse the ambiguity arises. Using that information, an STLearner can pull out useful pieces of information in order to update the developing grammar. That is, by paying attention to where ambiguity arises during the parse, a child can make a more informed grammar update. This grammar update then directly affects subsequent perceptual encoding and ambiguity detection. The discoveries that Fodor and Sakas have made in their STLearner investigations are primarily about how children could learn with syntactic parameters in the presence of ambiguity.

I'll now summarize the main components of the STLearner approach, in the modeling terms from before. The initial state of the learner includes knowledge of the parametric treelets, the ability to deploy these treelets during parsing, and the ability to detect ambiguity during parsing. Sakas (2016) notes that the parametric treelets may be building blocks derived from a more fundamental generative system (such as that used by the Minimalist Program). This would align with the idea that the parametric treelets represent explicit structural hypotheses selected from the larger implicit

hypothesis space that this generative system can produce (see Perfors 2012 for helpful discussion about explicit and implicit hypothesis spaces more generally in cognition).

The input for the STLearner are utterances from the target language, abstracted into structural categories such as *Subject, Auxiliary, Verb, Object* and *Preposition*, along with observable features such as *finiteness* and *illocutionary force*. For example, the utterance *I saw penguins* would be represented as something like *Subject Verb*$_{+finite}$ *Object*$_{direct}$ *[declarative]*. Notably, the utterance types in the STLearner's input don't have the same frequencies as the ones in child-directed speech. For example, there is one utterance for *Subject Verb*$_{+finite}$ *Object*$_{direct}$ *[declarative]* in the STLearner's input, even though a child might encounter many sentences of this form (*I saw penguins, The penguins were cute, A penguin ate a fish*, etc.). So, this is an abstraction of the input children actually do encounter, and impacts the kinds of questions that can be addressed. In particular, STLearner investigations are often focused on treacherous ambiguities that are salient in this kind of abstracted input, and would presumably be even more damaging if the input frequencies were a more direct reflection of child-directed speech.

The STLearner's perceptual intake is typically the full parse of the utterances using parametric treelets as long as there's no ambiguity; when there's ambiguity, the perceptual intake may either be a sample full parse highlighting the points of ambiguity or a partial parse that ceases at the first point of ambiguity. The exact perceptual intake depends on the acquisition proposal being implemented, as does the acquisitional intake. STLearner acquisition proposals typically differ on whether the acquisitional intake includes ambiguous portions of the parse or not.

Inference implementations are typically at the algorithmic level, and split into deterministic variants that rely on triggering and non-deterministic variants that rely on trial and error search. The more common deterministic variant incrementally processes the acquisitional intake, and then makes a single irreversible decision about the correct parametric treelet for the language when the learner deems it appropriate. This decision can incorporate simplicity metrics for dealing with ambiguity, such as preferring treelets that don't involve movement (sometimes called the Minimal Chain Principle) and preferring treelets that don't involve null elements (sometimes called Avoid Empty Category) (Sakas and Fodor, 2012). Non-deterministic variants often incorporate a reinforcement learning component, where treelets used for successful parses have their probability increased.

Because STLearner investigations are typically concerned with relative learnability of different languages within the CoLAG domain, there are no *a priori* restrictions on the length of the learning period. Rather, what's interesting is how long the learning period would need to be in order to learn one language vs. another, or what the maximum learning period would need to be to learn any of the 3072 languages in the domain, given the ambiguities present. The target state is simply the correct set of parameter values (parametric treelets) for the target language. Learning periods are measured by the number of utterances processed, and have ranged from just a few utterances to hundreds of thousands to infinite (in the case of ambiguities that can't be overcome).

Key findings from the STLearner approach include the following:

1. The ability to detect ambiguity during parsing is crucial for a learning strategy relying on deterministic inference.

2. Having a way to learn something from ambiguous data is useful – if a deterministic learner

is waiting for perfectly unambiguous data, she may be waiting forever.

3. A potentially useful way to handle ambiguous data is to have additional learning biases that refine the inference done in the presence of ambiguous acquisitional intake. These biases include refining the structures considered informative, imposing default values, and ordering the acquisition of parameters.

3.1.2 Variational learning

The *variational learning* (VarLearn) approach (Yang, 2002, 2004, Legate and Yang, 2007, Yang, 2012) sidesteps explicit consideration of ambiguity and relies on probabilistic search of the parametric hypothesis space. More specifically, a VarLearner will have some number of parameters, with different possible values, and each of those values will be assigned some initial probability. (This is typically a uniform probability, so that if there are two options like +ItoC and -ItoC, each will start with probability 0.5.) When encountering a data point, the VarLearner probabilistically samples a complete grammar of parameter values, based on the probability of those values. So, in a parametric hypothesis space with the three parameters *head-first/head-last*, *+ItoC/-ItoC*, and *?Top/+Top*, a VarLearner pulls out a single grammar, such as [*head-first +ItoC,?Top*] with probability p(*head-first*)*p(*+ItoC*)*p(*?Top*). Whichever grammar is sampled, the VarLearner then sees if the grammar can account for the data point. For example, the grammar above would be able to account for the utterance *I saw penguins* (parse 4 from (9)), but not the utterance *I penguins saw*. If the grammar can account for the data point, all the participating parameter values are rewarded; if not, all parameter values are punished.

The VarLearner uses a reinforcement learning inference approach, and as mentioned in section 2.3.2, this means that unambiguous data are very impactful as they are only compatible with one parameter value. For example, data can be unambiguous for the *head-first* value, which means they will always reward grammars with *head-first* and always punish grammars with *head-last*. Over time, this means, the parameter value with more unambiguous data in the intake (an unambiguous data advantage) will succeed. As mentioned in section 2.3.2, this is why VarLearner approaches typically do an analysis of the acquisitional intake, focusing on the unambiguous data perceived by the VarLearner to be available. In particular, these approaches focus on the unambiguous advantage one parameter value has over another. The higher the unambiguous data advantage for a parameter value, the faster a child using the VarLearner strategy should converge on that parameter value. This means that age of acquisition predictions can be made from careful analysis of the acquisitional intake. Specifically, parameter values that have higher unambiguous data advantages are learned earlier.

With this in mind, let's consider the initial state, data intake, inference, learning period, and target state of VarLearners. The initial state consists of both prior knowledge and abilities. First, the VarLearner must know the parameters and their values. It must also be able to probabilistically sample a grammar to parse with, based on the current probabilities associated with the parameter values, and then parse data using that grammar. This allows the VarLearner to recognize that a data point is compatible or incompatible with this grammar. The VarLearner must also have the ability to do probabilistic inference (here, reinforcement learning) to update parameter value probabilities.

The VarLearner's input is typically utterances of the target language, encountered with the same frequency that they're encountered in child-directed speech. The perceptual intake of an utterance is the representation that results from parsing with the probabilistically sampled grammar. This then yields the acquisitional intake: either successful parsing (and a signal to reward all participating parameter values) or unsuccessful parsing (and a signal to punish all participating parameter values). Inference is an algorithmic-level probabilistic update using reinforcement learning, based on the acquisitional intake. The learning period is measured in the number of utterances encountered before reaching the target state, which is the set of correct parameter values for the language.

Again, as mentioned, because the data perceived by the VarLearner as unambiguous for a parameter value are massively impactful on the acquisition outcome (in fact determining it), VarLearner studies typically focus on the nature of the acquisitional intake and what unambiguous data advantages emerge. All the VarLearner studies reviewed below are of this kind.

Yang (2002, 2004, 2012) uses a VarLearner approach to successfully predict the relative order of acquisition of certain parameter values in different languages, based on the unambiguous data advantage a VarLearner would perceive for each. Table 1 demonstrates the striking qualitative fit between the perceived unambiguous data advantage in child-directed speech and the age of acquisition for six different parameter values. This suggests that both the VarLearner representations of the unambiguous data and the VarLearner learning mechanism may be on the right track for understanding how children identify the correct values for parameters in their language.

Table 1: The qualitative fit Yang discovered between the unambiguous data advantage (Adv) perceived by a VarLearner in its acquisitional intake and the observed age of acquisition (AoA) in children for six parameter values across different languages.

Param Value	**Language**	**Unambiguous Form**	**Unambiguous Ex**	**Adv**	**AoA**
+*wh*-fronting	English	*wh*-fronting in questions	*Who did you see?*	25%	<1;8
+topic-drop	Chinese	null objects	*Wǒ méi kànjiàn* *I not see* "I didn't see (him)"	12%	<1;8
+pro-drop	Italian	null subjects in questions	*Chi hai visto* *who have seen* "Who have you seen?"	10%	<1;8
+verb-raising	French	*Verb Adverb*	*Jean voit souvent Marie* *Jean sees often Marie* "Jean often sees Marie"	7%	1;8
-pro-drop	English	expletive subjects	There's a penguin on the ice.	1.2%	3;0
+verb-second	German Dutch	*Object Verb Subject*	*Pinguine liebe ich.* *penguins like I* "I like penguins"	1.2%	3;0-3;2
-scope-marking	English	long-distance *wh* questions without medial-*wh*	*Who do you think is on the ice?*	0.2%	>4.0

Legate and Yang (2007) use a VarLearner approach to explain the different rates of *optional infinitives* (OIs) in child-produced speech, which are observed to vary by language. OIs, sometimes called *root infinitives*, are instances where children appear to use the non-finite verb form in main ("root") clauses, and specifically in places where adults can't. Some cross-linguistic examples

highlighted in Legate and Yang (2007) are shown in Table 2.

Table 2: Optional infinitive examples in child-produced speech in different languages, and their intended meaning.

English	**Dutch**	**French**	**German**	**Hebrew**
Papa have it.	*Thee drinken*	*Dormir petit bébé*	*Mein Kakao hinstelln*	*Lashevet al ha-shulxan*
Papa have$_{INF}$ *it*	*tea drink*$_{INF}$	*sleep*$_{INF}$ *little baby*	*my cocoa put*$_{INF}$	*sit*$_{INF}$ *on the-table*
"Papa has it"	"Drinks tea"	"Little baby sleeps"	"Puts my cocoa"	"Sits on the table"

Interestingly, children's frequency of OI use seems to vary by language, with some children using them very infrequently and tapering off OI use prior to age two (e.g., Spanish children), while other children still use OIs fairly frequently into age three and beyond (e.g., English children). Legate and Yang (2007) investigate a parametric account involving one key linguistic parameter, which they refer to as *+/-Tense*. As its name suggests, this parameter is sensitive to the linguistic expression of tense in a language: *+Tense* languages like Spanish express tense morphosyntactically (e.g., Spanish *duerme*=*sleep*$_{pres+3rd+sg}$ = "He/she/it sleeps"); *-Tense* languages like Mandarin Chinese don't, relying on other linguistic mechanisms to communicate tense (e.g., Mandarin Chinese *Zhangsan zai da qiu* = *Zhangsan ASPECT play ball* = "Zhangsan is playing ball."). So, this parametric account suggests that children using OIs may be using a *-Tense* grammar for an extended period of time, when their language is in fact a *+Tense* language. As the children encounter more unambiguous *+Tense* data in their acquisitional intake, the *+Tense* grammar is rewarded and the *-Tense* grammar generating the OIs is punished until it's no longer active. How fast this happens depends on how many unambiguous *+Tense* data are available.

From a VarLearner perspective, the *+Tense* option is rewarded (and *-Tense* punished) every time an utterance is encountered with overt tense morphology. So, the quantity of data with tense morphology in child-directed speech determines the *+Tense* unambiguous data advantage. This then determines how quickly children acquiring *+Tense* languages figure out that OIs are not allowed. Legate and Yang (2007) use the VarLearner approach to analyze the acquisitional intake of Spanish, French, and English children (who are all learning *+Tense* languages), and find a qualitative fit between the unambiguous data advantage and these children's production of OIs. More specifically, the unambiguous data advantage for *+Tense* in Spanish $>$ French $>$ English, while Spanish OI production $<$ French OI production $<$ English OI production. So, the greater the unambiguous data advantage for *+Tense* in a language's child-directed speech, the faster children acquiring that language stop using OIs.

3.2 Non-parametric approaches

There are also many syntactic phenomena whose acquisition process isn't represented with linguistic parameters. I review several studies of this kind below, beginning with a non-parametric account of children's cross-linguistic OI production. I'll then discuss some other phenomena relying only on syntactic input during the acquisition process, including the structure-dependence of

linguistic representations and constraints on *wh*-dependencies. I'll conclude with other studies that integrate syntactic acquisition with conceptual information. These include studies related to the interpretation of pronouns, which integrate the development of syntactic knowledge and referential knowledge, and studies of how children link conceptual information to its syntactic form more broadly (the linking problem, using semantic structure to infer syntactic structure).

3.2.1 Optional infinitives (OIs) again

Freudenthal and colleagues (Freudenthal et al., 2007, 2009, 2010, 2015) developed a non-parametric model of syntactic acquisition called MOSAIC (*Model of Syntactic Acquisition in Children*) which, like the parametric VarLearner, is able to account for the different cross-linguistic rates of OIs in children's speech. Instead of parameters, MOSAIC relies on an incremental probabilistic learning mechanism directly related to the child's perceptual intake. This mechanism keys into children's developing memory and parsing abilities.

More specifically, a MOSAIC learner has a strong utterance-final bias (in the spirit of a recency memory effect), which causes this learner to encode utterance-final phrases in child-directed input. As the MOSAIC learner's familiarity with the linguistic elements in utterances increases, utterances can be parsed and encoded more easily. This allows the encoded utterance-final phrases to become longer (e.g., instances of *Can Papa have it?* successively encoded as *it*, *have it*, and *Papa have it*).

The MOSAIC implementation of Freudenthal et al. (2015) also includes two additional biases. The first is a smaller utterance-initial bias, in the spirit of a primacy memory effect, which allows the learner to preferentially encode utterance-initial elements as well as utterance-final elements. The second is a bias to replace verb forms with what the child perceives to be the default form at that stage of development (e.g., *go* replacing *goes*), which can be viewed as a memory retrieval error during speech production. These two additional biases allow the MOSAIC learner to both (i) capture the rate of English OIs more precisely, while maintaining the quantitative fit to the other language OI production rates, and (ii) generate observed OI constructions that involve partial utterance pieces from both the beginning and end of utterances in child-directed speech (e.g., *Where he go?* coming from *Where did he go?*).

MOSAIC's algorithmic-level implementation highlights Freudenthal and colleagues' theoretical account of the observed OI production: OIs are really portions of longer utterances in children's input that include compound finite forms, i.e., a modal or auxiliary verb (like *can* or *did*) with a non-finite form (like *have* or *go*) for utterances like *Can Papa have it?* and *Where did he go?* Notably, this account of the linguistic representation development relies very much on other developing abilities. Representations of utterance pieces (rather than the whole utterance) result from children's developing memory and parsing abilities imperfectly encoding the input into the child's perceptual intake. The output from a MOSAIC learner is a fairly direct reflection of this underlying linguistic representation in the perceptual intake, potentially including memory retrieval errors that occur during speech production.

This approach, relying on simple implementations of children's memory and parsing limitations, captures additional observed behavior about the nature of OIs in child-produced speech. For example, Freudenthal et al. (2009) demonstrate that a MOSAIC learner captures the "Modal Refer-

ence Effect" in Dutch and German (and its absence in English), where Dutch and German children tend to use OIs more frequently in modal contexts while English children don't. For the MOSAIC learner, this is due to the frequent use of compound finite forms with modals in Dutch and German child-directed speech, and their relative scarcity in English child-directed speech. This MOSAIC learner also captures the presence of the "Eventivity Constraint" in Dutch and German and its reduced effect in English, where Dutch and German children tend to use OIs predominantly with eventive verbs like *go*, while English children do so less often. As with the Modal Reference Effect, this behavior comes directly from the composition of Dutch, German, and English child-directed speech: The vast majority of Dutch and German compound finite examples are used with eventive verbs, while fewer eventive uses occur in English.

As Freudenthal et al. (2010) note, both a VarLearner and a MOSAIC learner can capture the cross-linguistic patterning of OI errors reasonably well. Because a MOSAIC Learner has additional explanatory coverage about the specific verbs that appear as OI errors, we might prefer its account of OI development for now.

3.2.2 Structure dependence

The nature of linguistic representations is generally agreed to be hierarchical, with units inside other units (e.g., phrases inside other phrases). Moreover, rules for manipulating linguistic elements rely on this structure, rather than operating over simpler representations such as linear word order. A traditional example of this is complex yes/no question formation in English, where we can think about a yes/no question like (10a) as being a *structure-dependent* transformation of the declarative utterance in (10b) which involves the same core contentful elements.

(10) a. [$_{CP}$ Can the penguin [$_{CP}$ who is on the iceberg] t_{can} find a fish]?
b. [$_{CP}$ The penguin [$_{CP}$ who is on the iceberg] can find a fish].

Why do we call this structure-dependent? One common way to describe the transformation in (10) is with the rule "Move the auxiliary in the main clause to a position at the front of the utterance". This rule relies on the structural notion of "main clause". Importantly, there are other potential rules that *aren't* structure-based (i.e., they're *structure-independent*), such as "Move the last auxiliary". Structure-independent rules aren't able to account for the full range of English complex yes/no questions, though some (like the one mentioned above) are often compatible with most yes/no questions in child-directed speech.

From a developmental standpoint, children seem to know that syntactic rules like those controlling complex yes/no question formation in English ought to be structured-dependent: The classic study by Crain and Nakayama (1987) demonstrates this knowledge in children as young as three years old. Interestingly, from an input standpoint, it turns out that English children are likely to encounter very few examples of complex yes/no questions that would unambiguously indicate that a structure-dependent rule is required (Pullum and Scholz, 2002, Legate and Yang, 2002). So, English children's seemingly rapid development of this structure-dependent knowledge is an intriguing mystery.

Importantly, there are really two pieces to children's structure-dependent knowledge: (i) linguistic representations are hierarchically structured, and (ii) rules manipulating linguistic elements

utilize this structure. One hypothesis about where these two knowledge pieces come from is that children have an innate bias to use hierarchical structure, both in how they represent language and how they hypothesize rules that manipulate linguistic elements. In the simplest form, this bias has often been described as children innately knowing to only consider structure-dependent representations and to only hypothesize structure-dependent rules (Chomsky, 1971, Berwick, Pietroski, Yankama and Chomsky, 2011).

A competing hypothesis is that children can converge on this knowledge – in particular, the nature of the correct rules (ii above) – without an innate bias for structure-dependence. Reali and Christiansen (2005) demonstrate that a probabilistic learning model sensitive only to bigram and trigram word frequencies in English child-directed speech seems to distinguish between grammatical and ungrammatical forms of complex yes/no questions. That is, without a bias for structure-dependent rules (and in fact without relying on structure-dependent representations at all), this algorithmic-level model can distinguish between grammatical yes/no questions like *Is the boy who is watching Mickey Mouse happy?* and ungrammatical variants like **Is the boy who watching Mickey Mouse is happy?*

However, Kam, Stoyneshka, Tornyova, Fodor and Sakas (2008) demonstrated that the modeled learner of Reali and Christiansen (2005) benefited from a "lucky fluke" in the particular corpus used as input and the particular sentences chosen as test sentences. More specifically, bigram information irrelevant to complex yes/no question formation in general was used to distinguish the tested complex yes/no questions. When tested on a wider range of complex yes/no questions in English (such as discriminating the grammatical *Is the wagon your sister is pushing red?* from the ungrammatical **Is the wagon your sister pushing is red?*), this modeled learner did not succeed. So, this potential structure-independent learning strategy does not in fact work (at least as implemented).

A very interesting variant of the innate bias hypothesis that focuses on the structure-dependent nature of linguistic representations was explored by Perfors et al. (2011). They considered that children might not need to innately restrict their hypotheses to structure-dependent representations. Rather, children's innate bias might be to *allow* structure-dependent representations into the hypothesis space in the first place – not to disallow other competing hypotheses from also being there. That is, children's innate bias is about considering structure-dependent representations as one of the potential representations at all. This is a less specific innate bias than one which requires children to *only* allow structure-dependent representations in their hypothesis space.

Using a computational-level hierarchical Bayesian model[5], Perfors et al. (2011) showed that a modeled learner can leverage English child-directed speech input (abstracted into sequences of syntactic categories like DETERMINER NOUN VERB DETERMINER NOUN for *The penguin ate a fish*) to determine that hierarchically structured linguistic representations are preferable to other possible linguistic representations. The key is that evidence in favor of structure-dependent representations comes not just from the examples of complex yes/no questions, but also from many other kinds of utterances. So, other utterances favoring structure-dependent representations serve as indirect positive evidence in favor of structure-dependent representations when representing

[5]Note that "hierarchical" for Bayesian models refers to having overhypotheses as discussed in section 2.3.4, not to having hierarchical structure in the hypothesized linguistic representations.

complex yes/no questions. Through the hierarchical Bayesian learner's use of overhypotheses, it was able to leverage this evidence.

Moreover, using the specific structure-dependent representations the modeled learner converged on, it would be able to parse certain grammatical complex yes/no questions and be unable to parse other ungrammatical ones. Perfors et al. (2011) interpreted this to mean that the modeled learner would have the correct judgments about which complex yes/no questions were grammatical. This is based on the intuitive idea that if the learner can't parse the utterances with its current grammar, these utterances aren't grammatical for it. So, this model demonstrates that it's *in principle* possible to have innate knowledge that is more general about the structure-dependence of linguistic representations, yet which still allows acquisition to proceed successfully from the data English children typically encounter.

Why only in principle? Remember that computational-level models (like this one) aren't necessarily using inference procedures we think humans use – rather, the point is to discover if the conceptualization of the acquisition task and the learning assumptions involved lead to successful acquisition. This hierarchical Bayesian model demonstrates that these assumptions about structure-dependence will in fact do that if children are capable of doing optimal inference to identify the optimal representational hypothesis. However, this type of model abstracts away from some of the cognitive limitations children have – such as memory and processing limitations – which potentially impact children's ability to do optimal inference. It remains for future research to determine if the learning assumptions used in this model still allow successful acquisition of structure-dependent representations to occur even when there are cognitive limitations on the inference process.

3.2.3 Constraints on *wh*-dependencies: Syntactic islands

One hallmark of the syntax of human languages is the ability to have long-distance dependencies: relationships between two words in a sentence that are not adjacent to each other. Long-distance dependencies, such as the dependency between *what* and *stole* in (11a), can be potentially infinite in length.[6] However, there are specific syntactic structures that long-distance dependencies can't cross. These structures are known as *syntactic islands*. Four examples of syntactic islands are shown in (11b)-(11e) (Chomsky, 1965, Ross, 1967, Chomsky, 1973). During acquisition, children must infer the constraints on long-distance dependencies that allow them to recognize that the *wh*-dependencies in (11b)-(11e) are not allowed, while the *wh*-dependency in (11a) is fine.

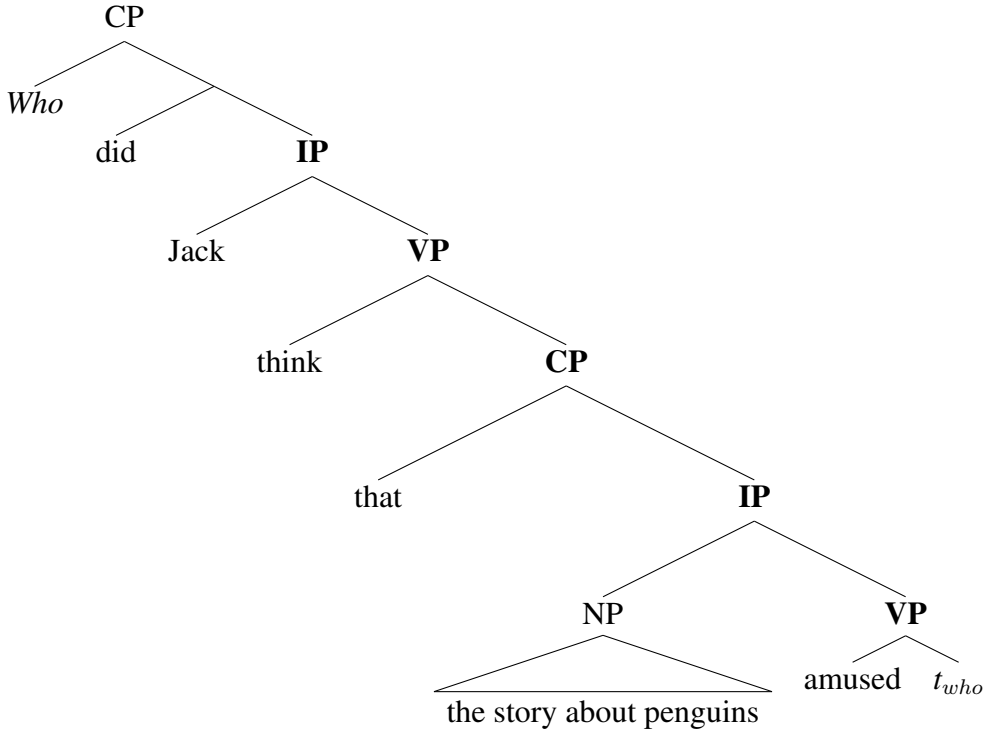
(11) a. What does Jack think that Lily said that the goblins stole t_{what}?
b. *What do you wonder [whether Jack bought t_{what}]? (whether island)
c. *What did you make [the claim that Jack bought t_{what}]? (complex NP island)
d. *What do you think [the joke about t_{what}] was hilarious? (subject island)
e. *What do you worry [if Jack buys t_{what}]? (adjunct island)

[6] Of course, there's clearly an upper bound on the number of words and/or clauses that an English speaker can keep track of during language processing. However, this restriction appears to be based on the limited nature of human working memory capacity rather than an explicit structural restriction on the length of English *wh*-dependencies.

Pearl and Sprouse (2013a,b, 2015) investigated an algorithmic-level model for learning these constraints on long-distance dependencies. This probabilistic learning model relies on the idea that children can characterize a long-distance dependency as a path from the head of the dependency (e.g., *Who* in (12)) through the phrasal nodes that contain the tail of the dependency, as shown in (12a)-(12b). Under this view, children simply need to learn which long-distance dependencies have licit syntactic paths and which don't. The probabilistic learning algorithm of Pearl & Sprouse tracks local pieces of these syntactic paths. More specifically, it breaks the syntactic path into a collection of syntactic trigrams that can be combined to reproduce the original syntactic path, as shown in (12c).

(12) Who did Jack think that the story about penguins amused t_{who}?

a. Phrasal node structure containing the *wh*-dependency headed by *Who*:



Who did [$_{IP}$ Jack [$_{VP}$ think [$_{CP}$ that [$_{IP}$ the story about penguins [$_{VP}$ amused t_{who}]]]]]?

b. Syntactic path of *wh*-dependency:
start-IP-VP-CP$_{that}$-IP-VP-*end*

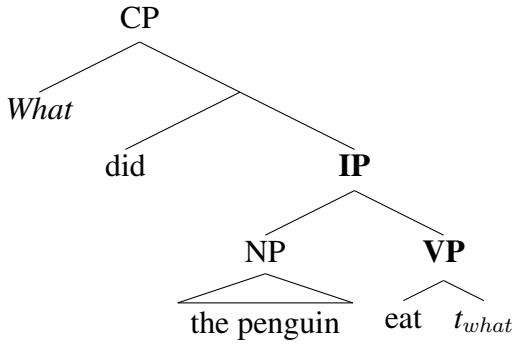
c. Syntactic trigrams $\in Trigrams_{start-IP-VP-CP_{that}-IP-VP-end}$:
= *start*-IP-VP
IP-VP-CP$_{that}$
VP-CP$_{that}$-IP
CP$_{that}$-IP-VP
IP-VP-*end*

The learning strategy implemented in the model tracks the frequencies of these syntactic tri-

grams in children's acquisitional intake, which come from every instance of a *wh*-dependency that the modeled learner perceives. The learner later uses these frequencies to calculate probabilities for all syntactic trigrams comprising a *wh*-dependency. This allows the modeled learner to generate the probability of any *wh*-dependency because any *wh*-dependency can be broken into a subset of these same syntactic trigrams. For example, the *wh*-dependency in *What did the penguin eat* t_{what}*?* can be characterized as in (13), and its probability generated from some of the same syntactic trigrams observed in (12).

(13) What did the penguin eat t_{eat}?

a. Phrasal node structures containing the *wh*-dependency:



What did [$_{IP}$ the penguin [$_{VP}$ eat t_{what}]]?

b. Syntactic path of *wh*-dependency:
start-IP-VP-*end*

c. Syntactic trigrams $\in Trigrams_{start-IP-VP-end}$:
= *start*-IP-VP
IP-VP-*end*

The generated probability corresponds to whether that dependency is allowed, with higher probabilities indicating grammatical dependencies and lower probabilities indicating ungrammatical dependencies. Pearl and Sprouse (2013a,b, 2015) let the modeled learner's input be a quantitatively and qualitatively realistic sample of English child-directed speech. With this input encountered incrementally, the modeled learner estimated syntactic trigram probabilities and could then generate probabilities for any desired *wh*-dependency.

Pearl & Sprouse mapped the learner's generated probability for a *wh*-dependency to its assessment of how acceptable a *wh*-dependency is. In particular, Pearl & Sprouse drew on acceptability judgment data from Sprouse, Wagers and Phillips (2012) about the four syntactic islands in (11b)-(11e) to provide a behavioral target for the model's output. The modeled learner was able to replicate the observed acceptability response pattern that indicated knowledge of these syntactic islands. So, Pearl & Sprouse interpreted this to mean that this learning strategy relying on the syntactic trigrams of *wh*-dependencies was a reasonable way for English children to acquire knowledge of these islands.

Interestingly, this is another example of less-specific prior knowledge being sufficient to explain acquisition of certain linguistic phenomena. In the Government and Binding framework of

the 1980s, syntacticians proposed a constraint called the Subjacency Condition to explain the acquisition of syntactic islands. This constraint basically says that dependencies can't cross two or more *bounding nodes* (Chomsky 1973, Huang 1982, Lasnik and Saito 1984, among others); if a dependency crosses two or more bounding nodes, a syntactic island effect occurs. What counts as a bounding node varies cross-linguistically, though bounding nodes are always drawn from the set {NP, IP, CP}. So, when using this representation, children needed to learn which of these are bounding nodes in their language, though they already know (via UG) about the restriction the Subjacency Condition imposes and the set of possible bounding nodes. This was fairly specific prior knowledge.

The strategy of Pearl and Sprouse (2013a,b, 2015) shares the intuition with Subjacency that there's a local structural anomaly when syntactic islands occur. However, instead of characterizing this anomaly with bounding nodes, Pearl & Sprouse suggested that it could be described as a low probability region with respect to the phrase structure nodes that characterize the syntactic path of *wh*-dependency. The learner recognizes this low probability region via low probability syntactic trigrams. There was no need for the learner to know about a hard constraint prohibiting the crossing of two bounding nodes, and in fact no need for the learner to identify bounding nodes at all. Instead, children would need to identify the phrase structure nodes containing the *wh*-dependency and break that syntactic path into syntactic trigrams – certainly no trivial matter and perhaps requiring innate, linguistic knowledge as well. Importantly however, the knowledge of syntactic trigrams and how to use them doesn't seem to be as specific to learning islands as the knowledge about bounding nodes and how to use them. Syntactic trigrams may be useful for other syntactic phenomena, and probabilistic learning is useful for both linguistic and non-linguistic phenomena.

So, the investigations of Pearl & Sprouse offer an alternative for both representing syntactic islands knowledge and acquiring this knowledge from the available input. On the representation side, syntactic trigrams comprised of simple phrase structure nodes will do instead of bounding nodes, and no hard-coded constraint about which abstract units a *wh*-dependency can cross is required. On the acquisition side, Pearl & Sprouse provide a concrete account of how English children would leverage the proposed syntactic trigram representation to acquire a constraint about the units a *wh*-dependency can cross. Importantly, this constraint derives directly from a simple dislike of low probability items, rather than being specific to the acquisition of syntactic islands.

3.2.4 Incorporating reference and meaning

The first three examples I'll discuss below involve the acquisition of pronouns. Interpreting a pronoun in context typically involves both syntactic knowledge and referential knowledge, and so pronoun acquisition also requires the integration of different knowledge types. The last two examples involve incorporating meaning more broadly into syntactic acquisition, with the idea that conceptual and syntactic information are intimately linked for certain fundamental aspects of linguistic development.

3.2.4.1 English anaphoric *one*. The first example is about learning to interpret English anaphoric *one* (Regier and Gahl, 2004, Foraker et al., 2009, Pearl and Lidz, 2009, Pearl and Mis, 2011, Payne, Pullum, Scholz and Berlage, 2013, Pearl and Mis, 2016). Suppose the following scenario occurs:

You see a red bottle, and your friend comments, "*Oh look – a red bottle! Do you see another one around anywhere?*"

What kind of bottle is your friend looking for? The use of *one* in your friend's utterance is anaphoric – that is, it's the pronoun use of *one*. To understand this utterance and know what bottle to look for, you have to interpret the pronoun: Is the *one* your friend referring to a second red bottle or just a second bottle of whatever color? Both interpretations are generally available to adults, though there may be some preference for referring to a second red bottle in this context. How does this work?

The interpretation has to do with the antecedent of *one* – that is, what linguistic string it stands in for. The phrase "*another one*" can be interpreted in two ways: "*another red bottle*" or "*another bottle*", given the previous utterance's use of "*a red bottle*". So, the antecedent of *one* can be either *red bottle* or *bottle*. When the antecedent is *red bottle*, *one* refers only to a red bottle; when the antecedent is *bottle*, *one* can refer to a bottle of any color.

This highlights one way we think pronoun interpretation operates: We look for the antecedent of the pronoun, and then look for a referent compatible with that antecedent. Importantly, one prominent approach to how we determine a pronoun's antecedent is that the antecedent depends on the syntactic category of the pronoun – that is, pronouns take antecedents of the same category. So, the reason *one* can take either *red bottle* or *bottle* as its antecedent is because *one* is a category that includes both these strings. One name for this category is N′ (Chomsky, 1970, Jackendoff, 1977), and it includes both modifier+noun strings like *red bottle* and noun-only strings like *bottle*.[7]

With respect to syntactic category, there are plenty of other options for what a pronoun's category might be – perhaps it's a simple Noun category, or a phrasal category like NP. These categories allow a different range of potential antecedents. If *one* were a Noun, its antecedents could only be nouns – this would make interpreting "*another one*" as "*another red bottle*" impossible because "*red bottle*" is more than just a noun. But of course, that interpretation is perfectly possible, so this rules out category Noun. If *one* were category NP in this context, its antecedent would be *a red bottle* and we would expect the phrase "*another one*" to be disallowed – after all, we can't say "*another a red bottle*". But again, this is perfectly possible, so this rules out *one* being category NP in this context.

The acquisition question for interpreting English anaphoric *one* therefore concerns what syntactic category *one* is – this is a syntactic acquisition task. However, pronoun interpretation also concerns how to choose between potential antecedents in context to determine the correct referent. That is, in the example scenario from before, what kind of bottle are we actually referring to when we say "*another one*" – another red bottle or just another bottle? Both *red bottle* and *bottle* are potential antecedents because they're included in category N′. So, to interpret *one* correctly, we also need intuitions about the intended referent – this highlights that learning about English anaphoric *one* is an acquisition task about referential knowledge as well.

Courtesy of experiments with 18-month-olds by Lidz, Waxman and Freedman (2003) (LWF), we have some insight about how toddlers interpret English anaphoric *one* in this kind of context. More specifically, LWF examined the looking behavior of 18-month-olds to gauge how these tod-

[7]Note that this category may have other names, depending on the particular syntactic theory being used – the important thing for our purposes is that the category includes both modifier+noun strings and noun-only strings.

dlers interpret anaphoric *one* in scenarios fairly similar to the one described above. LWF found that children of this age behave as if they interpret both "*Do you see another one?*" and "*Do you see another red bottle?*" the exact same way (preferring to look at a red bottle). This notably contrasts with their behavior when asked "*Do you see another bottle?*", where they preferred to look at a non-red bottle. LWF interpreted this to mean that 18-month-olds select *one*'s antecedent to be the N′ string *red bottle*. This suggests that 18-month-olds recognize *one* as category N′ and prefer its antecedent to be the modifier+noun N′ string in this context.

Several computational modeling studies investigated how English children could converge on this pronoun interpretation knowledge – either the syntactic knowledge alone (Foraker et al. 2009), or both the syntactic and referential knowledge (Regier and Gahl 2004, Pearl and Lidz 2009, Pearl and Mis 2011). This meant assessing modeled learners on whether they achieved the target knowledge state. Building on these studies, Pearl and Mis (2016) decided to leverage the available behavioral data from LWF. In the modeling terms from before, Pearl and Mis (2016) assessed whether their algorithmic-level Bayesian model could achieve the appropriate target behavior observed in the 18-month-olds. When their modeled learner was capable of reproducing the 18-month-old looking behavior, Pearl and Mis could then examine the underlying knowledge state generating that behavior.

Pearl and Mis found two learning strategies capable of generating the 18-month-old target behavior when given a quantitatively and qualitatively realistic sample of English child-directed speech as input. The first strategy restricted the learner's acquisitional intake to a subset of of data including only certain utterances using anaphoric *one*. This strategy generated the target behavior by using syntactic and referential knowledge similar to adult knowledge, though it was less robust to different learning scenarios. More specifically, it required certain additional properties to be true about the learning environment in order to achieve the target behavior.

In contrast, the second strategy did not require these additional restrictions on the learning scenario – it succeeded in producing the target behavior no matter what. Its approach was to expand the child's acquisitional intake to include utterances using any anaphoric pronoun (not just those using anaphoric *one*). The indirect positive evidence coming from these additional anaphoric data points stabilized acquisition of the target behavior. Interestingly, however, the target behavior was the result of an immature context-dependent representation of anaphoric *one*. Specifically, a learner using this strategy only assumed *one*'s syntactic category was N′ in specific contexts where the potential antecedent mentioned a property (such as the experimental setup of LWF, where the previous utterance was "*Look - a red bottle!*"); in other contexts with no property mentioned (e.g., "*Look – a bottle!*"), *one*'s category would be N.

This is a very subtle thing – a child using the second strategy would have a knowledge state that looks quite adult-like in many contexts, but would be revealed as non-adult-like in a few others. This contrasts with a child using the first strategy, who would always have an adult-like knowledge state no matter the context. Importantly, we don't yet know how young children behave in the contexts that would reveal whether they *always* have adult-like knowledge of *one*'s syntactic category. This remains for future experimental work. Still, by doing careful computational modeling work of this kind, we now have much more concrete ideas for the specific contexts that need investigating. We also have concrete ideas for the learning strategies that enable English 18-month-olds to

acquire the anaphoric *one* knowledge that they do.

3.2.4.2 Anaphora resolution. When interpreting pronouns, the type of pronoun also places restrictions on its potential antecedents. Plain pronouns like *her* behave differently from reflexives like *herself*, as we can see in (14) – these pronouns take complementary antecedents.

(14) a. Lily, who adores Sarah, admired her in the mirror. (*her* $\neq$ *Lily*, and probably = *Sarah*)
b. Lily, who adores Sarah, admired herself in the mirror. (*herself* = *Lily*, and $\neq$ *Sarah*)

Each pronoun type has certain restrictions on where its potential antecedents can be located in the utterance, and these are typically described structurally (e.g., using the notion of c-command: Chomsky 1973). So, pronoun interpretation (also sometimes called *anaphora resolution*) is another example of linguistic knowledge involving structure-dependent rules. Frank, Mathis and Badecker (2013b) investigated whether a certain type of probabilistic learning strategy, as implemented by simple recurrent networks (SRNs) (Elman, 1993, Lewis and Elman, 2001), is capable of inferring the appropriate structure-dependent rules of anaphora resolution for plain and reflexive pronouns.

SRNs are a type of probabilistic learning model where the learner's representations are distributed across a number of units, rather than being explicitly available and manipulable. This is a bit easier to understand with a concrete example, so let's consider a specific instance of anaphora resolution and how it might be implemented in a Bayesian model versus an SRN. To interpret *herself* in (14b), a Bayesian model might consider a hypothesis such as "*herself* refers to *Lily*" and an overhypothesis such as "*A reflexive must have a c-commanding NP within its own clause as the antecedent*".[8] This hypothesis and overhypothesis are explicit "units" within the Bayesian learner's reasoning process (for example, the modeled learner can explicitly assign probabilities to the hypothesis and overhypothesis).

In contrast, the SRN in Frank et al. (2013b) has 28 units to encode the current word (such as *herself*) and 28 units to predict the referent of that word, in addition to numerous hidden units and context units that recode the current word and its context in a distributed fashion. This underscores how the meaning of a single word isn't encapsulated anywhere. It's also not obvious where the "units" of the hypothesis or overhypothesis that the Bayesian learner uses are. This is because an SRN approach represents a very different idea of what prior knowledge is, how knowledge can be encoded, and how knowledge can be used to generate the observable linguistic behavior (such as that of anaphora resolution). As Frank et al. (2013b) note, SRNs have typically been used to investigate whether structure-dependent knowledge can be learned from a probabilistic learning approach relying on distributed representations that don't explicitly include any notion of structure. All the prior "knowledge" of the SRN is in the modeled architecture connecting the distributed representations to each other (e.g., that there are 28 input units that connect to 10 hidden units). Learning involves probabilistically tuning the connection weights between the units of the distributed representations, based on the input received.

As you may imagine from the preceding description, this can make it difficult to decode what exactly is causing an SRN to generate the behavior it does. For example, if the output units map

[8] And perhaps even an over-over-hypothesis of "*Rules should be structure-dependent*" for generating that specific overhypothesis, which generates that specific hypothesis.

to *Lily* when the input units map to *herself* in a particular utterance, is it because the SRN has a distributed representation of the appropriate structure-dependent rule embedded in it? Or is it because the SRN's internalized some other mapping rule that generates the appropriate output in this case? Frank et al. (2013b), in my opinion, do a particularly fine job of uncovering the inner workings of their modeled SRN with clever analysis of the distributed representation patterns, specific test cases, and targeted manipulations of the SRN's architecture. In order to have the explanatory power we want from computational models, this is especially important to do for less-transparent computational models like SRNs.

Because of this effort, Frank et al. (2013b) were able to uncover some striking findings about the acquisition of anaphora resolution using SRNs of this kind. The SRN that they investigated represented a computational-level learner as it focused on the impact of the SRN's learning assumptions rather than empirical details of children's learning periods. This modeled learner was attempting to determine the optimal representation given its distributed architecture and a selection of potentially informative English sentences containing embedded clauses, reflexives, and pronouns. After learning from these input data, the SRN was fairly good at identifying the appropriate referent for both reflexives and pronouns in the test cases – for instance, the SRN-preferred referent for reflexives is the correct one around 90% of the time, and SRN-preferred referents for plain pronouns are in fact allowed by humans around 72% of the time.

However, as Frank et al. (2013b) note, the way the SRN accomplishes anaphora resolution "diverges in key respects" from the way we think humans do. As one example, when a noun is linearly adjacent to the main verb (such as *Sarah* in "*Lily, who adores Sarah, admires herself*"), the SRN pays special attention to that noun. That is, the SRN is relying on a structure-independent rule for accomplishing anaphora resolution. This contrasts with humans, who don't appear to have that kind of bias for anaphor resolution – it doesn't matter that *Sarah* is close to *admires* because these words aren't in the same clause.

Another example of this same issue – that is, paying attention to details that don't matter for humans – has to do with sentence types. The SRN processes anaphor resolution differently when the intervening embedded clauses differ in structure. For instance, it treats the resolution of *herself* differently in all three of these utterances: "*Lily admires herself*", "*Lily, who Sarah adores, admires herself*", "*Lily, who adores Sarah, admires herself*". This contrasts with how we believe humans interpret *herself*, as only the elements of the main clause are salient to humans.

Frank et al. (2013b) note that the key is "how the network chooses, during learning, to represent the context" and "whether this representation collapses contexts that are relevantly similar". This seems to be the sticking point for the SRN they evaluated – it lacks the prior knowledge necessary to filter out irrelevant information from its acquisitional intake. This leads to it hypothesizing "spurious linear generalizations" that human children seem able to avoid. That is, the SRN goes awry because it lacks prior knowledge about structure that would allow it to filter its input in a useful way. So, we might reasonably interpret this study as a demonstration that the absence of structure-dependent knowledge can cause children to perceive their input in an unhelpful way when learning about anaphora resolution in English.

3.2.4.3 Pronoun classes. As a nice conceptual companion to Frank et al. (2013b)'s study, Orita, McKeown, Feldman, Lidz and Boyd-Graber (2013) consider the acquisition of anaphora resolution from a structurally-based standpoint. Their starting premise is that children are biased to assume that structurally-defined properties are relevant for anaphora resolution. So, when perceiving their data, children automatically focus on structurally-defined information, such as what structural location a pronoun's antecedent appears at. The acquisition task is then about identifying classes of pronouns whose antecedents seem to appear in different distributions. That is, if children already know to be on the lookout for structurally-defined antecedent positions, how many classes of pronouns are there and which words belong to which classes? Once children know this, they can then interpret any pronoun they encounter – they simply identify which pronoun category the pronoun belongs to, and consider the antecedents allowed by that pronoun class.

As a concrete example, let's look again at *Lily, who adores Sarah, admires herself in the mirror.* A child using this learning strategy would recognize that *herself* is a specific pronoun class (which we adults term reflexives) whose antecedents appear in certain structurally-defined positions. The only potential antecedent in one of these positions is *Lily*, and so *Lily* must be the referent of *herself*.

Interestingly, while this learning approach seems to solve some filtering problems related to the child's acquisitional intake, Orita et al. (2013) note that there's still a knotty learning problem remaining. The problem is one that highlights the interdependence of acquisition and processing: Children must sort pronouns into classes based on where their antecedents appear (acquisition), but children need to know how to interpret a pronoun in order to know where its antecedent appears (processing). That is, when given "*Lily, who adores Sarah, admires* PRONOUN *in the mirror*", if the child doesn't already know what class PRONOUN belongs to, how can she know whether the antecedent is *Lily* or *Sarah*? If PRONOUN is a reflexive (like *herself*), *Lily* is the right choice; if PRONOUN is a plain pronoun (like *her*), *Lily* can't possibly be the right choice.

Orita et al. (2013)'s insight was that there are other cues to the intended antecedent for a pronoun in context. That is, even if the child has no idea what the structural constraints are for a pronoun, she may be able to guess the intended antecedent from the context available. For example, if the utterance "*Lily, who adores Sarah, admires herself in the mirror*" is used to describe a scenario where Lily is clearly looking at her own reflection, a child may be able to guess that *Lily* is the antecedent of *herself*. As another example, if that same utterance were embedded in a conversation that made it clear Lily tends to do a lot of gazing at her own reflection, the discourse context could also allow a child to infer that *herself* refers to *Lily*. Based on these guesses about pronoun antecedents, the child can then get a sense of what the structural distribution of antecedents is for any given pronoun, and get the acquisition process started for pronoun classes.

Orita et al. (2013) focused on discourse cues, and estimated how guessable a pronoun's referent was by taking snippets of recorded conversations between adults and children from the North American English section of the CHILDES database. An utterance containing one of three NP categories (reflexive pronouns like *herself*, plain pronouns like *her*, or lexical names like *Lily*) had that NP deleted and replaced with a blank, with 12 utterances before and 6 utterances after provided for conversational context. Adult participants were asked to guess what went in the blank, which provided an estimate of how informative the naturalistic discourse context was for

guessing the identity of the missing NP. Orita et al. (2013) used this as a proxy for how informative the discourse context would be for a child.

More specifically, adults of course already know the structural rules and discourse cues that determine the antecedent for pronouns of different kinds. The idea was that if an adult can correctly guess the identity of the missing NP, the discourse information containing the potential antecedents must be informative enough to do so. That is, if an adult is given "*Lily, who adores Sarah, admires* _", and can correctly determine that the missing word is *herself*, the discourse context must be what's driving that guess. Since the adult already knows *herself* must refer to *Lily*, the discourse context is indicating that the missing word must refer to *Lily* too. This is also what a child would be able to infer from the discourse context. This means that a child given this very same conversational snippet – but who doesn't yet know how to interpret *herself* – would be able to infer that *herself* refers to *Lily*. This is the key information for learning about the distribution of antecedents for *herself*.

The estimate of discourse informativity turned out to differ somewhat by pronoun class – reflexives were harder to guess from discourse context than plain pronouns (68% correct vs. 81%), though both were far more guessable than chance. Notably, the discourse context seemed to distinguish fairly well between reflexive pronouns and plain pronouns – rarely was a plain pronoun guessed where a reflexive actually occurred, and vice versa. This supports the idea that the discourse cues available in naturalistic child-directed dialogues are informative for bootstrapping the acquisition of pronoun classes.

To demonstrate this, Orita et al. (2013) implemented several variants of a computational-level hierarchical Bayesian model for learning pronoun classes. The input to the model was a subset of the same dialogues used to derive estimates of discourse informativity, with the modeled learner's acquisitional intake consisting of three pieces of information: (i) the pronoun used in the dialogue, (ii) a distribution over possible antecedents for that pronoun, either harnessing discourse informativity or not, and (iii) the structural position of the possible antecedents for that pronoun. Using this approach, Orita et al. (2013) discovered that a learner harnessing discourse informativity learns exactly the right pronoun classes and their distributions over antecedent positions. This allows the learner to correctly interpret both reflexive and plain pronouns in context.

Moreover, without that discourse information, pronoun class acquisition is quite poor. Interestingly, learning how to interpret particular pronouns is just as poor even if the modeled learner *already* knows there are two classes of pronouns (reflexive vs. plain) and also knows exactly the right rules about antecedent distribution for each class. That is, the acquisition process for anaphora resolution *still* can't get off the ground without some way to guess what a particular pronoun's antecedent might be. This might seem surprising at first but makes intuitive sense when you think about the acquisition problem for a concrete case. For "*Lily, who adores Sarah, admires* _", do we – as adults – know what kind of word belongs in the _ without some additional context? Even though we know there are two kinds of pronouns and we know which antecedents each can take, we just don't know which pronoun type that word ought to be without additional information. So, through a computational model, Orita et al. (2013) were able to formally demonstrate that intuition and also show exactly how integrating discourse information of a particular kind can help solve the acquisition problem for anaphora resolution.

3.2.4.4 The linking problem. To successfully learn how to use verbs correctly – that is, both how to order the words related to a verb (syntax) and how to interpret those words (semantics) – children must solve what's known as *the linking problem*: they have to learn the mapping between the thematic roles specified by a verb's lexical semantics (e.g., AGENT and PATIENT for *break*) and the syntactic argument positions specified by a verb's syntactic frame (e.g., *subject breaks object*). So, for example, if an English child hears *The penguin breaks the ice*, she realizes that *The penguin*, which is the *subject* position, links to the AGENT of *break*; in contrast, *the ice*, which is in the *object* position, links to the PATIENT of *break*. This in turn allows her to interpret this utterance to mean the penguin is the one doing the breaking and the ice is the thing being broken, rather than any number of alternative interpretations (e.g, the ice doing the breaking while the penguin is being broken, the penguin doing the breaking on behalf of the ice, the penguin doing the breaking by using the ice to do it, etc.).

Once a child realizes the linking patterns for her language, she can use these patterns to cluster verbs together that behave similarly with respect to these patterns. This is useful because it allows her to make generalizations across verbs in the same class – and this in turn allows her to predict helpful properties of those verbs: their syntactic profile (how these verbs will appear with respect to word order and number of arguments) and their interpretation profile (how these verbs will be used with their arguments to convey compositional meaning).

Let's consider a concrete example of the useful generalization process that comes from verb clustering like this. Suppose a child observed the following:

1. *appear* can be used in constructions like *The boy appeared to fall*, *The rock appeared to fall*, *It appeared that the boy fell*, and *It appeared that the rock fell*.

2. In contrast, while *try* can be used in constructions like *The boy tried to fall*, it doesn't seem to be used in constructions like **The rock tried to fall*, **It tried that the boy fell*, or **It tried that the rock fell.*

With these observations, a child might reasonably infer that *appear* and *try* belong to different verb classes, despite them both allowing the construction *The* $NP_{+animate}$ V_1*ed to* V_2 (e.g., *The boy appeared/tried to fall*). In fact, these verbs are typically considered as belonging to the verb classes of *subject-raising* (*appear*) vs. *subject-control* (*try*) (Becker, 2006, 2007, 2009, Mitchener and Becker, 2010, Becker, 2014). Subject-raising verbs like *appear*, *seem*, and *happen* allow inanimate subjects and non-finite complements beginning with *to*, and so allow constructions like "*The rock seemed/happened to fall*". The syntactic intuition is that the subject in the main clause has "raised" from its original position in the embedded clause (*The rock appeared* $\_\__{The\ rock}$ *to fall*), and so doesn't really meaningfully connect with the main clause verb (*appear*). This contrasts with subject-control verbs like *try* and *want*, which originate in the main clause and "control" the subject of the embedded clause too – they *do* need to meaningfully connect with the main verb, which is why "**The rock tried to fall*" isn't acceptable (except perhaps in extraordinary interpretations involving animate rocks).

Becker demonstrates that children as young as three or four have already started classifying verbs this way, and are sensitive to the conceptual cue of animacy when doing so. In particular,

when an inanimate subject (like *The rock*) appears as the subject of a verb with a non-finite complement, children seem to place the verb in the subject-raising verb class. This then allows children to predict how new verbs will behave (i.e., their syntactic and interpretation profiles) on the basis of which verb class they get sorted into. So, *The rock daxed to fall* makes children believe *dax* is a verb like *appear*, *seem*, or *happen*. (Becker and Mitchener explicitly demonstrate how this process could successfully occur in a computational model sensitive to animacy distributional information.) Then, these same children would predict that *It daxed that the rock fell* should be allowed, as this is a construction subject-raising verbs allow. They would also predict that the construction *The rock daxed to fall* should be interpreted as *The rock* only meaningfully connecting to *fall* and not to *dax*, just as other subject-raising verbs are interpreted in this construction.

Pearl and Sprouse (2019) investigated the acquisition of a variety of verb classes, using conceptual cues like animacy and thematic roles (e.g., AGENT and PATIENT), as well as syntactic cues like the distribution of syntactic constructions a verb occurs in. More specifically, Pearl and Sprouse (2019) use the acquisition of verb classes to evaluate different theories of how children solve the linking problem. That is, solving the linking problem involves learning how to map thematic roles to syntactic positions for different verbs; verb classes are collections of verbs that accomplish this linking the same way. For example, subject-raising verbs like *appear* don't have a thematic role associated with the main clause *subject* position (*The boy*$_{\varnothing}$ *appeared to fall*); in contrast, subject-control verbs like *try* do (*The boy*$_{\text{EXPERIENCER}}$ *tried to fall*). So, the verb classes children learn (and the behavior of the verbs in those classes) indicate the linking patterns children have learned for different clusters of verbs. These verb classes serve as an expected output for a modeled child solving the linking problem, and are something that we can behaviorally observe in actual children. In particular, Pearl and Sprouse (2019) used 12 types of syntactic or interpretation behavior surveyed from a large collection of child behavioral studies in order to identify verb classes that three-, four-, and five-year-old English children have. These behaviors included unaccusativity (*The plate broke*), ditransitivity (*Jack gave her a cookie*), allowing a non-finite complement beginning with *to* (*Jack wanted to see the penguin*), appearing in the passive (*The plate was carried*), allowing a finite complement beginning with *that* (*Jack knew that Lily was right*), allowing a finite complement beginning with *whether* or *if* (*Jack knew if Lily was right*), subject-control (*Jack wanted to see the penguin*), subject-raising (*Jack seemed to see the penguin*), object-control (*Lily asked Jack to hug the penguin*), object-raising (*Lily wanted Jack to hug the penguin*), psych subject-experiencer (*Jack loved Lily*), and psych object-experiencer (*Jack surprised Lily*). From these verb behaviors at ages three to five, Pearl and Sprouse (2019) derived age-specific verb classes that a modeled learner should attempt to match when it learns from the same data that three-, four-, or five-year-olds learn from.

With this in mind, Pearl and Sprouse (2019) investigated two prominent theories for how children solve the linking problem: the Uniformity of Theta Assignment Hypothesis (UTAH) and the relativized form of that theory, relativized UTAH (rUTAH). The key difference between these two approaches has to do with how they deal with individual thematic roles. UTAH groups these roles into larger thematic categories (like AGENT-ish, PATIENT-ish, or OTHER-ish) that then *always* map to a specific syntactic position (e.g., AGENT-ish $\leftrightarrow$ *subject*). In contrast, rUTAH assumes there's a relative ordering of thematic roles (e.g, AGENT$>$PATIENT), and whichever thematic roles are

present are sorted according to this ordering. Then, the highest role available maps to the highest syntactic position (e.g., *subject*). This gives rUTAH more flexibility than UTAH, for example allowing it to easily handle unaccusative constructions like *The ice*$_{\text{PATIENT}}$ *breaks*. In this case, UTAH would expect the *subject* to have an AGENT-ish role, which PATIENT certainly isn't. In contrast, rUTAH would note there was only one role available (PATIENT) and expect that role to appear in the highest syntactic position available (*subject*), which it does.

Notably, Pearl and Sprouse (2019) also investigated whether those linkings – UTAH or rUTAH – were known already by children or were instead learned over time. If the linkings were already known (perhaps because they were innately specified the way UTAH and rUTAH theorists had previously assumed), we would expect modeled three-year-olds using this linking knowledge to create verb classes that best match observed three-year-old verb classes. In contrast, if the linkings were learned over time (perhaps because they were derived from the children's language experience), we might expect that older modeled children using this linking knowledge would best match the relevant observed verb classes from actual older children; however, younger modeled children who lack this linking knowledge would best match the observed verb classes from actual younger children.

Pearl and Sprouse (2019) used a computational-level hierarchical Bayesian model that learned from realistic samples of speech directed at three-, four-, or five-year-old English children. A modeled child's perceptual intake for each utterance involving a verb was the syntactic phrase structure, the animacy of a verb's arguments, and the thematic roles of a verb's arguments. The acquisitional intake involved several pieces of information. First, the modeled child extracted the syntactic frame a verb was used in (e.g., *The ice broke* would have the frame NP _ for *break*). Second, the modeled child heeded the animacy of the verb's arguments and information specific to the particular linking theory assumptions being used: a modeled child using UTAH would map the thematic roles to their respective thematic categories; a modeled child using rUTAH would map the thematic roles to their order (HIGHEST, 2ND-HIGHEST, and so on). A modeled child who didn't have linking theory knowledge would then simply track where the thematic category appeared syntactically (e.g., PATIENT-ish or HIGHEST in *subject* for *The ice broke*); a modeled child who had linking theory knowledge would instead note if the thematic categories appeared where they were expected to or seemed to have moved (e.g., PATIENT-ish in *subject* is unexpected for UTAH, while HIGHEST in *subject* is expected for rUTAH). So, based on the linking theory assumptions a modeled child was using (UTAH vs rUTAH, not having linking knowledge vs. having it already), the acquisitional intake could look very different.

Using this approach, Pearl and Sprouse (2019) found support for two compelling ideas. First, *both* thematic representation options seem to be used by English children at different ages: either UTAH or rUTAH explained five-year-old verb classes, but only UTAH best explained four-year-old verb classes, and only rUTAH best explained three-year-old verb classes. So, children's thematic system knowledge isn't static, but rather something that evolves over development. Second, children's verb classes are best explained by children not having built-in knowledge of linking theories. Instead, only five-year-old verb classes are best matched by modeled children with existing linking knowledge; three- and four-year-old verb classes are best matched by modeled children without this linking knowledge. This suggests that children may derive knowledge of linking patterns over

time, rather than knowing it *a priori*.

Given this, Pearl and Sprouse (2021) investigated which linking theory, UTAH or rUTAH, would be easier to derive from English children's input. Using the same realistic sample of speech directed at three-, four-, and five-year-old children, Pearl and Sprouse (2021) concretely defined how children might construct the linkings specified by both UTAH and rUTAH. In particular, they modeled children who would examine individual links in the input (e.g., AGENT-ish ↔ *subject* vs. PATIENT-ish ↔ *subject*) in order to derive a complex linking pattern that connected all thematic categories to syntactic positions, and all syntactic positions to thematic categories. Then, that complex linking pattern would be evaluated against the data from the verbs in children's input.

The evaluation process at both stages (i.e., evaluating whether individual links were strongly enough attested and whether complex linking patterns were strongly enough attested) was accomplished by using the Tolerance Principle. Recall from 2.3.3 that the Tolerance Principle is a computational-level decision criterion for deciding when a generalization has enough support (i.e. few enough exceptions) to support its adoption by the child. So, Pearl and Sprouse (2021) used the Tolerance Principle to assess which links were strong enough to generalize and which complex linking pattern was strong enough to generalize, based on children's acquisitional intake. This acquisitional intake was a subset of the acquisitional intake used by the modeled learners in Pearl and Sprouse (2019): it was just the thematic categories for the syntactic positions associated with a verb (e.g., *The ice broke* would have PATIENT-ish ↔ *subject* (UTAH) or HIGHEST ↔ *subject* (rUTAH)).

Their results suggested that a child using the Tolerance Principle to learn linking theories from English child-directed speech data has two advantages for learning rUTAH instead of UTAH. First, a rUTAH-learning child will have a significantly easier time explicitly generating complex linking patterns for solving the linking problem from individual links. Second, rUTAH is the only complex linking pattern that can be successfully generalized from these realistic child-directed data using the Tolerance Principle. So, taken together with Pearl and Sprouse (2019), these results suggest that English children may solve the linking problem by deriving a relativized linking theory like rUTAH over time, with this complex linking theory in place by five years old.

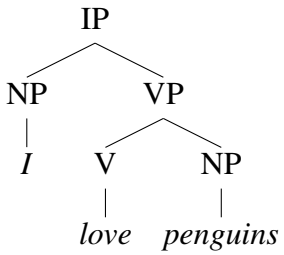
3.2.4.5 Which structured representations exactly? Recall from section 3.2.2 that a child using Bayesian inference may be able in principle to infer that hierarchichally structured representations of utterances are preferable to ones that aren't. This then leaves open the question of exactly *which* hierarchically structured representations should be selected out of all the ones possible, with the idea that the correct representations will allow the child to successfully parse the utterances of her language – this is also the approach assumed by the Structural Triggers Learner (STLearner) from section 3.1.1 and the hierarchical Bayesian learner from Perfors et al. (2011) of section 3.2.2. However, in contrast to both these prior learning approaches, which only rely on syntactic information, the approach of Abend, Kwiatkowski, Smith, Goldwater and Steedman (2017) incorporates information from both semantics and syntax; leveraging both information sources allows the learner to simultaneously learn the meanings associated with individual words and the syntactic structures that will yield the correct utterance-level meaning for utterances in the child's input. This kind of learning, when representations of different kinds are learned simultaneously,

is often referred to as *joint learning* or *bootstrapping*, with the idea that partial information of one type can help the child learn about representations of another type, and vice versa. For example, syntactic bootstrapping would involve partial syntactic information constraining hypotheses about what individual words mean; semantic bootstrapping would involve partial semantic information constraining what syntactic structures are compatible with the observed utterance. More generally, this bootstrapping approach is an example of syntax viewed within the context of the larger linguistic system; so, learning about syntax involves learning both *from* non-syntactic (though related) information and learning *about* non-syntactic (though related) information.

For Abend et al. (2017), a correct parse of an utterance doesn't just include all and only the observable words; it also allows the child to associate the correct utterance-level meaning with the parsed utterance. The key assumption the child must have is that there's a transparent mapping from the syntactic structure to the logical form (i.e., the meaning) of an utterance. So, if the utterance has the right parse, it will yield the right logical form, and the logical form will match the utterance-level meaning, as shown in (15) below.

(15) A parsed utterance[9], corresponding logical form, and intended meaning

a. Parsed utterance:



b. Associated logical form: *love*(*I*, *penguins*)

c. Intended meaning: *the speaker has happy feelings towards penguins*

With this mapping firmly in place in the initial state, the modeled child of Abend et al. (2017) can make serious progress with only a few other components. First, the modeled child has some ability to perceive the utterance-level meaning from non-syntactic context – that is, based on the sorts of things happening in the child's environment and what's salient to the child, the child can infer some kind of utterance-level meaning that's not too far off from the true meaning. This may seem surprising to assume, but it may be more reasonable in context than we think. For instance, the utterance "*I love penguins*" might occur when the child and parent are both looking at a picture of penguins, and the parent is smiling and pointing at the penguins. The parent may have also previously said things about penguins, such as "*Look at those penguins!*" or "*Those penguins are so cute!*" It turns out that speech directed at children under the age of two is often about the "here and now" (Frank, Tenenbaum and Fernald, 2013a, Clerkin, Hart, Rehg, Yu and Smith, 2016), and so likely to refer to things that are visually salient to the child. Therefore, this kind of scenario isn't implausible. So, given this context, the child might then be able to infer that the penguins are

[9]Note that the notation in (15a) uses phrase structure grammar categories for ease of comparison with the other examples in the rest of the chapter, but the actual notation should be that of Combinatory Categorial Grammar (CCG), where, for example, VP might be notated in English as S\NP. See Abend et al. (2017) for an accessible overview of CCG.

being talked about and the parent is expressing some kind of positive sentiment towards them – this translates into a meaning whose logical form is something like *love(I, penguins)*. Abend et al. (2017) assume that the modeled child has access to a universal conceptual structure, which allows the child to generate that logical form from the perceived utterance meaning.

The next component the modeled child relies on is a pre-defined hypothesis space of syntactic structure rules. Importantly, rather than explicitly defining all possible rules (which would be difficult, as there are infinitely many!), the hypothesis space is implicitly defined via a set of constraints: rules are hierarchical, rules are constrained in how their elements combine, and rule elements are composed from a pre-defined set of building blocks. That is, with these constraints, the *latent* hypothesis space is defined;[10] when inferring which syntactic structures are best, the modeled child explicitly generates specific hypotheses for consideration from this latent hypothesis space, using the building blocks defined by that latent hypothesis space. The utility of this approach is that very large latent hypothesis spaces can be considered by the child, but still feasibly searched for the best hypothesis. In fact, the idea of large latent hypothesis spaces is something familiar to those who work in the parameters framework. For instance, n binary parameters implicitly define a latent hypothesis space of 2^n grammars; the variational learning approach used by Yang (2002, 2004) is one way to search this latent hypothesis space by generating explicit hypotheses about grammars to test as the data come in. Abend et al. (2017) harness this same idea for defining and searching a latent hypothesis space of syntactic structure rules.

The last two components are biases the modeled child has. The first is to use Bayesian inference when evaluating hypotheses about potential syntactic structures. The second is a predisposition to make use of structural parts that have been used before – that is, when considering a potential syntactic rule, the child is biased in favor of structural components that have been used in other syntactic rules. This is similar in spirit to an STLearner preference for treelets that have helped the STLearner successfully parse utterances before. The idea is that reusing pieces that have been used before is efficient, and therefore preferred in potential syntactic rules. This is represented implicitly in the Bayesian learner as an overhypothesis about structural pieces – so, any time a particular structural piece is used in a syntactic rule, the learner takes note of that. The learner then considers the popularity of the pieces involved in any syntactic rule under consideration. In this way, utterances that seem unrelated on the surface may in fact be connected by a commonly used structural piece (similar to how linguistic parameters are meant to work).

With this initial state set, the modeled child of Abend et al. (2017) is then ready to learn. Its acquisitional intake involves both the words of the utterance in order (e.g., *I love penguins*), and the logical forms of both that utterance and some number of utterances that occur around it (e.g., something like *love(I, penguins)*, *be(those(x, penguins(x)), so cute)*, and *look(at(those(x, penguins(x))))*). The fact that multiple (presumably related) logical forms are taken in is meant to represent the child's uncertainty about the true logical form of the utterance, something that seems cognitively plausible in the contexts children encounter. That is, while the child is able to extract some core conceptual structure (which this set of utterances may have in common because of discourse context and the child's here-and-now environment), there's uncertainty in exactly which

[10] See Perfors (2012) for a helpful discussion of latent vs. explicit hypothesis spaces, and Pearl and Sprouse (2021) for discussion targeted at a latent hypothesis space of linking theories.

conceptual structure corresponds to the utterance under consideration.

The data themselves come from a subset of the Brown-Eve corpus in the CHILDES database (MacWhinney, 2000), and are encountered by the modeled child in the order the actual child encountered them, one at a time. Idealized Bayesian inference is done for each utterance, updating the child's set of word meanings, syntactic categories, and syntactic rules using those categories. So, this is an example of a computational-level model that nonetheless learns incrementally (one utterance at a time). It's computational-level because inference is idealized, and so the model investigates what's possible in principle, given this conceptualization of the child's acquisition task. However, by learning incrementally, this learner can generate a developmental trajectory, where its knowledge and behavior can be inspected at any point during learning. This property is particularly useful for comparing the modeled child's developmental patterns against known developmental patterns in actual children, which Abend et al. (2017) in fact do.

That is, in addition to evaluating the modeled child on the knowledge it achieves (target knowledge), Abend et al. (2017) also evaluate the modeled child on qualitative behavior patterns observed via naturalistic or experimental setups (target behavior). For target knowledge, the modeled child is tested on whether it can generate the correct logical form for an utterance it hasn't heard before. The only way for it to do this is to use its acquired syntactic representations to parse the utterance, and then map that parsed form to a logical form. This can be thought of as syntactic bootstrapping, because the syntactic structure is what allows the modeled child to generate a meaning for the utterance (and all its parts). The modeled child achieved up to 60% accuracy at inferring the correct logical form after only a few thousand utterances of input. While this is fairly impressive given the modeled child's limited input and many competing alternative parses, what's even more striking is the qualitative behavior of the modeled child.

As mentioned above, the modeled child is capable of syntactic bootstrapping – just like actual children in many experiments. This in turn allows the model to do something else actual children do in numerous experiments: one-shot learning of words when the words are embedded in a syntactic context (i.e., an utterance that includes other words) and the child has enough syntactic knowledge internalized to use that syntactic context. Because of this syntactic bootstrapping, the modeled child also shows accelerated learning of individual vocabulary items corresponding to specific syntactic categories (like nouns and transitive verbs). As the child acquires more syntactic knowledge, it becomes faster and faster at acquiring new word meanings. Related to this, it turns out that it's easier for the modeled child to acquire the syntactic structure associated with nouns, and so nouns are bootstrapped earlier in development. This then leads to the commonly-observed "noun bias" in children's early vocabularies (Braginsky, Yurovsky, Marchman and Frank, 2015).

The modeled child is also capable of sudden "inspiration" – that is, sudden jumps in learning, where many different aspects of knowledge are now present. This is in part due to the overhypotheses about structure – information about the right representations can come from many different sources. So, one piece of information can be the tipping point for a particular popular structure, and every utterance using that popular structure then benefits. This again may remind us of linguistic parameters. The main difference is that the modeled child of Abend et al. (2017) doesn't hardcode prior knowledge of specific linguistic structures connected together under a linguistic parameter. Instead, the popular structures emerge as explicit hypotheses from the latent hypothesis

space because the data themselves show these structures get reused a lot. All the modeled child has built in is the overhypothesis to look for structural pieces in common across utterances and the bias to prefer popular ones.

Taken together, the results from the computational-level learner of Abend et al. (2017) suggest that many observed developmental effects for children's syntactic knowledge may not require syntax-specific mechanisms. Children *do* need structured representations (here: the semantics encoded in a logical form) and constraints on how representations can be built (here: constraints on the form of syntactic structures). With this knowledge, the knowledge that there's an intimate connection between logical form and syntactic form, and a few domain-general biases, children can in principle bootstrap their way to sophisticated syntactic knowledge and a variety of observed acquisition behaviors.

4 Where we are and where we're headed

4.1 Where we are

As language scientists interested in understanding syntactic acquisition, we can use computational modeling to formalize specific proposals about how syntactic acquisition could work and empirically evaluate those proposals. The model allows us to look inside the acquisition process in a very precise way so that we can explain exactly how and why the proposal works (or doesn't). This gives us a lot of explanatory power, particularly for aspects of syntactic acquisition that are difficult to understand through other methods (for example, what the data intake can be or which learning strategies can succeed). Using computational modeling, we can understand more about both the developing representations in children during syntactic acquisition and the learning mechanisms that work in tandem with those developing representations.

The key to building an informative computational model is characterizing the syntactic acquisition task with as much precision as we can. This involves defining several task aspects, including the modeled learner's initial state, the data intake used for learning, the inference process that updates the learner's internal representations, the learning period during which acquisition occurs, and the learner's target state. To make our models match the human language acquisition process as much as possible, we try to ground all these aspects in empirical data. This includes considerations of cognitive plausibility, so that the modeled learner is doing something that we think children are potentially capable of doing.

When we do this, we have the best chance of creating informative computational models of syntactic acquisition. As discussed earlier in the chapter, there are a variety of ways for computational models to be informative. Computational-level models investigate whether a specific learning strategy (and, importantly, the learning assumptions that strategy includes) will work in principle – if inference were optimal, are these assumptions enough to get the job done? Algorithmic-level models investigate if a learning strategy will work in practice for children, who have cognitive limitations – when inference is approximate, are these learning assumptions good enough? Implementational-level models investigate if the strategy will work in practice in children's brains, which have biological limitations – when the representations and inference are encoded in the

wetware of the brain, are these learning assumptions good enough?

Something important to remember is that a model's inference process is about how the modeled learner updates the underlying representations, based on the data intake. Typically, this involves more or less fancy ways of counting things, and reasoning about those counts. Where linguistic theory usually comes in – and what often drives a modeled learner to acquisition success or failure – is what specifically in the input is being counted. To put it simply, it's not *that* things are being counted, it's *what* things are being counted. This is often what varies from acquisition theory to acquisition theory, and what does the explanatory work.

In this chapter, I surveyed several syntactic acquisition models that demonstrated exactly this, highlighting the importance of the modeled learner's assumptions about what data from the input were relevant to learn from, based on what was being learned. These models included both parametric and non-parametric approaches to syntactic acquisition, and all the models evaluated specific proposals about the exact knowledge and abilities necessary to solve certain syntactic acquisition tasks. Common elements of the parametric approaches were that (i) data perceived as unambiguous are particularly impactful, and (ii) the acquisitional intake is generated by parsing the input data with the currently available parametric options. Notably, parametric approaches rely on very precise, language-specific innate knowledge, in the form of linguistic parameters. Non-parametric approaches don't necessarily have to, although they *can* incorporate precise, language-specific prior knowledge. Interestingly, a common element of the non-parametric models surveyed here is that less-specific prior linguistic knowledge can often be effectively used to solve different syntactic acquisition tasks. Interestingly, this less-specific prior knowledge often involves a bias for structured linguistic representations.

Many syntactic acquisition models I discussed also encode very precise ideas about how developing representations and developing processing abilities work together during syntactic acquisition. An exciting direction in the syntactic acquisition modeling community is to develop more articulated models of this kind (Lidz and Gagliardi, 2015, Omaki and Lidz, 2015), which recognize how syntactic acquisition occurs as part of a larger cognitive system.

4.2 Where we're headed

This idea of syntactic acquisition occurring as part of broader linguistic (and cognitive) development underscores other sources of information we'll want to integrate into our syntactic acquisition theories, including other types of syntactic, linguistic, and non-linguistic information. That is, even for what seems to be an acquisition task that targets a specific piece of syntactic knowledge, children may well be using a variety of data sources to either constrain possible hypotheses or helpfully search through those hypotheses (or both).

We saw this in the case studies of pronoun interpretation, where other syntactic data (English anaphoric *one*: data coming from other pronouns) or other linguistic data (anaphora resolution: discourse information) were successfully leveraged. We also saw this in case studies of the linking problem and learning about syntactic structure more generally, where other types of conceptual information were incorporated: animacy, thematic roles, and logical forms of utterances.

Relatedly, it may be useful to consider how syntactic acquisition may be bootstrapped by other representations developing at the same time. This kind of joint learning often allows a child to learn

better and more quickly than sequential learning (in sequential learning, first a "foundational" representation is learned completely, and then the representation that builds on that foundational representation is learned). For example, in the realm of phonology, learning about phonetic categories first and then word forms using those categories second is harder than learning about phonetic categories and word forms at the same time (Feldman, Griffiths, Goldwater and Morgan, 2013). It's also harder to learn phonetic categories first and phonological rules using the phonetic categories second, rather than learning the correct phonological rules at the same time as learning phonetic categories (Dillon, Dunbar and Idsardi, 2013). In the realm of syntax, it's typically been assumed by modelers that syntactic categories are known before the syntactic rules that depend on those categories are learned. Yet, Abend et al. (2017) demonstrate that great acquisition strides can be made when both of these representations (categories and the rules that use them) are learned at the same time. This is all due to the power of bootstrapping, where partial information from one representation is helpful for learning about another representation.

Exploring a broader range of data sources for children's acquisitional intake may also help us better understand how children solve the induction problems that occur in syntactic acquisition (and language acquisition more generally). The key idea is that a wealth of indirect positive evidence may exist for the specific syntactic knowledge children need to acquire, simply because children are learning a linguistic system as a whole and not just isolated pieces of it. More concretely, by thinking about syntactic acquisition data this way, we may develop additional answers for the representations and learning mechanisms necessary for successful syntactic acquisition. These answers may complement or extend existing proposals for solving acquisition tasks that seem to require prior knowledge in children (e.g., see a handy list of examples from Hsu and Chater (2010), along with estimates of which ones are more likely to cause problems for acquisition without additional prior knowledge).

Of course, children have to know to look (and more importantly, where to look) for just the right sources of information in the vast input signal around them. This just-right search of the input likely requires prior knowledge about what to look for (e.g., Perfors et al. 2011, Pearl and Sprouse 2013a, Pearl and Mis 2016). Moreover, children have to be able to reliably extract out the just-right information from their input. This is where their developing language processing abilities and extralinguistic abilities have a huge impact, as the articulated models of Lidz and Gagliardi (2015) and Omaki and Lidz (2015) show. This is why understanding more about children's developing parsing and extralinguistic abilities is paramount for building more informative and integrated computational models of syntactic acquisition. We can also implement more informative models once we have more precise empirical data about what children know when, and what abilities are available then.

Another exciting theoretically-oriented takeaway relates to linguistic parameters. While traditional linguistic parameters are very intriguing from an acquisition perspective for all the reasons I've discussed in this chapter, they've run into empirical coverage problems as comparative linguists have surveyed more and more data from the world's languages.[11] For one thing, it's difficult to identify a single parameter that doesn't have exceptions – that is, the structures that are meant to be connected to a parameter are *often* connected as expected, but sometimes they're not (e.g., the

[11] Some of these data are helpfully available at http://wals.info, the World Atlas of Language Structures.

headedness parameter: German has prepositions before their objects (head-initial) but verbs after their complements (head-final)). This is a problem if linguistic parameters are hardcoded into a child's prior knowledge because exceptions shouldn't be allowed at all; however, if linguistic parameters emerge as overhypotheses from structural pieces that keep getting reused as the child tries to understand her language, then it's not surprising to find exceptions. The "linguistic parameter" for that language just doesn't happen to involve the exceptional structural piece. Importantly, the exception doesn't have to be pre-defined – and neither does the linguistic parameter. Linguistic parameters can have more variation in how they get instantiated from language to language, but what's in common is how they're derived from the prior knowledge the child has about structured representations and using those representations to parse and interpret the input. So, less-specific prior knowledge of structured representations coupled with the available language data and the child's learning mechanisms may allow us to derive what linguistic parameters actually look like, and how much variation there is in what they look like.

One current empirical hurdle is the lack of large-scale datasets of structurally-annotated child-directed speech from different languages. The CHILDES Treebank[12] currently provides approximately 201K utterances of speech directed at North American English children which is annotated with phrase structure, as well as some animacy and thematic role information. This has been enough to start evaluating English syntactic acquisition models. Right now, many computational models of syntactic acquisition focus on English because that's where the empirical data are easily available. But that means we only have an English-focused modeling snapshot of the universal process of syntactic acquisition that all typically-developing children are supposed to go through. To evaluate our syntactic acquisition theories more thoroughly with computational modeling techniques, we need structurally annotated data from other languages.

We also need data from other populations that may have quantitatively or qualitatively different input than those whose input data we've seen so far. For example, we know that there are both quantitative and qualitative differences in children's linguistic input across socioeconomic status (SES) (e.g., see Schwab and Lew-Williams (2016) for an overview). However, we currently don't know if the input needed for syntactic acquisition differs quantitatively, qualitatively, or both across SES. We also don't know to what extent any such differences impact the development of the syntactic knowledge itself. This, however, is exactly what a computational model could tell us, if only we had the data to feed into it.

With this in mind, my student Alandi Bates recently syntactically annotated just under twenty thousand utterances of low-SES child-directed speech.[13] We then investigated the quantity and quality of the input with respect to learning about syntactic islands (Bates and Pearl, 2019, 2020). Our results suggest first that the *wh*-dependency data in low-SES's children is quantitatively similar to the *wh*-dependency data in high-SES children's data. This is a welcome finding, as it suggests that this kind of complex linguistic input doesn't differ that much across SES, unlike many other more fundamental input types. Moreover, by applying the same syntactic islands learning model used by Pearl and Sprouse (2013a), we found that low-SES children's *wh*-dependency input doesn't differ qualitatively either. That input will allow a child to successfully internalize knowledge of the

[12] Available at http://www.socsci.uci.edu/~lpearl/CoLaLab/CHILDESTreebank/childestreebank.html

[13] Available as part of the CHILDES Treebank.

same islands that high-SES children's data does, assuming the child is using the same probabilistic learning strategy. This is an even more welcome finding, as it suggests that any linguistic gap across SES doesn't extend to this complex syntactic level – once children have the ability to take in the available *wh*-island data in their input and process it, they'll be able to learn syntactic islands knowledge. There may not be a complex syntax gap across SES, so to speak, but much more investigation of this type remains to be done.

And that's really the big picture about using computational models for syntactic acquisition: They're informative tools for answering certain kind of questions, provided we have the right empirical data to base them on. When we understand what knowledge and abilities children have available at each stage of syntactic development, what data they have available, and what exactly they know when, we can better determine what building blocks they use to acquire syntax as well as they do.

References

ABEND, OMRI; TOM KWIATKOWSKI; NATHANIEL J SMITH; SHARON GOLDWATER; and MARK STEEDMAN. 2017. Bootstrapping language acquisition. *Cognition* 164.116–143.

ALISHAHI, AFRA. 2010. Computational modeling of human language acquisition. *Synthesis Lectures on Human Language Technologies* 3(1).1–107.

ALISHAHI, AFRA; and SUZANNE STEVENSON. 2008. A computational model of early argument structure acquisition. *Cognitive science* 32(5).789–834.

BAR-SEVER, GALIA; RACHAEL LEE; GREGORY SCONTRAS; and LISA PEARL. 2018. Little lexical learners: Quantitatively assessing the development of adjective ordering preferences. *Proceedings of the 42nd Annual Boston University Conference on Child Language Development*, ed. by A. Bertolini and M. Kaplan, 58–71. Somerville: Cascadilla Press.

BAR-SEVER, GALIA; and LISA PEARL. 2016. Syntactic categories derived from frequent frames benefit early language processing in English and ASL. *Proceedings of the 40th Annual Boston University Conference on Child Language Development*, 32–46. Somerville: Cascadilla Press.

BATES, ALANDI; and LISA PEARL. 2019. **What do you think that happens?* A quantitative and cognitive modeling analysis of linguistic evidence across socioeconomic status for learning syntactic islands. *Proceedings of the 43rd Annual Boston University Conference on Language Development*, ed. by Megan Brown and Brady Dailey, 42–56. Somerville: Cascadilla Press.

BATES, ALANDI; and LISA PEARL. 2020. Identifying if input differences are developmentally meaningful: A look at complex syntactic input across socio-economic status. https://ling.auf.net/lingbuzz/004687.

BATES, ALANDI; LISA PEARL; and SUSAN BRAUNWALD. 2018. I *can* believe it: Quantitative evidence for closed-class category knowledge in an English-speaking 20- to 24-month-old child. *Proceedings of the Berkeley Linguistics Society*, Berkeley, CA: BLS.

BECKER, MISHA. 2006. There began to be a learnability puzzle. *Linguistic Inquiry* 37(3).441–456.

BECKER, MISHA. 2007. Animacy, expletives, and the learning of the raising-control distinction. *Generative approaches to language acquisition North America* 2.12–20.

BECKER, MISHA. 2009. The role of np animacy and expletives in verb learning. *Language Acquisition* 16(4).283–296.

BECKER, MISHA. 2014. *The acquisition of syntactic structure: Animacy and thematic alignment*, vol. 141. Cambridge University Press.

BERWICK, ROBERT; PAUL PIETROSKI; BERACA YANKAMA; and NOAM CHOMSKY. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35.1207–1242.

BOECKX, CEDRIC; and EVELINA LEIVADA. 2014. On the particulars of Universal Grammar: Implications for acquisition. *Language Sciences* 46.189–198.

BONAWITZ, ELIZABETH; STEPHANIE DENISON; ANNIE CHEN; ALISON GOPNIK; and THOMAS L GRIFFITHS. 2011. A Simple Sequential Algorithm for Approximating Bayesian Inference. *Proceedings of the Annual Conference of the Cognitive Science Society*, Cognitive Science Society.

BRAGINSKY, MIKA; DANIEL YUROVSKY; VIRGINIA A MARCHMAN; and MICHAEL C FRANK. 2015. Developmental Changes in the Relationship Between Grammar and the Lexicon. *Proceedings of the Cognitive Science Society*, 256–261.

BUSH, ROBERT R; and FREDERICK MOSTELLER. 1951. A model for stimulus generalization and discrimination. *Psychological Review* 58(6).413.

CHEMLA, EMMANUEL; TOBEN H MINTZ; SAVITA BERNAL; and ANNE CHRISTOPHE. 2009. Categorizing words using 'Frequent Frames': What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science* 12(3).396–406.

CHOMSKY, NOAM. 1965. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.

CHOMSKY, NOAM. 1970. Remarks on monimalization. *Reading in English Transformational Grammar*, ed. by R. Jacobs and P. Rosenbaum, 184–221. Waltham: Ginn.

CHOMSKY, NOAM. 1971. *Problems of knowledge and freedom*. London: Fontana.

CHOMSKY, NOAM. 1973. Conditions on transformations. *A Festschrift for Morris Halle*, ed. by S. Anderson and P. Kiparsky, 237–286. New York: Holt, Rinehart, and Winston.

CHOMSKY, NOAM. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

CLARK, ROBIN. 1992. The selection of syntactic knowledge. *Language Acquisition* 2(2).83–149.

CLERKIN, ELIZABETH M; ELIZABETH HART; JAMES M REHG; CHEN YU; and LINDA B SMITH. 2016. How everyday visual experience prepares the way for learning object names. *Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 126–131. IEEE.

CRAIN, STEPHEN; and MINEHARU NAKAYAMA. 1987. Structure dependence in grammar formation. *Language* 522–543.

DE VINCENZI, MARICA. 1991. *Syntactic parsing strategies in Italian: The minimal chain principle*, vol. 12. Dordrecht: Springer Science & Business Media.

DILLON, BRIAN; EWAN DUNBAR; and WILLIAM IDSARDI. 2013. A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science* 37.344–377.

ELMAN, JEFFREY L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48(1).71–99.

ERKELENS, MARIAN. 2009. *Learning to categorize verbs and nouns: studies on dutch*. Netherlands Graduate School of Linguistics.

FELDMAN, NAOMI; THOMAS GRIFFITHS; SHARON GOLDWATER; and JAMES MORGAN. 2013. A role for the developing lexicon in phonetic category acquisition. *Psychological Review* 120(4).751–778.

FODOR, JANET D. 1998a. Parsing to Learn. *Journal of Psycholinguistic Research* 27(3).339–374.

FODOR, JANET D. 1998b. Unambiguous Triggers. *Linguistic Inquiry* 29.1–36.

FODOR, JANET D.; and WILLIAM G. SAKAS. 2017. Learnability. *The oxford handbook of universal grammar*, ed. by Ian Roberts, xxx–xxx. Oxford, UK: Oxford University.

FODOR, JANET DEAN. 2017. Ambiguity, parsing, and the evaluation measure. *Language Acquisition* 24(2).85–99.

FODOR, JANET DEAN; and WILLIAM GREGORY SAKAS. 2005. The subset principle in syntax: Costs of compliance. *Journal of Linguistics* 41(03).513–569.

FODOR, JANET DEAN; WILLIAM GREGORY SAKAS; and ARTHUR HOSKEY. 2007. Implementing the subset principle in syntax acquisition: Lattice-based models. *Proceedings of the Second European Cognitive Science Conference*, 161–166. Hove, UK: Lawrence Erlbaum.

FORAKER, STEPHANI; TERRY REGIER; NAVEEN KHETARPAL; AMY PERFORS; and JOSHUA TENENBAUM. 2009. Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One. *Cognitive Science* 33.287–300.

FORSYTHE, HANNAH; and LISA PEARL. 2019. Immature representation or immature deployment? Modeling child pronoun resolution. *Proceedings of the Society for Computation in Linguistics 3*, 59.

FRANK, MICHAEL C; MIKA BRAGINSKY; DANIEL YUROVSKY; and VIRGINIA A MARCHMAN. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language* 44(3).677–694.

FRANK, MICHAEL C; JOSHUA B TENENBAUM; and ANNE FERNALD. 2013a. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development* 9(1).1–24.

FRANK, ROBERT; DONALD MATHIS; and WILLIAM BADECKER. 2013b. The acquisition of anaphora by simple recurrent networks. *Language Acquisition* 20(3).181–227.

FRANK, STELLA; SHARON GOLDWATER; and FRANK KELLER. 2009. Evaluating models of syntactic category acquisition without using a gold standard. *Proceedings of the 31st Annual Conference of the Cognitive Science Sociey*, 2576–2581.

FRANK, STELLA; SHARON GOLDWATER; and FRANK KELLER. 2013c. Adding sentence types to a model of syntactic category acquisition. *Topics in Cognitive Science* 5(3).495–521.

FREUDENTHAL, DANIEL; and AFRA ALISHAHI. 2014. Computational models of language development. *Encyclopedia of language development*, ed. by Patricia J Brooks and Vera Kempe, SAGE Publications.

FREUDENTHAL, DANIEL; JULIAN PINE; and FERNAND GOBET. 2010. Explaining quantitative variation in the rate of Optional Infinitive errors across languages: a comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language* 37(03).643–669.

FREUDENTHAL, DANIEL; JULIAN M PINE; JAVIER AGUADO-OREA; and FERNAND GOBET. 2007. Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science* 31(2).311–341.

FREUDENTHAL, DANIEL; JULIAN M PINE; and FERNAND GOBET. 2009. Simulating the referential properties of Dutch, German, and English root infinitives in MOSAIC. *Language Learning and Development* 5(1).1–29.

FREUDENTHAL, DANIEL; JULIAN M PINE; GARY JONES; and FERNAND GOBET. 2015. Defaulting effects contribute to the simulation of cross-linguistic differences in Optional Infinitive errors. *Proceedings of the 37th annual meeting of the Cognitive Science Society*, 746–751. Pasadena.

GIBSON, EDWARD; and KENNETH WEXLER. 1994. Triggers. *Linguistic Inquiry* 25(4).407–454.

GOODMAN, NELSON. 1955. *Fact, fiction and forecast*, vol. 74. JSTOR.

GUTMAN, ARIEL; ISABELLE DAUTRICHE; BENOÎT CRABBÉ; and ANNE CHRISTOPHE. 2015. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition* 22(3).285–309.

HSU, ANNE S; and NICK CHATER. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34(6).972–1016.

HUANG, C.-T. JAMES. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: MIT dissertation.

HYAMS, NINA. 1987. The theory of parameters and syntactic development. *Parameter setting*, 1–22. Springer.

JACKENDOFF, RAY. 1977. *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.

KAM, XUÂN NGA CAO; IGLIKA STOYNESHKA; LIDIYA TORNYOVA; JANET D. FODOR; and WILLIAM G. SAKAS. 2008. Bigrams and the Richness of the Stimulus. *Cognitive Science* 32(4).771–787.

KEMP, CHARLES; AMY PERFORS; and JOSHUA TENENBAUM. 2007. Learning overhypotheses with hierarchical bayesian models. *Developmental Science* 10(3).307–321.

KEMP, CHARLES; and JOSHUA B TENENBAUM. 2008. The discovery of structural form. *Proceedings of the National Academy of Sciences* 105(31).10687–10692.

KOL, SHELI; BRACHA NIR; and SHULY WINTNER. 2014. Computational evaluation of the Traceback Method. *Journal of Child Language* 41(1).176–199.

LASNIK, HOWARD; and MAMURO SAITO. 1984. On the nature of proper government. *Linguistic Inquiry* 15.235–289.

LEGATE, JULIE; and CHARLES YANG. 2002. Empirical re-assessment of stimulus poverty arguments. *Linguistic Review* 19.151–162.

LEGATE, JULIE; and CHARLES YANG. 2007. Morphosyntactic Learning and the Development of Tense. *Linguistic Acquisition* 14(3).315–344.

LEGATE, JULIE; and CHARLES YANG. 2013. Assessing Child and Adult Grammar. *Rich Languages from Poor Inputs*, ed. by Robert Berwick and Massimo Piatelli-Palmarini, 168–182. Oxford, UK: Oxford University Press.

LEWIS, JOHN D; and JEFFREY L ELMAN. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. *Proceedings of the 26th annual conference on language development*, .

LIDZ, JEFFREY; and ANNIE GAGLIARDI. 2015. How Nature Meets Nurture: Universal Grammar and Statistical Learning. *Annual Review of Linguistics* 1(1).333–352.

LIDZ, JEFFREY; SANDRA WAXMAN; and JENNIFER FREEDMAN. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition* 89.B65–B73.

MACWHINNEY, BRIAN. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MARR, DAVID. 1982. *Vision*. San Francisco, CA: W.H. Freeman.

MINTZ, TOBEN. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90.91–117.

MITCHENER, WILLIAM GARRETT; and MISHA BECKER. 2010. Computational models of learning the raising-control distinction. *Research on Language and Computation* 8(2-3).169–207.

NGUYEN, EMMA; and LISA PEARL. 2019. Using Developmental Modeling to Specify Learning and Representation of the Passive in English Children. *Proceedings 43rd annual Boston University Conference on Language Development*, ed. by Megan Brown and Brady Dailey, 469–482. Cascadilla Press.

NIYOGI, PARTHA; and ROBERT C. BERWICK. 1996. A language learning model or finite parameter spaces. *Cognition* 61.161–193.

OMAKI, AKIRA; and JEFFREY LIDZ. 2015. Linking parser development to acquisition of syntactic knowledge. *Language Acquisition* 22(2).158–192.

ORITA, NAHO; REBECCA MCKEOWN; NAOMI H FELDMAN; JEFFREY LIDZ; and JORDAN BOYD-GRABER. 2013. Discovering pronoun categories using discourse information. *Proceedings of the 35th annual conference of the cognitive science society*, .

PAYNE, JOHN; GEOFFREY PULLUM; BARBARA SCHOLZ; and EVA BERLAGE. 2013. Anaphoric *one* and its implications. *Language* 90(4).794–829.

PEARL, LISA. 2010. Using computational modeling in language acquisition research. *Experimental Methods in Language Acquisition Research*, ed. by Elma Blon and Sharon Unsworth, 163–184. Amsterdam, The Netherlands: John Benjamins.

PEARL, LISA. 2011. When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition* 18(2).87–120.

PEARL, LISA. 2014. Evaluating learning strategy components: Being fair. *Language* 90(3).e107–e114.

PEARL, LISA. 2020. Poverty of the Stimulus Without Tears. https://ling.auf.net/lingbuzz/004646.

PEARL, LISA; and SHARON GOLDWATER. 2016. Statistical Learning, Inductive Bias, and Bayesian Inference in Language Acquisition. *The Oxford Handbook of Developmental Linguistics.*, ed. by Jeffrey Lidz, William Snyder and Joe Pater, 664–695. Oxford, UK: Oxford University Press.

PEARL, LISA; TIMOTHY HO; and ZEPHYR DETRANO. 2017. An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. *Language Acquisition* 24.307–342.

PEARL, LISA; and JEFFREY LIDZ. 2009. When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. *Language Learning and Development* 5(4).235–265.

PEARL, LISA; and JEFFREY LIDZ. 2013. Parameters in Language Acquisition. *The Cambridge Handbook of Biolinguistics*, ed. by Kleanthes Grohmann and Cedric Boeckx, 129–159. Cambridge, UK: Cambridge University Press.

PEARL, LISA; and BENJAMIN MIS. 2011. How Far Can Indirect Evidence Take Us? Anaphoric One Revisited. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, ed. by L. Carlson, C. Höschler and T. Shipley, 879–884. Austin, TX: Cognitive Science Society.

PEARL, LISA; and BENJAMIN MIS. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language* 92(1).1–30.

PEARL, LISA; and JON SPROUSE. 2013a. Computational Models of Acquisition for Islands. *Experimental Syntax and Islands Effects*, ed. by Jon Sprouse and Norbert Hornstein, 109–131. Cambridge: Cambridge University Press.

PEARL, LISA; and JON SPROUSE. 2013b. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition* 20.19–64.

PEARL, LISA; and JON SPROUSE. 2015. Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research* 58.740–753.

PEARL, LISA; and JON SPROUSE. 2019. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language* .

PEARL, LISA; and JON SPROUSE. 2021. The acquisition of linking theories: A Tolerance and Sufficiency Principle approach to learning UTAH and rUTAH. https://ling.auf.net/lingbuzz/004088.

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

PERFORS, AMY; JOSHUA TENENBAUM; and TERRY REGIER. 2011. The learnability of abstract syntactic principles. *Cognition* 118.306–338.

PERFORS, AMY; JOSHUA TENENBAUM; and ELIZABETH WONNACOTT. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language* 37(3).607 – 642.

PERKINS, LAUREL. 2019. *How Grammars Grow: Argument Structure and the Acquisition of Non-Basic Syntax"*: University of Maryland, College Park dissertation.

PERKINS, LAUREL; NAOMI FELDMAN; and JEFFREY LIDZ. 2017. Learning an input filter for argument structure acquisition. *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, 11–19.

PHILLIPS, COLIN; and LARA EHRENHOFER. 2015. The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism* 5(4).409–453.

PHILLIPS, LAWRENCE; and LISA PEARL. 2014a. Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. *Proceedings of the computational and cognitive models of language acquisition and language processing workshop*, Gothenberg, Sweden: EACL.

PHILLIPS, LAWRENCE; and LISA PEARL. 2014b. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. *Proceedings of the 36th annual conference of the cognitive science society*, 2775–2780. Quebec City, CA: Cognitive Science Society.

PHILLIPS, LAWRENCE; and LISA PEARL. 2015a. Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. *Proceedings of CMCL* 68–78.

PHILLIPS, LAWRENCE; and LISA PEARL. 2015b. The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science* doi:10.1111/cogs.12217.

PULLUM, GEOFFREY; and BARBARA SCHOLZ. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19.9–50.

RÄSÄNEN, OKKO. 2012. Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication* 54(9).975–997.

REALI, FLORENCIA; and MORTEN CHRISTIANSEN. 2005. Uncovering the Richness of the Stimulus: Structure Dependence and Indirect Statistical Evidence. *Cognitive Science* 29.1007–1028.

REGIER, TERRY; and SUSANNE GAHL. 2004. Learning the unlearnable: The role of missing evidence. *Cognition* 93.147–155.

ROSS, JOHN. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation.

SAKAS, WILLIAM. 2003. A Word-Order Database for Testing Computational Models of Language Acquisition. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 415–422. Sapporo, Japan: Association for Computational Linguistics.

SAKAS, WILLIAM. 2016. Computational approaches to parameter setting in generative linguistics. *The oxford handbook of developmental linguistics*, ed. by Jeffrey Lidz, William Snyder and Joe Pater, 696–724. Oxford, UK: Oxford University Press.

SAKAS, WILLIAM; and JANET FODOR. 2012. Disambiguating Syntactic Triggers. *Language Acquisition* 19(2).83–143.

SAKAS, WILLIAM; and JANET D. FODOR. 2001. The Structural Triggers Learner. *Language Acquisition and Learnability*, ed. by Stefano Bertolo, 172–233. Cambridge, UK: Cambridge University Press.

SAKAS, WILLIAM; and EIJI NISHIMOTO. 2002. Search, Structure or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Manuscript.

SAVINELLI, K.J.; GREGORY SCONTRAS; and LISA PEARL. 2017. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. *Proceedings of the 39th annual meeting of the Cognitive Science Society*, London, UK: Cognitive Science Society.

SAVINELLI, K.J.; GREGORY SCONTRAS; and LISA PEARL. 2018. Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, Salt Lake City, UT.

SCHWAB, JESSICA F; and CASEY LEW-WILLIAMS. 2016. Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science* 7.264––275.

SCONTRAS, GREGORY; and LISA PEARL. 2021. When pragmatics matters more for truth-value judgments: An investigation of quantifier scope ambiguity. https://ling.auf.net/lingbuzz/005287.

SPROUSE, JON; MATT WAGERS; and COLIN PHILLIPS. 2012. A test of the relation between working memory capacity and syntactic island effects. *Language* 88(1).82–124.

WANG, HAO; BARBARA HÖHLE; NF KETREZ; AYLIN C KÜNTAY; TOBEN H MINTZ; N DANIS; K MESH; and H SUNG. 2011. Cross-linguistic distributional analyses with Frequent Frames: The cases of German and Turkish. *Proceedings of 35th Annual Boston University Conference on Language Development*, 628–640. Cascadilla Press Somerville, MA.

WANG, HAO; and TOBEN MINTZ. 2008. A dynamic learning model for categorizing words using frames. *Proceedings of the 32nd annual Boston University Conference on Language Development [BUCLD 32]*, ed. by Harvey Chan, Heather Jacob and Enkeleida Kapia, 525–536. Somerville, MA: Cascadilla Press.

WEISLEDER, ADRIANA; and SANDRA R WAXMAN. 2010. What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in spanish and english. *Journal of Child Language* 37(05).1089–1108.

XIAO, LING; XIN CAI; and THOMAS LEE. 2006. The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. *Proceedings of*

the Seventh Tokyo Conference on Psycholinguistics, ed. by Yukio Otsu, 299–322. Tokyo: Hitizi Syobo.

YANG, CHARLES. 2002. *Knowledge and Learning in Natural Language*. Oxford, UK: Oxford University Press.

YANG, CHARLES. 2004. Universal grammar, statistics or both? *Trends in Cognitive Science* 8(10).451–456.

YANG, CHARLES. 2005. On productivity. *Yearbook of Language Variation* 5.333–370.

YANG, CHARLES. 2011. A statistical test for grammar. *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics*, 30–38. Association for Computational Linguistics.

YANG, CHARLES. 2012. Computational models of syntactic acquisition. *WIREs Cognitive Science* 3.205–213.

YANG, CHARLES. 2015. Negative knowledge from positive evidence. *Language* 91(4).938–953.

YANG, CHARLES. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.

Modeling syntactic acquisition: The future

For the Oxford Handbook of Experimental Syntax

Lisa S. Pearl
University of California, Irvine
lpearl@uci.edu

It turns out that I had a lot more thoughts than I realized about the future of syntactic acquisition modeling – this is why my chapter has a subsection devoted to "where we're headed" that spans several pages. Here, I'll provide a more condensed version of those thoughts. From my perspective, there are four main takeaways from the current state of syntactic acquisition modeling.

First, children may be able to accomplish quite a lot with general-purpose prior knowledge about syntax. For instance, a reappearing element in successful syntactic acquisition models is the ability to generate structured representations of certain kinds – not to *prefer* these representations – but simply the ability to generate them at all. Coupled with domain-general learning mechanisms, (e.g., allowing for overhypotheses, preferring structural pieces that get reused), this ability can be used to converge on a variety of syntactic representations. Perhaps most interestingly, recent work has tantalizingly suggested how this kind of general-purpose linguistic prior knowledge could be used to generate syntactic knowledge that looks a lot like traditional linguistic parameters. It remains for us to work out exactly how close the derived knowledge is to traditional parameters, which will likely require much more cross-linguistic work.

Second, viewing syntax as part of a larger linguistic (and cognitive) system of knowledge makes other sources of information relevant for syntactic acquisition. That is, even for what seems to be an acquisition task that targets a specific piece of syntactic knowledge, children may well be using a variety of data sources to either constrain possible hypotheses or helpfully search through those hypotheses (or both). I've reviewed examples that use other syntactic, linguistic, and non-linguistic information, but many more relevant information sources may be available than we currently realize. Relatedly, it may be useful to consider how syntactic acquisition may be bootstrapped by other representations developing at the same time. That is, how can partial information about other representations help children learn about syntactic representations, and how can partial information about syntactic representations help children learn about those other representations? The key idea is that a wealth of indirect positive evidence may exist for the specific syntactic knowledge children need to acquire, simply because children are learning a linguistic system as a whole and not just isolated pieces of it.

Related to this, the third takeaway is the need to develop more articulated syntactic acquisition models that recognize the impact of developing language processing and extralinguistic abilities

on the information available to children as they acquire their syntactic representations.

Fourth, a current empirical hurdle is the lack of large-scale datasets of structurally-annotated child-directed speech from different languages and different socioeconomic statuses (SESes). Right now, many computational models of syntactic acquisition focus on high-SES English because that's where the empirical data are easily available. But that means we only have a high-SES English-focused modeling snapshot of the universal process of syntactic acquisition that all typically-developing children are supposed to go through. To evaluate our syntactic acquisition theories more thoroughly with computational modeling techniques, we need the structurally-annotated developmental data to do it.

Modeling syntactic acquisition: An annotated bibliography

For the Oxford Handbook of Experimental Syntax

Lisa S. Pearl
University of California, Irvine
lpearl@uci.edu

In my experience, after you have a general idea about how syntactic acquisition modeling works, the best way to learn how to model yourself is to really walk through the nitty-gritty details of specific models. You'll particularly benefit from pouring over ones which (i) are modeling phenomena that are similar to the phenomena you're interested in, or (ii) are using approaches that are similar to the approach you want to use. Because of that, the references below are models of syntactic acquisition that either provide a lot of detail about their implementation, the motivation for that implementation, or both. I've broken them down by modeling approach.

- **Bayesian inference**

  - ABEND, OMRI; TOM KWIATKOWSKI; NATHANIEL J SMITH; SHARON GOLDWATER; and MARK STEEDMAN. 2017. Bootstrapping language acquisition. *Cognition* 164.116–143.
    This demonstrates a computational-level Bayesian model of learning syntactic structure rules, with detail about the motivation for the model and a lot of detail in the appendices about model implementation.

  - ORITA, NAHO; REBECCA MCKEOWN; NAOMI H FELDMAN; JEFFREY LIDZ; and JORDAN BOYD-GRABER. 2013. Discovering pronoun categories using discourse information. *Proceedings of the 35th annual conference of the Cognitive Science Society*, vol. 35, Berlin.
    This demonstrates a computational-level Bayesian model of learning pronoun classes, with good detail about the motivations for the model design and empirical grounding of model variables.

  - PEARL, LISA; and BENJAMIN MIS. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language* 92(1).1–30.
    This demonstrates an algorithmic-level Bayesian model for learning English anaphoric *one*, with extensive supplementary material that walks through the modeling implementation.

- PEARL, LISA; and JON SPROUSE. 2018a. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. URL https://ling.auf.net/lingbuzz/003913.
  This demonstrates a computational-level Bayesian model of learning linking theories, with extensive information in the appendices about how the model is implemented and why the specific modeling decisions were made.
- PERFORS, AMY; JOSHUA TENENBAUM; and TERRY REGIER. 2011. The learnability of abstract syntactic principles. *Cognition* 118.306–338.
  This demonstrates a computational-level Bayesian model for learning to prefer structure-dependent representations, with detailed appendices about the model implementation.

- **The Tolerance Principle**

  - PEARL, LISA; and JON SPROUSE. 2018b. The acquisition of linking theories: A Tolerance Principle approach to learning UTAH and rUTAH. URL https://ling.auf.net/lingbuzz/004088.
    This demonstrates a computational-level model of deriving linking theories using the Tolerance Principle, with a lot of detail about the motivation for the implementation choices.
  - YANG, CHARLES. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.
    Chapters 3-6 provide motivation and an explicit walk-through of how the Tolerance Principle was derived (chapter 3), along with several examples of the application of the Tolerance Principle (chapters 4-6).

- **Reinforcement learning**

  - YANG, CHARLES. 2002. *Knowledge and Learning in Natural Language*. Oxford, UK: Oxford University Press.
    Chapter 2 provides the cognitive motivations for variational learning, along with its implementation, and both chapters 2 and 4 offer several examples of its application.

- **Structural Triggers Learner**

  - FODOR, JANET DEAN. 2017. Ambiguity, parsing, and the evaluation measure. *Language Acquisition* 24(2).85–99.
    This walks through the nitty-gritty about how parsing with treelets is meant to work, providing the motivation for modeling decisions.

- **Simple Recurrent Networks**

  - FRANK, ROBERT; DONALD MATHIS; and WILLIAM BADECKER. 2013. The acquisition of anaphora by simple recurrent networks. *Language Acquisition* 20(3).181–227.
    This provides a detailed walk-through of a simple recurrent network for learning how to resolve anaphora, with a special focus on how to tell what the model is doing internally to generate its observable behavior.

- **Other probabilistic approaches**

  - FREUDENTHAL, DANIEL; JULIAN PINE; and FERNAND GOBET. 2010. Explaining quantitative variation in the rate of Optional Infinitive errors across languages: a comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language* 37(03).643–669.
    This provides a detailed walk-through of the algorithmic-level MOSAIC model, compared against variational learning for the acquisition of optional infinitives.
  - PEARL, LISA; and JON SPROUSE. 2013. Computational Models of Acquisition for Islands. *Experimental Syntax and Islands Effects*, ed. by Jon Sprouse and Norbert Hornstein, 109–131. Cambridge: Cambridge University Press.
    This demonstrates an algorithmic-level model for learning syntactic islands, with details about the motivation for model implementation choices and an appendix describing the implementation.