Theory and predictions for the development of basic morphology and syntax: A Universal Grammar + statistics approach

Lisa Pearl
University of California, Irvine

Abstract

The key aim of this special issue is to make developmental theory proposals concrete enough to evaluate with empirical data. With this in mind, I discuss proposals from the "Universal Grammar + statistics" (**UG+stats**) perspective for learning several morphology and syntax phenomena. I briefly review why UG has traditionally been part of many developmental theories of language, as well as common statistical learning approaches that are part of UG+stats proposals. I then discuss each morphology or syntax phenomenon in turn, giving an overview of relevant UG+stats proposals for that phenomenon, predictions made by each proposal, and what we currently know about how those predictions hold up. I conclude by briefly discussing where we seem to be when it comes to how well UG+stats proposals help us understand the development of basic morphology and syntax knowledge.

1 Introduction

The goal of this special issue is to make different theoretical proposals concrete enough to provide testable predictions. If those predictions are then borne out, the proposal is supported; if not, the proposal isn't. Generating precise, testable predictions for theories is something I deeply support, and computational cognitive modeling (a methodology I use most often in my own work) provides one way to do exactly this (e.g., see Pearl (in press) for more detailed discussion on this point).

Here, I've been asked to represent the perspective of proposals that involve both Universal Grammar (**UG**) and statistics (so I'll refer to them as **UG+stats** proposals). In almost every case study where I'll present the UG+stats proposals I'm aware of, a proposal is implemented concretely in a computational cognitive model. Why the focus on computational cognitive models? This is because it's often hard to pin down a specific prediction that a UG+stats proposal makes without a concrete model using the proposed UG knowledge and implementing a specific learning strategy that relies on the proposed statistics. When we have a computational cognitive model, predictions about children's behavior can be generated that are precise enough to assess with empirical data that either already exist or can be obtained in the future.

So, computational cognitive modeling offers a way to implement a UG+stats developmental theory, which is typically a theory of both the linguistic representations the child is learning (this is usually the UG part) and the acquisition process the child undergoes (this is usually the statistics part). The computational model then becomes a "proof of concept" for the developmental theory, as implemented by that model (see Pearl (2014, in press) for more detailed discussion about this). This is in fact why an effective way to evaluate a UG+stats theory (or really, any developmental theory) is to implement it in a computational cognitive model; implementing the model involves embedding the relevant prior knowledge and learning mechanisms proposed for the child in the model, giving the modeled child realistic input to learn from, and generating output predictions from that modeled child that connect in some interpretable way to children's behavior. This is the approach that the proposals I'll review here have generally taken for investigating how children learn basic morphology and syntax knowledge.

I should also note that it's likely I've been asked to represent the UG+stats perspective because I often work on learning problems where UG representations are combined with statistical learning in some form. This is because I think UG approaches to development can often greatly benefit from integrating statistical learning approaches (see Pearl (in press) for more detailed discussion on this point). However, I should note that for some of the case studies in the development of basic morphology and syntax that will be discussed here, I don't necessarily agree that the UG+stats proposals I'm aware of are the best approaches. As relevant, I'll briefly note the caveats I have for the UG+stats proposals discussed.

In the remainder of this article, I'll first briefly review what UG is meant to be and why UG has traditionally been part of many developmental theories of language. I'll then discuss some common statistical learning approaches that are often part of UG+stats proposals. I'll then turn to specific morphology and syntax phenomena, and how UG+stats proposals account for each (or don't yet). More specifically, for each phenomenon, I'll first discuss specific UG+stats proposals for learning it, including a brief overview of both the UG part and the statistics part. I'll then present the predictions that the UG+stats proposals make, aiming to specify at least one prediction that would support a specific proposal and one that would undermine it. I'll then discuss whether the proposal predictions hold up, don't hold up, or if we just don't know yet. In cases where we don't know yet, we have a clear path forward for fruitful future avenues of behavioral research (namely, studies that would test specific proposal predictions). I'll conclude with a brief summary of where we seem to be when it comes to how well UG+stats proposals help us understand the development of basic morphology and syntax knowledge.

2 UG + statistics

2.1 The UG part

A key motivation for UG has always been developmental: UG could help children acquire the linguistic knowledge that they do as quickly as they do from the data that's available to them (Chomsky, 1981; Jackendoff, 1994; Laurence & Margolis, 2001; Crain & Pietroski, 2002). That is, UG would allow children to solve what's been called the *Poverty of the Stimulus* (see Pearl

(2019) for a recent review), where the available data often seem inadequate for pinpointing the right linguistic knowledge as efficiently as children seem to. So, without some internal bias, children wouldn't succeed at language acquisition. UG is then a proposal for what that internal bias could be that enables language acquisition to in fact succeed.

Typically, a UG proposal would provide a way to structure the child's hypothesis space with respect to a specific piece of linguistic knowledge – that is, UG can help define what explicit linguistic hypotheses are considered, and what building blocks allow children to construct those explicit hypotheses for consideration. For instance, traditional linguistic parameters (Chomsky, 1981, 1986) are building blocks that children can construct their linguistic system from. So, a language's system would be described by a specific collection of parameter values for these linguistic parameters. Having these parameter building blocks then allows a child to construct and consider explicit hypotheses about a language's system as she encounters her language's data. In some of the phenomena we'll discuss below (basic word order, lack of inflection, movement), linguistic parameters supplied by UG allow the child to construct a constrained set of possible hypotheses to navigate through, given her input.

More generally, a working definition of UG is that it's anything that's both innate and language-specific (Pearl, 2019, in press). So, linguistic parameters fit this definition because they would be innate knowledge and they're only used for learning language. In the specific linguistic phenomena reviewed in this article, we'll see a variety of examples of UG knowledge, as relevant for basic morphology and syntax.

2.2 The statistics part

In UG+stats proposals, the statistics part refers to statistical learning. That is, on the basis of the statistics of her input, the child is learning something. One reason statistical learning can work so well in combination with UG is that statistical learning is often used to navigate through a hypothesis space in order to identify the correct hypothesis for the language. Because UG can provide a hypothesis space to the child, statistical learning can then naturally complement UG proposals to language development.

How does this work exactly? At its core, statistical learning is about counting things (this is the "statistical" part), and updating hypotheses on the basis of those counts (this is the "learning" part, sometimes also called *inference* (Pearl, in press)). Counting things is a domain-general ability, because we can count lots of different things, both linguistic and non-linguistic (even as babies: Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998; Saffran, Johnson, Aslin, & Newport, 1999; Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002; Wu, Gopnik, Richardson, & Kirkham, 2011; Stahl, Romberg, Roseberry, Golinkoff, & Hirsh-Pasek, 2014; Ferry et al., 2016; Aslin, 2017; Fló et al., 2019). These counts can then be converted into probabilities – for example, seeing something 3 times out of 10 yields a probability of $\frac{3}{10}$=0.30. Then, things with higher probabilities can be interpreted as more likely than things with lower probabilities.

So, to effectively use statistical learning, a child has to know what to count. UG can identify what to count, because UG defines the hypothesis space. This means that the relevant things to count are the relevant things for determining which hypothesis in the hypothesis space is the right one for the language. For language acquisition, the relevant things are typically linguistic

things (though sometimes non-linguistic things might be relevant to count too, depending on what the child's trying to learn). Importantly, the statistical learning mechanism itself doesn't seem to change – once the child knows the units over which inference is operating, counts of the relevant units are collected and inference can operate. In the rest of this subsection, I'll briefly review some common approaches to doing inference over collected counts: Bayesian inference, reinforcement learning, and the Tolerance & Sufficiency Principles (for a more comprehensive overview of each, see Pearl (in press)). Table 1 summarizes which inference mechanisms are used by particular UG+stats proposals for the different basic morphology and syntax phenomena discussed in the rest of this article.

Table 1: Common inference mechanisms in statistical learning that are used by UG+stats proposals for different basic morphology and syntax phenomena: basic syntactic categories (**syn cat**), basic word order (**word order**), inflectional morphology (**infl mor**), showing a temporary lack of inflection (**no infl**), movement (**mvmt**), and constraints on utterance form and interpretation (**constr**).

	syn cat	**word order**	**infl mor**	**no infl**	**mvmt**	**constr**
Basic counts & probabilities	✓	✓	✓	✓	✓	✓
Bayesian inference	✓		✓		✓	✓
Reinforcement learning		✓		✓	✓	
Tolerance & Sufficiency			✓			

2.2.1 Bayesian inference

Bayesian inference operates over probabilities (as mentioned above, probabilities can be derived from counts). This inference mechanism involves both prior assumptions about the probability of different hypotheses and an estimation of how well a given hypothesis fits the data. A Bayesian model assumes the learner (for our purposes, the modeled child) has some space of hypotheses H, each of which represents a possible explanation for how the data D in the relevant part of the child's input were generated. For example, a UG+stats modeled child relying on a linguistic parameter to determine if her language has *wh*-movement might consider both a +wh-movement option and a -wh-movement option as two hypotheses ($\{$+wh-movement, -wh-movement$\} \in H$); the data might be the collection of questions in the child's input involving *wh*-words ($\{$*What did Jack climb?, Jack climbed what?!, ...*$\} \in D$).

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 called the *posterior* for that hypothesis. This is calculated via Bayes' Theorem as shown in (1).

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

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 where 30% of the data are compatible only with +wh-movement) have a lower likelihood; hypotheses with a good fit to the data have a higher likelihood.

$P(h)$ 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 (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, Tenenbaum, and Regier (2011) and Piantadosi, Tenenbaum, and Goodman (2012)). So, for example, more complex hypotheses will typically have lower prior probabilities. A hypothesis's prior is something that could be specified by UG – but all that matters is that the prior is specified beforehand somehow, wherever it comes from.

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 (e.g., is +wh-movement or -wh-movement more probable after seeing the data D?), calculating $P(D)$ can be skipped over and the numerator alone used (hence, the $\propto$ in (1)).

From a developmental perspective, there's a considerable body of evidence suggesting that young children are capable of Bayesian inference (3 years: Xu & Tenenbaum, 2007; 9 months: Gerken, 2006; Dewar & Xu, 2010; Gerken, 2010; 6 months: Denison, Reed, & Xu, 2011, among many others). Given this, Bayesian inference seems a plausible statistical learning mechanism for language acquisition.

2.2.2 Reinforcement learning

Reinforcement learning also operates over probabilities and is a principled way to update the probability of a categorical option which is in competition with other categorical options (see Sutton and Barto (2018) for a recent overview). So, if we're thinking in terms of linguistic parameters for *wh*-movement, a child might consider both a +*wh*-movement and a -*wh*-movement option. A common implementation used by UG+stats proposals is the linear reward-penalty scheme (Bush & Mosteller, 1951). 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). For instance, if the +wh-movement option is under consideration, and it's compatible with the current data point (like *What's Jack climbing* $\_\__{what}$*?*), the +wh-movement option is rewarded and its probability is increased. In contrast, if that same option is under consideration, but it's not compatible with the current data point (such as an echo question like *Jack's climbing what?!*), the +wh-movement option is penalized and its probability is decreased.

While applying reinforcement learning in UG approaches to language acquisition is a fairly recent innovation, reinforcement learning itself is well-supported in the child development literature more generally (sometimes under the name "operant conditioning"). In particular, we have evidence that very young children are capable of it (under 18 months: Hulsebus, 1974; 12 months: Lipsitt, Pederson, & Delucia, 1966; 10 months: de Sousa, Garcia, & de Alcantara Gil, 2015;

3 months: Rovee-Collier & Capatides, 1979; 10 weeks: Rovee & Rovee, 1969; Watson, 1969; among many others). So, it seems plausible that young children could use reinforcement learning for language acquisition.

2.2.3 Tolerance & Sufficiency Principles

The Tolerance and Sufficiency Principles (Yang, 2005, 2016) together describe a particular inference mechanism, and this mechanism operates over specific kinds of counts that have already been collected. More specifically, these principles together provide a formal approach for when a child would choose to adopt a "rule", generalization, or default pattern to account for a set of items. For example, these principles can be used to determine if there's a general rule for forming the past tense in English from a verb's root form (e.g., *kiss* $\rightarrow$ *kissed*).

Both principles are based on cognitive 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. The learning innovation of these principles is that they're designed for situations where there are exceptions to a potential rule. In the English past tense example above, there are certainly exceptions in the child's input: past tense forms like *drank* (rather than *drinked*) and *caught* (rather than *catched*).

So, these two principles help the child infer whether the rule is robust enough to bother with, despite the exceptions. In particular, a rule should be bothered with if it speeds up average retrieval time for any item. For instance, it's faster on average to have a past tense rule to retrieve a regular past tense form (like *-ed* for English). However, if the past tense is too irregular, it's not useful to have the rule: retrieving the target information (i.e., the correct past tense form) takes too long on average.

The Tolerance Principle determines how many exceptions a rule can "tolerate" in the data before it's not worthwhile for the child to have that rule at all; the Sufficiency Principle uses that tolerance threshold to determine how many rule-abiding items are "sufficient" in the data to justify having the rule. This means, of course, that the child needs to have previously counted how many items obey the potential rule and how many don't. With these counts in hand, the child can then apply the Tolerance and Sufficiency Principles to infer whether the data justify the adoption of the rule under consideration (or not).

Together, these two principles have been used for investigating a rule, generalization, or default pattern for a variety of linguistic knowledge types (Yang, 2005; Legate & Yang, 2013; Yang, 2015; Schuler, Yang, & Newport, 2016; Yang, 2016; Pearl, Lu, & Haghighi, 2017; Yang, 2017; Irani, 2019; Pearl & Sprouse, 2019). However, there isn't yet much evidence that children are capable of using the Tolerance and Sufficiency Principles – the main support comes from the study by Schuler et al. (2016), which demonstrates that 5- to 8-year-old behavior is consistent with children using these principles. Still, these principles seem like a promising statistical learning mechanism for UG+stats proposals, given their current success at predicting child behavior (more on this in the subsection on learning morphology in highly-inflected languages).

3 The phenomena

3.1 Basic syntactic categories

One foundational type of knowledge in any language is the syntactic categories of the language, and which words belong in each category. For instance, how does a child learn that her language has categories like NOUN, VERB, and DETERMINER? For English, how would a child learn that both *kitty* and *idea* are NOUNs, while *kisses* is a VERB and *the* is a DETERMINER? Ambridge (2017) notes that developmental researchers who are interested in UG have largely turned their attention away from investigating how children learn basic syntactic categories. I think why that is will become clearer when we look at the predictions that can be generated at this point in time. Relatedly, it can be tricky to tell if a proposal is really a UG+stats proposal; this is because the UG part would need to be innate, language-specific knowledge about syntactic categories, and it's not always clear the prior knowledge assumed by a proposal is necessarily UG-type knowledge (more on this below).

3.1.1 Specific UG+stats proposals (potentially)

Semantic bootstrapping. The main proposal I'm aware of that could potentially be a UG+stats proposal is semantic bootstrapping (Pinker, 1984, 1987). This proposal suggests that children have innate links between abstract syntactic categories and semantic relations (e.g., NOUN $\leftrightarrow$ name of a concrete thing). These innate links allow children to initially break into the syntactic category system, as children would expect that similar semantic relations (e.g., concrete things like *ball* and *kitty*) map to the same syntactic category (which we refer to as NOUN). Children would then rely on statistical learning to fine-tune which words really belong to which categories, on the basis of their input. So, children start with abstract syntactic categories, and via their input, they identify the true implementation of that category in their language (Valian, 2009, 2014). Importantly, the true implementation typically will go far beyond a specific semantic relation (e.g., the NOUN *idea* isn't a concrete thing).

The UG part. If children have innate links from innate abstract syntactic categories to certain semantic relations, then that would be UG knowledge – the child has innate knowledge that's specifically about language. However, it could be that the links emerge from the child considering the words that seem to be clustered together in her language in a particular category. That is, the child notices that the semantic relations encoded by the members of CATEGORY_1 (which we as adults recognize as a type of NOUN) seem to include a lot of concrete things. So, on the basis of that observation, the child constructs the hypothesis about the link ($\text{CATEGORY}_1 \leftrightarrow$ concrete things), and uses this hypothesized link to accomplish whatever the innate link would have accomplished.

Moreover, if there are innate abstract categories that are language-specific (i.e., something like NOUN and VERB), then these too would be UG knowledge. However, it's possible that the innate knowledge about categories may not necessarily be language-specific. For example, suppose a child innately knows that there are in fact categories of some kind, but doesn't have something as specific as NOUN and VERB in mind. Could we tell the difference between innate knowledge

that CATEGORY$_1$ and CATEGORY$_2$ exist, as opposed to innate knowledge that NOUN and VERB exist? What would that difference be? If the difference is about the links between categories (e.g., NOUN $\leftrightarrow$ concrete thing), then the innate knowledge is really about the links and not the categories themselves. That is, this link could just as easily be expressed as SOME_CATEGORY $\leftrightarrow$ concrete thing. As we saw above, it's not clear that this link is necessarily innate, rather than something that could be derived from the child's input. So, more generally, it's not obvious that the UG part for semantic bootstrapping is necessarily UG.

The statistics part. The learning mechanism for fine-tuning a language's syntactic category implementations is distributional learning. In distributional learning, items with the same distributions (that is, appearing in the same contexts, and so for instance preceded and following by the same elements) are perceived as the same kind of thing. The way a child might tell that two items have the same distributions is by tracking which elements precede and/or follow those items. One common implementation of distributional learning for discovering a language's syntactic categories, which relies on a statistical learning component, is an approach called Frequent Frames (Mintz, 2003, 2006; Xiao, Cai, & Lee, 2006; Wang & Mintz, 2008; Chemla, Mintz, Bernal, & Christophe, 2009; Erkelens, 2009; Weisleder & Waxman, 2010; Wang et al., 2011; Bar-Sever & Pearl, 2016). A child using Frequent Frames tracks which items appear between two elements (e.g., two words like *the_is* for a NOUN, or two morphemes like *is_-ing* for a VERB) – this is the "frames" part. The "frequent" part is that the child tracks how often frames appears and only really pays attention to those frames that are frequent (a simple way to do this is by counting how many instances of a frame have appeared). The frequent frames then form the foundation of the language-specific syntactic categories. Under a UG+stats approach, these language-specific categories can be matched against the innate, abstract categories, based on the semantic relations they encode. For instance, the *the_is* frame's items may map to NOUN, if these items correspond to concrete objects (Mintz, 2003). However, frequent frames are also compatible with a non-UG+stats approach, where the child finetunes her language's categories by using the frequent-frame-based categories as a starting point and noticing what semantic relations these categories encode.

3.1.2 Predictions made

A lot of syntactic categorization research is about when children seem to demonstrate knowledge of different syntactic categories in their language (Valian, 1986; Capdevila i Batet & Llinàs i Grau, 1995; Pine & Martindale, 1996; Pine & Lieven, 1997; Tomasello, 2000; Fisher, 2002; Tomasello & Abbot-Smith, 2002; Booth & Waxman, 2003; Tomasello, 2004; Kemp, Lieven, & Tomasello, 2005; Rowland & Theakston, 2009; Theakston & Rowland, 2009; Tomasello & Brandt, 2009; Valian, Solt, & Stewart, 2009; Yang, 2011; Shin, 2012; Pine, Freudenthal, Krajewski, & Gobet, 2013; Theakston, Ibbotson, Freudenthal, Lieven, & Tomasello, 2015; Ambridge, 2017; Meylan, Frank, Roy, & Levy, 2017; Bates, Pearl, & Braunwald, 2018). In general, very early knowledge of language-specific syntactic categories has been tacitly taken as a signal that children rely on innate (UG) knowledge to achieve that level of linguistic development so early. That is, from a UG perspective, the assumption has been that innate knowledge of abstract syntactic categories and

links from those categories to semantic relations should speed up the development of language-specific categories. So, when children seem to converge on language-specific syntactic categories very early (say, before age two), this has been interpreted as evidence for UG knowledge.

But this interpretation is difficult to be sure about without knowing what developmental trajectory we expect with vs. without the abstract category and linking knowledge. That is, how can we know that children's acquisition of syntactic categories is faster than it should have been if they didn't have this innate knowledge? For instance, it's not clear we have precise predictions about how long it should take children to identify their language-specific NOUN category if they did in fact have abstract knowledge of NOUN and linking rules like NOUN $\leftrightarrow$ concrete object. (We could assume children used something like Frequent Frames to create language-specific clusters of words and then mapped those clusters to abstract categories on the basis of the number of concrete objects named by the words in any given cluster.) Similarly, it's not clear we have precise predictions for how long it should take children if they didn't in fact have that innate knowledge, but used Frequent Frames to create language-specific clusters and then identified that some clusters seemed to have a lot of words that named concrete objects.

One option to generate these kind of precise predictions that map to specific ages of acquisition is to use an information-theoretic analysis, like the Minimum Description Length (MDL) approach leveraged by Chater & colleagues for syntactic rule acquisition (Hsu & Chater, 2010; Hsu, Chater, & Vitányi, 2011, 2013; Chater, Clark, Goldsmith, & Perfors, 2015). In essence, MDL quantifies how much space it takes to store information, with preference given to more compact storage options (see Pearl (2019) for a more detailed discussion of the MDL approach). For language acquisition, the information that needs to be stored is both the child's internal representation of some knowledge (like syntactic categories) and the data the child encounters, as encoded by using that representation. So, more complex representations (e.g., involving abstract categories and linking rules) may not be very compact compared to simpler representations (e.g., not involving either abstract categories or linking rules). However, as the child encounters data from her input, she encodes the data using the representation she has available – and a more complex representation may offer some storage savings on the incoming data, compared to a simpler representation. Over time, as the child encounters more data, those storage savings add up and can yield a "breakeven" point, where the more complex representation and the input data encoded so far take up less space than the simpler representation and the input data encoded so far. That breakeven point can be mapped to a specific age of acquisition, based on how frequently the child hears the data that the representation is encoding. I should note that I don't have a firm idea of how exactly to implement this for the problem of syntactic category representations. However, this approach seems like a promising avenue to explore if we want to try to generate precise predictions about expected ages of acquisition with vs. without UG knowledge. These expected ages could then be matched against observed ages of acquisition for different language-specific syntactic categories.

3.1.3 Prediction evaluation

As mentioned above, the basic problem of what we're predicting hasn't yet been solved, at least with respect to the expected age of acquisition. So, it hasn't yet been possible to really evaluate UG+stats proposals against the available data on age of acquisition. This may be why UG-friendly

researchers haven't spent as much energy on this area of linguistic development. I think it's still very worthwhile to understand the learning strategies that are capable of yielding language-specific adult syntactic category knowledge. But, this area is less interesting to researchers specifically interested in UG approaches to language development.

3.2 Basic word order

Another type of core syntactic knowledge is the basic canonical word order of languages that have (relatively) fixed word order. For example, English is canonically a *Subject-Verb-Object* (**SVO**) language, which is why the default way to express the idea that Lily likes penguins is *Lily*$_{Subject}$ *likes*$_{Verb}$ *penguins*$_{Object}$. In contrast, German has a canonical word order of *Subject-Object-Verb* (**SOV**); so, we might reasonably think that the way to express that same idea in German is *Lily*$_{Subject}$ *Pinguine*$_{Object}$ *liebt*$_{Verb}$. But, this isn't quite right, because in main clauses, another syntactic operation occurs called *Verb-second* (**V2**) movement, where the Verb moves to the second position in the clause and something else (like the Subject or Object) moves to the first position. This is why we're likely to hear either *Lily*$_{Subject}$ *liebt*$_{Verb}$ *Pinguine*$_{Object}$ or *Pinguine*$_{Object}$ *liebt*$_{Verb}$ *Lily*$_{Subject}$ to express the idea that Lily likes penguins, but not the canonical SOV order. More specifically, these two utterances have a structure something like what's in (2c-i) and (2c-ii), where $\_{element}$ represents the underlying position of the linguistic element:

(2) V2 movement with an underlying SOV canonical word order in German

a.				*Lily*$_{Subject}$	*Pinguine*$_{Object}$	*liebt*$_{Verb}$
b.			*liebt*$_{Verb}$	*Lily*$_{Subject}$	*Pinguine*$_{Object}$	$\_{Verb}$
c.	(i)	*Lily*$_{Subject}$	*liebt*$_{Verb}$	$\_{Subject}$	*Pinguine*$_{Object}$	$\_{Verb}$
	(ii)	*Pinguine*$_{Object}$	*liebt*$_{Verb}$	*Lily*$_{Subject}$	$\_{Object}$	$\_{Verb}$

These kind of complications, where multiple syntactic operations may be active, can make uncovering the canonical word order for a language difficult. For instance, if a child encounters an SVO utterance, and she doesn't know whether she's learning English or German, the canonical word order for her language could either be SVO (English, no V2 movement) or SOV (German, with V2 movement). This kind of ambiguity (and far more) is what children face when trying to identify the basic word order of their language.

3.2.1 Specific UG+stats proposals

The variational learning approach. The variational learning (**VarLearn**) approach (Yang, 2002, 2004; Legate & Yang, 2007; Yang, 2012) combines the UG idea of linguistic parameters with reinforcement learning; this combination allows a VarLearner to probabilistically search a hypothesis space defined by the linguistic parameters. For instance, one parameter may be VO vs. OV word order (corresponding to the SVO order of English vs. the SOV order of German), while another is -V2 vs. +V2 movement. With these two parameters and potential values, the hypothesis space consists of four possible language word orders: VO and -V2 (English), VO and +V2, OV and -V2, and OV and +V2 (German). More generally, L linguistic parameters with opt options each will

yield a hypothesis space of opt^L language word orders. In this small example, that's only 4 (2^2), but if we had 10 parameters with 2 possible values each, now we have 2^{10}=1024. So, even with linguistic parameters, the word order hypothesis space can get very large very quickly. This is why UG-oriented researchers have long been interested in how a child could navigate a hypothesis space defined by linguistic parameters (Clark, 1992; Gibson & Wexler, 1994; Niyogi & Berwick, 1996; Fodor, 1998b, 1998a; Sakas & Fodor, 2001; Sakas & Nishimoto, 2002; Yang, 2002; Sakas, 2003; Yang, 2004; Fodor & Sakas, 2005; Fodor, Sakas, & Hoskey, 2007; Sakas & Fodor, 2012; Boeckx & Leivada, 2014; Sakas, 2016; Fodor, 2017; Fodor & Sakas, 2017).

The VarLearn approach assigns probability to each parameter value for a given parameter, and typically these values are equal initially. For example, a VarLearner might start out with VO and OV each with probability 0.5, and -V2 and +V2 each with probability 0.5. When encountering a data point from the input, the VarLearner probabilistically samples a complete set of parameter values (which is equivalent to some language's word order), based on the probability of those values. So, in our example above, the VarLearner might select the VO and -V2 parameter values with probability 0.5*0.5 (prob(VO) * prob(-V2)) = 0.25. Whichever word order is sampled, the VarLearner then sees if that word order, as defined by the parameter values chosen, can account for the data point. In this example, the word order specified by VO and -V2 would be able to account for *Lily likes penguins* (*Subject Object Verb*), but not for *Pinguine liebt Lily* (*Object Verb Subject*). If the word order can account for the data point, all the participating parameter values are rewarded (and have their probability increased); if not, all parameter values are penalized (and have their probability decreased).

Over time (in particular, as the child encounters more input from her language), the idea is that the language's true parameter values will have their probabilities increased until they're near 1; the alternative parameter values will have their probabilities correspondingly decreased. Importantly, this means that unambiguous data for a parameter value are very impactful – these data will always reward the corresponding parameter value and always penalize the alternative parameter value(s). For example, data perceived by the child as unambiguous +V2 data will always reward the +V2 value and always penalize the -V2 value. This means that the parameter value perceived as having more unambiguous data (that is, an *unambiguous data advantage*) will be the one that has its probability increased to around 1 – it's the value the child will choose, given enough input. This is why VarLearn approaches typically do an analysis of the unambiguous data advantage a child might perceive from her input. The higher the unambiguous data advantage for a parameter value, the faster a child using the VarLearn strategy should converge on that parameter value. This means that age of acquisition predictions can be made from careful analysis of the child's input. Specifically, parameter values that have higher unambiguous data advantages are predicted to be learned earlier.

The UG part. Linguistic parameters are meant to be innate, language-specific knowledge. One reason linguistic parameters have been a core component of UG approaches to language development is that they're intended as extremely useful building blocks. More specifically, linguistic parameters allow a child to construct a (potentially very large) collection of explicit hypotheses about a language's word order, without having to specify all those hypotheses out beforehand.

Moreover, linguistic parameters are meant to constrain the child's possible hypotheses to those that correspond to actual languages the child may be learning. So, linguistic parameters are helpful for acquisition because they're a compact way to represent the space of possible hypotheses a child might reasonably need to consider (in this case, about word order). See Pearl and Lidz (2013), Pearl (in press), and Pearl (in press) for additional discussion about why UG approaches to acquisition like to incorporate linguistic parameters.

The statistics part. Reinforcement learning is a type of statistical learning, and forms the basis for the VarLearn learning mechanism.

3.2.2 Predictions made

As mentioned above, a VarLearn approach will often be able to analyze the unambiguous data advantage for one linguistic parameter value over another that a child would perceive from her input. On the basis of this advantage, a VarLearner can generate predictions about relative order of acquisition for different word order aspects related to different parameters. For instance, on the basis of one VarLearner analysis from Yang (2012) (shown in Table 2), it appears that English has an unambiguous advantage of 25% for *wh*-movement in questions. That is, in an English child's perceived input, the proportion of English *wh*-questions with *wh*-movement (e.g., *Who did you see?*) is .25 more than the proportion of English *wh*-questions without *wh*-movement (e.g., *You saw who?*). In contrast, it appears that German has an unambiguous data advantage of 1.2% for allowing V2 movement. So, we would then expect that +*wh*-movement in English would be learned earlier than +V2 movement in German for a VarLearn child. Based on the observed ages of acquisition shown in Table 2, that does seem to be true (+*wh*-movement in English is learned by 1 year 8 months (1;8), while +V2-movement in German is learned around 3 years old).

Perhaps more interestingly, the VarLearn approach predicts that similar unambiguous data advantages ought to lead to similar ages of acquisition. This then allows more precise predictions about what ages we ought to observe children acquiring certain word order options. More generally, Table 2 shows existing VarLearn child input analyses for several word order phenomena (see Yang (2012) and Pearl (in press) for more discussion about these individual word order phenomena).

As a concrete example, consider "pro-drop", which allows the optional omission of subjects. English isn't a language like this – while English speakers do sometimes leave out subjects in conversational speech (e.g., *Speaker 1: "Are you going?" Speaker 2: "Headed out now."*), the basic usage is that English speakers have to include the subject. This is why (unlike languages like Spanish and Italian), English speakers use what are called *expletive subjects*, which are subjects that aren't contentful; some examples of expletive subjects are the *it* in *It's raining* and *It seems that a penguin is on the ice*. In both cases, the "it" isn't referring to anything, the way the pronoun "it" typically does (e.g., *It's a penguin*, *Look what it's doing*). Instead, the "it" appears because English requires the subject to be there as a default, whether the subject refers to anything or not. Hence, English uses expletive subjects. So, expletive subjects serve as an unambiguous signal that English is not a pro-drop language that can optionally drop its subjects. The VarLearn analysis by Yang (2012) suggested that expletive subjects (unambiguously signalling -pro-drop) had a 1.2%

advantage in children's input over any -pro-drop signals (shown in Table 2). Notably, this is the same unambiguous data advantage for +V2 movement in both German and Dutch (i.e., 1.2%). When we look at the observed age of acquisition, -pro-drop in English – just like +V2-movement in German and Dutch – appears to be acquired around age 3. So, the same unambiguous data advantage (1.2%) seems to correlate with the same observed age of acquisition for these two word order phenomena.

Table 2: The relationship noted by Yang (2012) between the unambiguous data advantage (Adv) perceived by a VarLearn child in her input and the observed age of acquisition (AoA) in children for six word order parameter values across different languages.

Param Value	**Language**	**Unambiguous Form**	**Unambiguous Example**	**Adv**	**AoA**
+*wh*-movement	English	*wh*-movement 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*	*Jack voit souvent Lily* *Jack sees often Lily* "Jack often sees Lily"	7%	1;8
-pro-drop	English	expletive subjects	It seems a penguin is on the ice.	1.2%	3;0
+V2	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

This means that the VarLearn approach has the potential to generate fairly specific predictions about age of acquisition, on the basis on the unambiguous data advantage a VarLearn child would perceive in her input. So, for any language and any word order linguistic parameter, we need to decide what the unambiguous data would be for the parameter value of the language (e.g., +V2 or -pro-drop) as well as the unambiguous data for any alternative parameter values (e.g., -V2 or +pro-drop). I should note that this is by no means trivial – what counts as unambiguous very much depends on what the competing options are for the parameter in question, as well as what *other* word order parameters in the language may obscure the target value's observable signature in the input. For instance, consider that the unambiguous signal for +V2 movement involved the order *Object Verb Subject* but not *Subject Verb Object* – this is because *Subject Verb Object* could be generated by -V2 combined with an SVO basic word order. Still, with a concrete idea of what unambiguous data are for each parameter value under consideration, we can calculate how much unambiguous data the child would perceive for the target value vs. the other values, and so calculate the unambiguous data advantage perceived by the child for the target value.

Once we know the unambiguous data advantage for the target word order parameter values (in either the same language or across several languages), we then know their predicted relative acquisition trajectory: those with a higher unambiguous data advantage should be acquired earlier.

If we have enough of this kind of data, we may also be able to triangulate on a specific expected age of acquisition for any given parameter value. Parameter values with similar unambiguous data advantages are predicted to have similar observed ages of acquisition, like +V2 movement in German and -pro-drop in English.

3.2.3 Prediction evaluation

As mentioned above, from the available VarLearn analyses shown in Table 2, it seems that current predictions (both relative and absolute) are borne out. Of course, there are many more word order aspects that can be captured by linguistic parameters and many more languages where VarLearn analyses are yet to be done. The VarLearn approach would be supported any time the unambiguous data advantage aligns with the relative order order of acquisition; if the unambiguous data advantage also allows us to pinpoint a specific age of acquisition, then the VarLearn approach would be supported whenever that predicted age of acquisition is in fact observed.

In contrast, the VarLearn approach wouldn't be supported any time the unambiguous data advantage doesn't align with the relative order of acquisition or doesn't predict the observed age of acquisition. I do note that supporters of the VarLearn approach might then argue that the data considered unambiguous for the target parameter value might be the issue, rather than giving up on the VarLearn approach altogether (that is, the calculated unambiguous data advantage was incorrect). However, the burden of proof would be on those supporters to identify plausible unambiguous data that would lead to the appropriate unambiguous data advantage.

3.3 Inflectional morphology

For languages that have a rich inflectional morphology system, children need to learn how to indicate features like verb tense, person, and number, as well as noun case. Even for languages with sparser morphology (like English), children still need to learn to indicate some subset of features using morphology (e.g., past tense). Whether rich or sparse, morphology systems are harder to learn the more irregular they are – that is, the more exceptions there are to the default rule. This is because the default morphological rule(s) may well get obscured in children's input when there are many exceptions. So, a core aspect of morphological acquisition is how children figure out their morphology systems, particularly in the presence of exceptions.

3.3.1 Specific UG+stats proposals

The Tolerance and Sufficiency Principles. An approach to morphology acquisition proposed by Yang (2005, 2016) involves the Tolerance and Sufficiency Principles (**TolP+SuffP**), and has been used to account for the acquisition of a variety of semi-regular morphology in both English and German. More specifically, the TolP+SuffP learner identifies whether a morphological affix is productive, and so is applied to new word forms, or whether the affix is restricted to a certain subclass of words in the language (i.e., an exception to the productive rule). In English, this approach has been used to identify productive morphology for the past tense (+*ed* default: *kiss-kissed*), noun plurals (+*s* default: *penguin-penguins*), and derivational morphology (e.g., productive = *-ness*,

cute-cuteness; *-ment*, *enjoy-enjoyment*; *-er*, *teach-teacher*; *-ity*, *stupid-stupidity*; unproductive = *-age*, *pack-package*, *-th*, *true-truth*). In German, this approach has been used to identify productive noun plural morphology when the nouns have certain properties, such as a certain grammatical gender (e.g., being +feminine), a certain phonological property (e.g., a reduced final syllable), or a certain morphological property (e.g., being monosyllabic). When the nouns don't fit in any of these specified classes, the TolP+SuffP learner can also identify *-s* (*Auto-Autos*) as the productive plural, despite its infrequency.

The general approach a TolP+SuffP learner takes is to monitor the morphological forms in her input, and on the basis of that input, hypothesize potential rules that might be productive (e.g., for the English past tense, *+ed* and alternatives like "word rime becomes /ɔt/", as in *catch-caught* and *buy-bought*). Then, the TolP+SuffP learner identifies the relevant domain where these potential rules could apply (e.g., all English verbs for the English past tense). The learner then uses the Tolerance and Sufficiency Principles to identify how many exceptions a productive rule can tolerate while still being productive; if there are sufficient rule-following words (i.e., the exceptions are fewer than the specified number that a productive rule can tolerate), the TolP+SuffP learner identifies that rule as the productive rule for that domain. This process is done for every potential rule. Importantly, only one potential rule could be the productive rule, because of the implementation of the Tolerance and Sufficiency Principles – a productive rule requires a majority of the words that could obey it to actually obey it. (See Yang (2016) and Pearl (in press) for more detailed discussion on exactly why this is.) So, after this evaluation process, a TolP+SuffP learner could either (i) identify one of the potential rules (i.e., morphological affixes) to be productive within the specified domain of words, or (ii) identify that none of the potential rules are productive (and so there is no productive morphological affix for that domain of words).

The UG part. For the TolP+SuffP learner, it might be argued that these innate principles (i.e., the Tolerance and Sufficiency Principles) are language-specific, as they're derived from considerations of linguistic item storage and retrieval in real time (see Yang (2016) for discussion of this perspective).

The statistics part. The Tolerance and Sufficiency Principles operate over counts of relevant items (i.e., how many words obey a potential rule vs. how many are exceptions to that rule).

Biases with Bayesian inference. Another approach to morphology acquisition involves the child having biases to heed certain types of information over others in her input, combined with Bayesian inference (**Biases+Bayes**); this approach has been used by Gagliardi and colleagues (Gagliardi & Lidz, 2014; Gagliardi, Feldman, & Lidz, 2017) for morphology acquisition in Tsez. More specifically, Tsez has classes of nouns that take on different agreement morphology; these classes can be defined by different properties, including semantic properties like male vs. female or +/- animate, and phonological properties like the identity of the first or last segment of the noun. However, sometimes there doesn't seem to be a general property that holds of the nouns in the noun class – it's simply that this collection of nouns takes the same agreement morphology, and that's how we know these nouns belong to the same class.

The Biases+Bayes learner uses Bayesian inference to probabilistically aggregate the information from a specific noun and determine which class it belongs to for the purpose of agreement morphology. This is particularly useful when information from the noun conflicts – for example, perhaps the semantic information corresponds to one noun class while the phonological information corresponds to a different class. When Bayesian inference isn't tempered by biases to preferentially heed certain types of information, the learner simply relies on the statistically strongest information (e.g., the information that had the strongest correlation to a particular noun class in the child's input.) So, for example, consider that a semantic feature of +animate predicts the first noun class with probability 1 (i.e., $P(class_1|+animate)=1$) while a final segment of *-i* predicts the fourth noun class with probability .54 (i.e., $P(class_4|-i)=0.54$. Then, a Tsez word that refers to something animate but ends with an *-i* would be put in the first agreement class by unbiased Bayesian inference, because $P(class_1|+animate)>P(class_4|-i)$.

In contrast, if the child has biases, then Bayesian inference operates over whatever data the child extracts from her input, given those biases. Gagliardi and colleagues find that child agreement morphology behavior is best captured by a bias to preferentially attend to phonological information over semantic information. So, in our example above, if a Tsez word refers to something animate and ends with an *-i*, Tsez children seem to ignore the semantic component of animacy and instead focus on the final segment being *-i*. This causes them to place it in the fourth agreement noun class, rather than the first one. This behavior can be captured by Bayesian learners that ignore (or otherwise misperceive) semantic features some large percentage of the time (61-95%, depending on the exact implementation of ignoring or misperception).

The UG part. The origin of the proposed bias for phonological information over semantic information in Tsez children is unknown. It's possible that this bias is derived from the child's experience that phonological information has previously been more reliable. However, it's also possible that this bias is based on innate knowledge to heed information in the same linguistic domain, which would make this part of UG. In this case, that innate knowledge would mean that when a child is trying to learn about morphology information that's expressed with certain phonology (as the agreement affixes are), the child should pay (more) attention to morphophonological information (like phonological segments in the word) than other types of information (like animacy of the word's referent).

The statistics part. The Biases+Bayes learner uses Bayesian inference to learn which agreement morphology different Tsez nouns take.

3.3.2 Predictions made

TolP+SuffP. As mentioned above, a TolP+SuffP approach is able to capture the correct qualitative result for several cases of semi-regular morphology in English and German – that is, a TolP+SuffP child can identify the correct generalization for productive morphology. More generally, if a child acquires a productive rule for some piece of morphology, we would expect to see application of that morphology to new words that fall within the relevant domain. For example,

once the child acquires the *-ed* morphology rule for the English past tense, we would expect to see new words in the past tense with the *-ed* form (e.g., *Jack wugs today. He wugged yesterday.*). In fact, a productive rule might cause overregularizations in semi-regular systems where there are exceptions, but the child hasn't learned what all the exceptions are yet (e.g., *drink-drinked, go-goed*). We see both these kind of child outputs in English and German, as discussed by Yang (2016).

Similarly, we can consider lexical gaps, where certain forms with inflectional morphology don't seem to exist for adults. Some examples are the past participle of *stride* in English (*Jack has *stridden.*), the first person singular in the present tense of *abolish* in Spanish (**abuelo = I abolish*), and the first person singular of non-past verbs like *win* in Russian (**pobežu/*pobeždu = I win*). When asked to create these forms, adults in these languages don't quite know what to do because the relevant morphology isn't productive for that domain of words. Yang (2016) demonstrates how a TolP+SuffP learner can fail to identify a productive morphological rule in these cases.

However, we have yet to see precise predictions about exactly what age TolP+SuffP children should identify that certain morphology is productive (or not). In cases where the morphology is in fact productive, we might expect that the recognition of productivity depends on how frequently the individual words in the relevant domain appear in the child's input. The more often they do, the more likely the child is to notice them and be able to make the correct generalization using the TolP+SuffP approach. Importantly, applying the TolP+SuffP approach means the child has to also identify the relevant domain where the morphology would be productive, and it's unclear that we have precise predictions about when this would happen (or really, what might trigger this to happen).

In cases of lexical gaps where morphology isn't productive, we face a similar problem of not knowing precisely what age a child ought to figure out that there isn't a productive morphological rule for some domain of words. However, given that the target state at the end of development is the lack of a productive rule, we can at least see if children's input over time would lead a TolP+SuffP learner to decide there isn't a productive morphologial rule. What might be especially interesting is if a child's input could lead a TolP+SuffP learner to the temporary belief that there is in fact a productive rule, and we see evidence of that temporary state in children's behavior (either through application to novel words in the domain, or overregularization).

What's in common for generating more precise predictions about children's age of acquisition for morphology under the TolP+SuffP approach is a more incremental application of this approach to children's input. That is, we need to understand whether a TolP+SuffP child would predict a specific morphological affix to be productive when given realistic child input from specific ages (e.g., up to 12 months vs. 12-18 months vs. 18-24 months, and so on). With that kind of analysis, we would have precise predictions about whether a child of a particular age in a particular language should perceive a particular affix as productive or not. Then, we can assess whether these predictions are borne out in child linguistic behavior.

Biases+Bayes. At least one implementation of the Biases+Bayes approach seems to get the correct result with respect to Tsez children's observed behavior – in this case, matching rates of child production or comprehension of agreement morphology for novel nouns with certain semantic or

phonological features (Gagliardi et al., 2017). This seems to bode well for the Biases+Bayes approach, though it has yet to be applied to other morphological systems. If other morphological systems (whether in Tsez or other languages) also can be learned as children seem to learn them with a bias for phonological features over semantic features, in combination with Bayesian inference, then this would be support for this *particular* Biases+Bayes approach to learning morphology. If not, then it may be the case that other biases might be able to capture children's morphology learning behavior, in combination with Bayesian inference. If it turns out other biases work well for other systems, then the origin of these biases comes into question (in particular, whether they're likely to be UG – if so, then that Biases+Bayes proposal is still a UG+stats proposal).

3.3.3 Prediction evaluation

As mentioned above, it seems like a TolP+SuffP learner can get the right adult morphological generalizations for certain cases of semi-regular morphology in English and German. However, we don't yet have precise predictions about the expected age of acquisition for these generalizations, given children's input. It also seems like a Biases+Bayes learner can generate the right child morphological acquisition behavior for a specific case of Tsez agreement morphology when the bias is for phonological information over semantic information. However, we don't yet know how robust this particular bias is for learning morphology of other kinds in Tsez or in other languages. So, for both these UG+stats approaches, it seems that the way forward is to look for other morphology systems, especially semi-regular ones where there are exceptions and/or probabilistic associations of different types of information. Then, we can apply these approaches to the acquisition of those morphology systems to generate predictions about how acquisition ought to proceed, given realistic child input data.

3.4 A temporary lack of inflection

In many languages that have relatively less inflectional morphology (e.g., those shown in Table 3), children go through a stage where they seem to systematically leave off obligatory inflection on verbs. So, the verb appears to be in the non-finite (infinitive) form, where tense is missing. This stage is sometimes called the *optional infinitive* (**OI**) stage, as children optionally use what seems to be the infinitive form of the verb, instead of the appropriate inflected form.

Table 3: Optional infinitive examples in child-produced speech in different languages, and their intended meaning.

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

For example, in English, a child might want to express the idea that her father has something – the target form *Papa has it* is expressed as *Papa have it*, where the verb *have* is missing the 3rd

person singular present morphology. In Hebrew, a child might express the target form involving the present tense of *sit* by using the infinitive equivalent (*Lashevet - to sit*), which has additional morphology clearly indicating that the child used the infinitive form with infinitive morphology, rather than a root form with no morphology. This is also the case in Dutch, French, and German, where the form the child uses has clear infinitive morphology (e.g., *drinken-to drink* in Dutch, *dormir-to sleep* in French, and *hintelln-to put* in German from Table 3). Moreover, in these languages, the use of the infinitive is often accompanied by a word order that's appropriate for the infinitive form of the verb but not for the inflected form.

Interestingly, children's frequency of OIs 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). So, from an acquisition perspective, we want to understand why children across the world's languages show the amount of OI use that they do and how they break out of this stage to reach the adult use (which doesn't involve these OIs).

3.4.1 Specific UG+stats proposals

The variational learning approach. Legate and Yang (2007) propose a VarLearn approach to explain the different rates of OIs in child-produced speech, with the idea that children are relying on a linguistic parameter that determines whether their language is one that uses tense morphology (+Tense) or not (-Tense). +Tense languages like English, Hebrew, Dutch, French, and German express tense morphosyntactically (e.g., English *has*=*have*$_{present+3rd+sg}$); -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."). The OI stage of a +Tense language happens because children think the correct parameter value for their language is -Tense. As children perceive more unambiguous +Tense data in their input, the +Tense grammar is rewarded and the *-Tense* grammar generating the OIs is penalized until it's no longer active. How fast this happens depends on how many more unambiguous *+Tense* data are available than unambiguous -Tense data (i.e., the +Tense unambiguous data advantage).

The UG part. The Tense linguistic parameter is meant as UG knowledge, and children need to both know this parameter exists and that it has two values (+/-Tense).

The statistics part. As with the VarLearn approach for word order, reinforcement learning forms the basis of the learning mechanism.

3.4.2 Predictions made

As mentioned above, the VarLearn child is driven by the unambiguous data advantage she perceives from her input. So, for any given language, the perceived unambiguous data advantage for +Tense can be calculated. Then, unambiguous +Tense data advantages can be compared across languages for a relative order of acquisition. In particular, higher +Tense advantages indicate a shorter OI stage. Moreover, if the length of the OI stage is known for a specific language (i.e., what age

children leave the OI stage), the +Tense advantage can be correlated with that age. Similar +Tense advantages predict similar ages when children leave the OI stage.

3.4.3 Prediction evaluation

Legate and Yang (2007) use the VarLearn approach to analyze the perceived unambiguous data advantage for +Tense in Spanish, French, and English children (who are all learning *+Tense* languages); they 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 the Spanish rate of OI production < French OI production < English OI production. This in turn suggests that the OI stage for Spanish < French < English (i.e., the stage in English lasts the longest), and this seems to be true. 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.

Still to do is to evaluate the VarLearn approach on other languages where children have OI stages, such as Hebrew, Dutch, and German. I should also note an important caveat – an alternative non-UG+stats account for investigating OIs called MOSAIC (*Model of Syntactic Acquisition in Children*) has already been applied to a large number of languages (Freudenthal, Pine, Aguado-Orea, & Gobet, 2007; Freudenthal, Pine, & Gobet, 2009, 2010; Freudenthal, Pine, Jones, & Gobet, 2015), including those that the VarLearn approach has been applied to. (See Pearl (in press) for more discussion about the MOSAIC approach.) MOSAIC is also able to account for the different cross-linguistic rates of OIs in children, and additionally offers an explanation as to why certain specific verbs appear with OI errors. Currently, the VarLearn approach doesn't offer the same ability to explain OI errors with specific verbs in these languages. So, for this reason, the non-UG+stats MOSAIC account may be preferable for now to the VarLearn approach when it comes to OIs.

3.5 Movement: When things are understood in places where they don't seem to be

A more sophisticated type of syntactic knowledge involves "movement", where linguistic elements are understood in certain positions of an utterance and yet don't appear to be in those positions. So, the idea is that the linguistic elements have moved from the positions where they're understood. Some examples of this are *wh*-movement in questions, passives, and raising vs. control structures. I'll talk briefly about each in turn.

***Wh*-movement.** As we saw in the section of basic word order and Table 2, English is a *+wh*-movement language. In questions, like *What did the penguin do?*, the *wh*-word *what* is at the front of the question, but is interpreted as the object of the verb *do*. So, this question seems to have a structure more like *What did the penguin do* $\__{what}$*?*, where $\__{what}$ marks the position where *what* has moved from. Children then need to learn both that the *wh*-element moved and, more importantly, where it moved from so it can be interpreted correctly. English children seem to figure out pretty

quickly (earlier than 1 year 8 months) that *wh*-elements move, and also how to interpret simple *wh*-movement examples (such as *What did the penguin do?*).

Passives. In passive structures like *Jack was kissed*, the subject *Jack* is understood as the one being kissed (i.e., the thematic PATIENT), rather than the one doing the kissing (i.e., the thematic AGENT). So, this sentence seems to have a structure more like *Jack was kissed* $\__{Jack}$, where $\__{Jack}$ marks the position where *Jack* was moved from. Children then need to learn that this interpretation is possible, which involves understanding where the element in the subject position moved from (in this case, from a position where it can serve as PATIENT). Moreover, there are restrictions on which verbs allow passivization – for example, the active sentence *Jack stayed here* can't be passivized into **Here was stayed by Jack*. So, children also need to identify which verbs allow the passive structure. In English, there seems to be significant variation for the age of acquisition of the passive, with the passives of some verbs acquired as young as three while others appear delayed till age nine (de Villiers & de Villiers, 1973; Maratsos, 1974; Maratsos & Abramovitch, 1975; Pinker, 1987; Gordon & Chafetz, 1990; Fox & Grodzinsky, 1998; Hirsch & Wexler, 2006; O'Brien, Grolla, & Lillo-Martin, 2006; Crain, Thornton, & Murasugi, 2009; Messenger, Branigan, McLean, & Sorace, 2009, 2012; Orfitelli, 2012; Liter, Huelskamp, Weerakoon, & Munn, 2015; Nguyen, Lillo-Martin, & Snyder, 2016).

Raising vs. control structures. In subject-raising structures like *Jack seemed to kiss Lily*, the subject of the main clause *Jack* doesn't appear to have an AGENT thematic role with respect to the main clause verb *seem* – that is, *Jack* isn't a "seemer" (whatever that is). Instead, *Jack* is the AGENT of *kiss*, which is the embedded clause verb. That's why this utterance can be rephrased as *It seemed that Jack kissed Lily*, which has an expetive *it* as the main clause subject and *Jack* overtly as the embedded clause subject. So, the original sentence seems to have a structure more like *Jack seemed* $\__{Jack}$ *to kiss Lily*, where $\__{Jack}$ marks the position where *Jack* moved (or "raised") from.

Subject-raising structures contrast with subject-control structures like *Jack wanted to kiss Lily* – here, the main clause subject *Jack* seems to take on two thematic roles: the AGENT of main clause verb *wanted* and the AGENT of embedded clause verb *kiss*. (This is why we can't rephrase this utterance as **It wanted that Jack kissed Lily* – expletive *it* can't be the AGENT of *wanted*.) Because traditional linguistic theory disliked linguistic elements having more than one thematic role, a solution was for this utterance to have a structure more like *Jack wanted PRO to kiss Lily*, where *Jack* is connected to the silent pronoun *PRO*; this allows *Jack* to be the AGENT of *wanted* while *PRO* is the AGENT of *kiss*. So, unlike raising structures, there's no movement associated with control structures. Instead, the child has to recognize the connection between the main clause subject and the silent pronoun *PRO*.

The same raising vs. control distinction also happens for objects – that is, there are object-raising verbs and object-control verbs. In object-raising structures like *Jack wanted Lily to laugh*, the main clause object *Lily* is only the AGENT of the embedded clause verb *laugh*, rather than also having a thematic role with respect to main clause verb *want*. So, the structure is something like *Jack wanted Lily* $\__{Lily}$ *to laugh*, with *Lily* raised from the embedded clause position. In contrast, in object-control structures like *Jack asked Lily to laugh*, the main clause object *Lily* takes two

thematic roles: the AGENT of embedded clause verb *laugh* and the GOAL of main clause verb *asked*. So, the structure is something like *Jack asked Lily PRO to laugh*, with *Lily* and *PRO* connected to each other.

For raising and control verbs, children therefore need to learn that these interpretations are possible (i.e., the main clause subject or object effectively gets associated with either one thematic role or two). This involves learning where the main clause subject or object moved from (raising) or that the main clause subject or object is connected to the silent *PRO* in the embedded clause (control). Moreover, children need to identify which verbs allow which types of structures (e.g., *seem* is a subject-raising verb, *want* is a subject-control verb and also an object-raising verb, and *ask* is a subject-control verb and also an object-control verb). Current behavioral evidence suggests that English four- and five-year-olds have these interpretation options available and have sorted some frequent raising and control verbs into relevant classes that allow adult-like interpretation of these verbs (Becker, 2006, 2007, 2009; Kirby, 2009a, 2009b, 2010; Becker, 2014).

3.5.1 Specific UG+stats proposals

***Wh*-movement.** The VarLearn approach (Yang, 2002, 2004; Legate & Yang, 2007; Yang, 2012) discussed previously is the UG+stats proposal that I'm aware of. In this case, reinforcement learning operates over a *wh*-movement linguistic parameter with two options (+/-*wh*-movement).

The UG part. A *wh*-movement linguistic parameter would be innate, language-specific knowledge. In particular, the limited range of options specified as possible movement options (i.e., "move to the front of the relevant structural unit (like the main clause)" or "don't move") would be UG knowledge. This contrasts with allowing a range of other seemingly plausible options, like "move to the first position, irrespective of whether that corresponds to a clause" or "move to the third position" or "move in front of the first noun phrase".

The statistics part. Reinforcement learning forms the basis of the VarLearn learning mechanism.

Passives. For how children learn to identify where the main clause subject moved from in passives, I'm unaware of UG+stats proposals. However, there's a possible UG+stats proposal for how children identify which verbs allow passivization; this proposal involves children attending to specific lexical semantic features in their input and using Bayesian inference to determine which verbs allow passivization (Nguyen & Pearl, 2019) – I'll call this the **Lex+Bayes** approach. In particular, children decide whether a verb is likely to allow passivization based on (i) its lexical semantic features, and (ii) the input encountered about how often verbs with those lexical semantic features allow passivization. Lexical semantic features include whether a verb is "actional" like *eat*, and so can be easily observed, and whether the verb is "object-experiencer" like *annoy*, so that the object of a transitive use (e.g., *Jack* in *The giant annoyed Jack*) experiences the mental or emotional state the verb describes (see Nguyen and Pearl (2018, 2019) for more details on the lexical semantic features currently being investigated).

The UG part. The Lex+Bayes approach is agnostic as to why the child considers any given set of lexical semantic features (the ones investigated by Nguyen & Pearl were derived empirically from prior behavioral work, rather than being theoretically-motivated). However, if it turns out that the set of lexical semantic features is constrained innately in a language-specific way, this would be UG knowledge.

The statistics part. The learning mechanism for Lex+Bayes is Bayesian inference.

Raising vs. control. The potential UG+stats approaches I'm aware of involve children attending to certain features of verbs and their arguments (e.g., whether the subject is animate, or what syntactic contexts a verb can appear in), and then using Bayesian inference to cluster together verbs that behave the same way with respect to these features (Mitchener & Becker, 2010; Becker, 2014; Pearl & Sprouse, in press). For instance, verbs that take inanimate subjects are more likely to be subject-raising verbs (e.g., *The rock seemed to fall* (*seem* is subject-raising) vs. **The rock wanted to fall* (*want* is subject-control)). The approach of Becker and Mitchener (Mitchener & Becker, 2010; Becker, 2014) focuses primarily on the animacy of the subject, while the approach of Pearl and Sprouse (in press) considers the animacy of all verb arguments, the thematic roles the verb arguments take (e.g., whether the subject is an AGENT or a THEME), and the syntactic contexts a verb can appear in (e.g., a transitive frame like *Jack kissed Lily* or a frame that involves a non-finite embedded clause like *Jack wanted to kiss Lily*).

The UG part. In these approaches, the main place where I see a role for UG is which features children use to sort verbs into relevant classes. In particular, it could be that innate, language-specific knowledge causes children to focus on animacy when clustering verbs together into classes, as opposed to other salient conceptual features of verb arguments. The Pearl & Sprouse approach considers a wider range of verb and verb argument features than the Becker & Mitchener approach, but still restricts the range of possibilities for the thematic role distinctions and the syntactic positions that children perceive; these restrictions are based on current theoretical proposals in the syntactic literature. If these thematic role and syntactic position distinctions are innate, language-specific knowledge, then they would come from UG.

The statistics part. The learning mechanism for these approaches is Bayesian inference.

3.5.2 Predictions made

***Wh*-movement.** The VarLearn approach predicts that the unambiguous data advantage for a language's *wh*-movement option correlates with the observed age of acquisition. If the correlation is there, the VarLearn approach is supported.

Passives. The Lex+Bayes approach predicts the age when children will correctly interpret or produce the passive for verbs that have various lexical semantic features. In particular, this approach can highlight which lexical semantic features children must attend to vs. ignore in order to match

the observed age of acquisition (assuming the observed age of acquisition can in fact be matched with any combination of the lexical semantic features considered). If children's observed age of acquisition can be predicted by some combination of lexical semantic features, then the Lex+Bayes approach is supported.

Raising vs. control. The Bayesian approaches to clustering verbs into classes that allow different raising and control constructions can predict the classes that children of different ages ought to cluster their verbs into. These predicted verb classes can then be checked against behavioral data from children of different ages. For example, if children treat two verbs the same way (e.g., both verbs allowing subject-raising, but not subject-control, object-raising, or object-control), then the Bayesian approaches ought to have clustered those two verbs together into the same class. This prediction check can be done for all verbs where we have empirical data about how children treat the verbs (i.e., as belonging to the same class or not). As a concrete example from Pearl and Sprouse (in press), one model variant predicts that English five-year-olds treat *want*, *like*, and *need* as belonging to the same class, while another variant predicts only *want* and *like* belong to the same class. We can check these predictions to see if English five-year-olds treat *want, need,* and *like* the same (e.g., interpreting them as subject-control verbs that take two thematic roles in instances like *Jack wants/needs/likes to go*). If five-year-olds do treat these the same, the first model variant is supported; if they treat only *want* and *need* as the same, the second model variant is supported; if they don't treat any of these verbs the same, then no model variant is supported – and maybe different features need to be considered for verb classification.

3.5.3 Prediction evaluation

***Wh*-movement.** The prediction for a very early age of acquisition for +*wh*-movement in English appears to hold (recall Table 2 showing acquisition by 1 year 8 months for a 25% unambiguous data advantage). This is promising, and suggests that it's worth trying the VarLearn approach on other languages (both +*wh*-movement and -*wh*-movement languages). Once we have a collection of language predictions, we can see if the relative order of *wh*-movement acquisition correlates with a VarLearn approach's predictions; if we're fortunate enough that the unambiguous data advantage for a language is similar to some linguistic knowledge the VarLearn has already been used to analyze (as summarized in Table 2), then we would have a predicted age of acquisition as well – the predicted age of acquisition would be the same as the other linguistic knowledge's age of acquisition.

Passives. The Lex+Bayes approach is currently able to match existing empirical data on which verbs English children will passivize by age five. It accomplishes this by using a filter that only heeds the lexical semantic feature +object-experiencer (and also the syntactic feature of transitivity). So, this approach then predicts that English five-year-olds pay special attention to these features (and ignore the rest) when deciding if a verb can be passivized. A useful next evaluation step is to see if English five-year-olds use these features when deciding whether to passivize novel verbs – if so, then we have stronger evidence that these features are key for understanding English

children's passivization knowledge. Moreover, we can see if the Lex+Bayes approach can match existing empirical data on younger children's passivization behavior (and if so, whether they also need to pay special attention to certain verb features). More generally, this approach can be used for any language where we have data about (i) the verbs children hear in the passive form, and (ii) which verbs they passivize at different ages.

Raising vs. control. The Bayesian approaches to clustering verbs into classes that involve raising vs. control interpretations appear to match children's verb classifications fairly well (Pearl & Sprouse, in press). So, these approaches seem promising, particularly when we allow children to consider a range of features (conceptual, thematic, and syntactic). A useful aspect of a model predicting verb classes is that we have a variety of ways to evaluate if children in fact have similar verb classes. One way is what's been done already – derive children's verb classes from their aggregated behavioral data and compare those against the model's verb classes. However, another way is to use the model's predicted verb classes to predict child behavior in specific experiments. For instance, given a specific context (i.e., animacy of the verb arguments, thematic roles of the verb arguments, and syntactic context of the verb), what's the probability that a child will interpret a novel verb as raising vs. control? This quantitative prediction about interpretation rate can be compared against the rates children actually do interpret a verb a particular way in context. Becker and Kirby (Becker, 2006, 2007, 2009; Kirby, 2009a, 2009b, 2010; Becker, 2014) have already conducted several behavioral experiments like these that can provide precise testing grounds for these Bayesian approaches.

3.6 Constraints: Things we just can't say

Another more sophisticated type of syntactic knowledge involves "constraints" – constraints disallow certain structures (and their accompanying interpretations), rather than specifying which structures are allowed. Two prominent examples of constraints investigated by UG+stats proposals are *syntactic islands* (sometimes called *subjacency*) and *binding*. I'll talk briefly about each in turn.

Syntactic islands: Constraints on *wh*-dependencies. As we saw with *wh*-movement, a *wh*-word appears at the front of a question in English. The relationship between the overt position of the *wh*-word and where it's understood can be called a *dependency*, and so (3a) shows a *wh*-dependency between *What* and where it's understood at the position marked by $\_what$. It turns out that there are constraints on the *wh*-dependencies that are allowed – one way to think about this is that there are certain structures called *syntactic islands* that *wh*-dependencies can't cross (Chomsky, 1965; Ross, 1967; Chomsky, 1973). Four examples of syntactic islands in English are shown in (3b)-(3e), with the proposed syntactic island structure in square brackets ([...]). During acquisition, English children have to learn the constraints on *wh*-dependencies that allow them to recognize that the *wh*-dependencies in (3b)-(3e) aren't allowed, while the *wh*-dependency in (3a) is fine.

(3) a. What does Jack think that Lily said that the goblins stole $\_what$?

b. *What do you wonder [whether Jack bought $\__{what}$]? (whether island)
c. *What did you make [the claim that Jack bought $\__{what}$]? (complex NP island)
d. *What do you think [the joke about $\__{what}$] was hilarious? (subject island)
e. *What do you worry [if Jack buys $\__{what}$]? (adjunct island)

Binding: Constraints on interpreting referential elements (like pronouns). When interpreting pronouns, the type of pronoun affects what interpretations are possible. This is because a pronoun is often "bound to" (i.e., associated with) a linguistic antecedent (i.e., another linguistic element in the utterance), and this binding depends on structural constraints that vary based on the type of pronoun. In particular, "plain" pronouns like *her* in English are interpreted differently than reflexive pronouns like *herself*, as we can see in (4). More specifically, when we don't know what type of pronoun the object of *admire* is, as in (4a), either *Lily* or *Sarah* seems like a plausible antecedent. However, if we have the plain pronoun *her* as in (4b), then the antecedent can't be *Lily* – it must be someone else, which could be *Sarah*. In contrast, if we have the reflexive pronoun *herself* as in (4c), then the antecedent must be *Lily* and can't be *Sarah*.

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

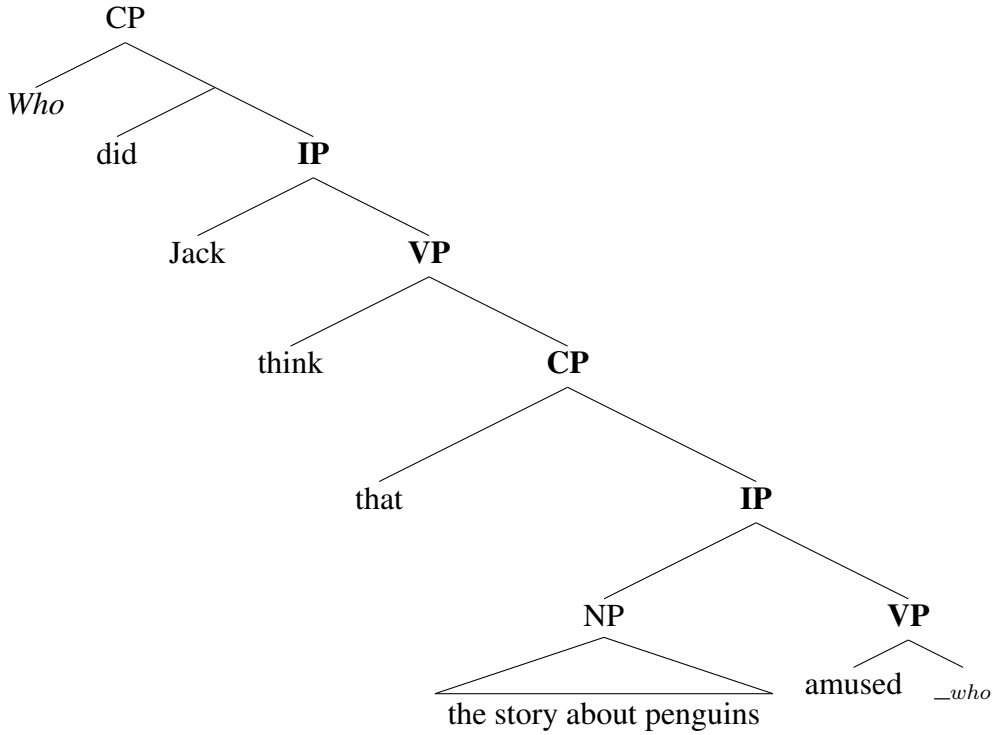
So, plain pronouns and reflexive pronouns appear to have a complementary distribution when it comes to which linguistic antecedents they can take, which in turn affects how they're interpreted. Importantly, the potential positions that their antecedents can appear in (for reflexive pronouns) or can't appear in (for plain pronouns) are structurally defined (Chomsky, 1973). So, during acquisition, children have to learn the structurally-based constraints on pronoun interpretation that allow them to recognize the appropriate interpretations for utterances like (4b)-(4c). Children also have to learn which pronouns correspond to which constraints (i.e., which pronouns are plain vs. reflexive in English).

3.6.1 Specific UG+stats proposals

Syntactic islands. Pearl and Sprouse (2013a, 2013b, 2015) investigated a probabilistic learning strategy that relies on trigrams (i.e., sequences of three elements) constructed from certain pieces of syntactic structure involved in *wh*-dependencies. So, we can think of this as a probabilistic *syntactic trigrams* approach (**SynTrigrams**). The SynTrigrams strategy relies on children viewing a *wh*-dependency as a path from the head of the dependency (e.g., *Who* in (5)) through the phrasal nodes that contain the tail of the dependency, as shown in (5a)-(5b). So, a SynTrigrams child just needs to learn which *wh*-dependencies have grammatical syntactic paths and which don't. The SynTrigrams child does this by tracking smaller building blocks of these syntactic paths – the syntactic trigrams. More specifically, a SynTrigrams learner breaks the syntactic path of a *wh*-dependency into a collection of syntactic trigrams that can be combined to reproduce the original syntactic path, as shown in (5c).

(5) Who did Jack think that the story about penguins amused $\__{who}$?

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 $\__{who}$]]]]]?

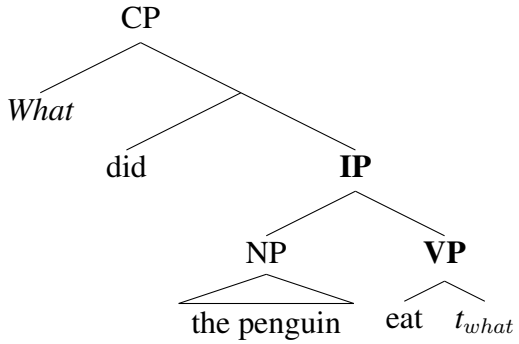
b. Syntactic path of *wh*-dependency:
start-IP-VP-CP$_{that}$-IP-VP-*end*

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 SynTrigrams child then tracks the frequencies of syntactic trigrams that the child perceives in her input. Importantly, every instance of a *wh*-dependency is composed of some set of syntactic trigrams, so a child can potentially learn about a specific syntactic trigram (e.g., *start*-IP-VP) from a variety of *wh*-dependencies. That is, the building blocks of a particular *wh*-dependency syntactic path can come from other *wh*-dependencies, not just that particular *wh*-dependency. The SynTrigrams child can later use the syntactic trigram frequencies to calculate the probability of any *wh*-dependency she likes, whether she's encountered it before or not; this is because all *wh*-dependencies can be broken into syntactic trigram building blocks, and the child has a sense from her input of how probable any particular syntactic trigram is, based on its frequency in her input. For example, the *wh*-dependency in *What did the penguin eat* $\__{what}$*?* can be characterized as in (6), and its probability generated from some of the same syntactic trigrams observed in (5).

(6) What did the penguin eat $\__{eat}$?

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



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

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 predicted probability of a *wh*-dependency's syntactic path corresponds to the grammaticality of the dependency, with higher probabilities indicating more grammatical dependencies. These predictions can then be compared to judgments of how allowable different *wh*-dependencies are.

The UG part. A key component of the SynTrigrams approach is what elements the trigrams are constructed from. In the implementation by Pearl and Sprouse (2013b, 2013a, 2015), the elements are the phrasal nodes that contained the *wh*-dependency. How the child determines what these nodes are (e.g., the labels CP$_{that}$ or VP) is currently unknown. It could be that this kind of phrasal structure representation requires the child to rely on innate, language-specific knowledge; if so, this would be UG knowledge.

The statistics part. The SynTrigrams learner relies on tracking the frequencies of syntactic trigrams, converting these frequencies to probabilities, and combining these probabilities into a single probability for any *wh*-dependency's syntactic path.

Binding. Orita, McKeown, Feldman, Lidz, and Boyd-Graber (2013) propose a UG+stats approach to binding that involves children being sensitive to the structurally-defined positions where a pronoun's linguistic antecedent could be. That is, when perceiving input like (4b) and (4c), children focus on both *Lily* and *Sarah* as potential linguistic antecedents and consider the structural positions these potential antecedents are in, where *Lily* is in one type of structurally-defined position while *Sarah* is in a different type of structurally-defined position. Children's acquisition task is then about identifying classes of pronouns whose true linguistic antecedents appear in certain positions. 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? The approach by Orita et al. (2013) relies on Bayesian inference to identify the pronoun classes.

Once children know their pronoun classes, they can then interpret any pronoun they encounter – they simply identify which pronoun class the pronoun belongs to, and consider the antecedents allowed by that pronoun class. As a concrete example, let's consider *Lily, who adores Sarah, admires herself in the mirror.* A child using this learning approach would recognize that *herself* is a specific pronoun class (which 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 linguistic antecedent of *herself*.

One interesting problem this learning approach tackles is how children get started. In particular, children are sorting pronouns into pronoun classes based on the true linguistic antecedents of these pronouns – but without knowing what class a pronoun belongs to, how do children know what the true linguistic antecedent is? For instance, let's consider *Lily, who adores Sarah, admires* PRONOUN *in the mirror*. Without knowing if PRONOUN is *herself* or *her*, we don't immediately know whether the true linguistic antecedent must be *Lily* (*herself*) or can't be *Lily* (*her*).

Orita et al. (2013)'s insight was that contextual cues – in particular, discourse cues – may indicate the true antecedent for PRONOUN. 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 discourse context available. For example, if the utterance *Lily, who adores Sarah, admires* PRONOUN *in the mirror* is embedded in a conversation that made it clear Lily tends to do a lot of gazing at her own reflection, the discourse context could allow a child to infer that PRONOUN refers to *Lily*. Because this UG+stats learning approach relies on discourse cues and Bayesian inference to infer pronoun classes, I'll refer to it as **Discourse+Bayes**.

The UG part. A key premise for a Discourse+Bayes learner is that children view the positions of potential linguistic antecedents according to particular structurally-defined properties (e.g., c-command). If these properties are both language-specific and innate (rather than being derivable from non-UG knowledge and prior linguistic experience), then this would be UG knowledge. Another core component for this learner is to harness the discourse context when trying to identify the intended linguistic antecedent of a pronoun. This bias to pay attention to discourse context as a source of information may be possible to derive from a non-UG source (e.g., a domain-general preference to focus on similar information sources – here, to pay attention to other salient linguistic information, like discourse context, when trying to learn new linguistic information, like a pronoun's antecedent). However, if it's not posbble to derive this bias to heed discourse information from a non-UG source, then this bias would also be UG knowledge.

The statistics part. The Discourse+Bayes learner uses Bayesian inference to infer pronoun classes.

3.6.2 Predictions made

Syntactic islands. The SynTrigrams learner of Pearl and Sprouse (2013a, 2013b, 2015) learned from a realistic sample of English child-directed speech, estimated syntactic trigram probabilities from that sample, and then generated probabilities for a specific set of *wh*-dependencies that pre-

vious work (Sprouse, Wagers, & Phillips, 2012) had collected acceptability judgments for. More specifically, Sprouse et al. (2012) had judgments about the relative acceptability of the four syntactic island types in (3b)-(3e), as well as control *wh*-dependencies that varied with respect to their syntactic path. These judgments served as a target for the SynTrigram learner – if the SynTrigrams learner could generate the same relative judgment pattern (based on the probability the learner calculated for each *wh*-dependency), then we could conclude that the learner would have internalized a representation that's similar to what humans used to generate their judgments. If instead the SynTrigrams learner failed to generate the same relative judgment pattern for these *wh*-dependencies, then we could conclude that the representation it internalized wouldn't be similar enough to the one humans use.

Binding. The Discourse+Bayes learner of Orita et al. (2013) used a realistic sample of English child-directed speech and adult judgments to estimate the informativity of discourse context when trying to identify a pronoun's intended antecedent. From this, a discourse-based distribution over potential antecedents could be obtained for any given utterance; for example, in *Lily, who adores Sarah, admired herself in the mirror*, there might be a 68% chance that *Lily* is the intended antecedent, with the remaining 32% distributed between *Sarah* and some other person not mentioned. The Discourse+Bayes learner then learned from both this discourse-based information about potential antecedents and the structural position of potential antecedents to sort pronouns into different classes. Importantly, there was no built-in restriction that there should only be two classes – instead, the learner could make as many classes as it thought it needed (though it had a bias for fewer rather than more pronoun classes). The inferred pronoun classes can then be compared to actual pronoun classes (in particular, the two classes of reflexive vs. plain pronouns). If there are two inferred classes whose members seem to divide out appropriately into reflexive vs. plain pronouns, then the Discourse+Bayes learning approach can be viewed as successful.

3.6.3 Prediction evaluation

Syntactic islands. The SynTrigrams learner of Pearl and Sprouse (2013a, 2013b, 2015) was in fact able to replicate the observed judgment pattern that indicated knowledge of the four syntactic islands investigated by Sprouse et al. (2012). This suggests that the learning strategy of the SynTrigrams learner is a plausible way for English children to acquire knowledge of these islands. What remains to be investigated is how well this learning strategy fares cross-linguistically, as there's variation on the syntactic islands that languages seem to have (even among the four in (3b)-(3e)). Moreover, there are other types of *wh*-dependency constraints (e.g., see discussion in Pearl and Sprouse (2013a) about *wh*-dependencies with multiple gaps), and it's unknown if a SynTrigrams strategy can handle these cases as well.

Binding. The Discourse+Bayes approach of Orita et al. (2013) did in fact lead to exactly the right two pronoun classes, with a class corresponding to reflexive pronouns and a class corresponding to the plain pronouns, each with its appropriately associated linguistic antecedents. This therefore suggests that the Discourse+Bayes learning strategy is a plausible way for English children to

acquire knowledge of these two pronoun classes; knowing these two pronoun classes (and which pronouns belong to which) means English children would know how to interpret both reflexive and plain pronouns in context. As with syntactic islands, what remains to be investigated is how well this approach fares cross-linguistically, and whether this approach can be adapted to handle a wider range of binding constraints (e.g., those that involve names in addition to pronouns, as in *Lily thinks that Lily will win*, which is acceptable in Thai but not in English (Ud Deen & Timyam, 2018)).

4 Conclusion

I've reviewed several UG+stats approaches to the acquisition of different specific morphology and syntax phenomena, with the idea that these approaches make the developmental theories they implement concrete enough to evaluate. In common across nearly all these approaches is that the UG part helps determine what's being counted by the child from the vast array of information available in the input, while the statistics part determines both how the counting is in fact done and how the counts are used to update the child's hypotheses about her language's morphology or syntax. Importantly, these UG+stats proposals have been specified in enough detail to make specific predictions about child acquisition, which can then be evaluated against available empirical data or data that can be obtained in the future. In the cases I discussed, the predictions of the UG+stats proposals have generally held up – this suggests that these proposals are worth pursuing more fully, and I've also suggested possibilities for future exploration (often looking cross-linguistically or at related morphology or syntax phenomena). With this in hand, I hope we can continue making progress from the UG+stats perspective on understanding how children learn all the things they do about morphology and syntax.

References

Ambridge, B. (2017). Syntactic Categories in Child Language Acquisition: Innate, Induced, or Illusory? In *Handbook of Categorization in Cognitive Science* (pp. 567–580). Elsevier.

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1-2), e1373.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Bar-Sever, G., & Pearl, L. (2016). Syntactic categories derived from frequent frames benefit early language processing in English and ASL. In *Proceedings of the 40th Annual Boston University Conference on Child Language Development* (pp. 32–46). Somerville: Cascadilla Press.

Bates, A., Pearl, L., & Braunwald, S. (2018). I *can* believe it: Quantitative evidence for closed-class category knowledge in an English-speaking 20-to 24-month-old child. In K. Garvin and N. Hermalin and M. Lapierre and Y. Melguy and T. Scott and E. Wilbanks (Ed.), *Proceedings of the Berkeley Linguistics Society* (pp. 1–16). Berkeley, CA.

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

Becker, M. (2007). Animacy, expletives, and the learning of the raising-control distinction. *Generative approaches to language acquisition North America*, *2*, 12–20.

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

Becker, M. (2014). *The acquisition of syntactic structure: Animacy and thematic alignment* (Vol. 141). Cambridge University Press.

Boeckx, C., & Leivada, E. (2014). On the particulars of Universal Grammar: Implications for acquisition. *Language Sciences*, *46*, 189–198.

Booth, A., & Waxman, S. (2003). Mapping words to the world in infancy: On the evolution of expectations for nouns and adjectives. *Journal of Cognition and Development*, *4*(3), 357–381.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413.

Capdevila i Batet, M., & Llinàs i Grau, M. (1995). The acquisition of negation in english. *Atlantis: Revista de la Asociación Española de Estudios Anglo-Norteamericanos*, *17*(1), 27–44.

Chater, N., Clark, A., Goldsmith, J., & Perfors, A. (2015). *Empiricism and language learnability*. Oxford University Press.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'Frequent Frames': What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12*(3), 396–406.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 237–286). New York: Holt, Rinehart, and Winston.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Barriers* (Vol. 13). MIT press.

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

Crain, S., & Pietroski, P. (2002). Why language acquisition is a snap. *The Linguistic Review*, *19*, 163–183.

Crain, S., Thornton, R., & Murasugi, K. (2009). Capturing the evasive passive. *Language Acquisition*, *16*(2), 123–133.

Denison, S., Reed, C., & Xu, F. (2011). The emergence of probabilistic reasoning in very young infants. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33).

de Sousa, N. M., Garcia, L. T., & de Alcantara Gil, M. S. C. (2015). Differential reinforcement in simple discrimination learning in 10-to 20-month-old toddlers. *The Psychological Record*, *65*(1), 31–40.

de Villiers, J. G., & de Villiers, P. A. (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research*, *2*(4), 331–341.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871–1877.

Erkelens, M. (2009). *Learning to categorize verbs and nouns: studies on dutch*. Netherlands Graduate School of Linguistics.

Ferry, A. L., Fló, A., Brusini, P., Cattarossi, L., Macagno, F., Nespor, M., & Mehler, J. (2016). On the edge of language acquisition: Inherent constraints on encoding multisyllabic sequences

in the neonate brain. *Developmental Science*, *19*(3), 488–503.
Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.
Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello (2002). *Cognition*, *82*(3), 259–278.
Fló, A., Brusini, P., Macagno, F., Nespor, M., Mehler, J., & Ferry, A. L. (2019). Newborns are sensitive to multiple cues for word segmentation in continuous speech. *Developmental Science*, e12802.
Fodor, J. D. (1998a). Parsing to Learn. *Journal of Psycholinguistic Research*, *27*(3), 339–374.
Fodor, J. D. (1998b). Unambiguous Triggers. *Linguistic Inquiry*, *29*, 1–36.
Fodor, J. D. (2017). Ambiguity, parsing, and the evaluation measure. *Language Acquisition*, *24*(2), 85-99. Retrieved from https://doi.org/10.1080/10489223.2016.1270948 doi: 10.1080/10489223.2016.1270948
Fodor, J. D., & Sakas, W. G. (2005). The subset principle in syntax: Costs of compliance. *Journal of Linguistics*, *41*(03), 513–569.
Fodor, J. D., & Sakas, W. G. (2017). Learnability. In I. Roberts (Ed.), *The oxford handbook of universal grammar* (pp. 249–269). Oxford, UK: Oxford University.
Fodor, J. D., Sakas, W. G., & Hoskey, A. (2007). Implementing the subset principle in syntax acquisition: Lattice-based models. In *Proceedings of the Second European Cognitive Science Conference* (pp. 161–166). Hove, UK.
Fox, D., & Grodzinsky, Y. (1998). Children's passive: A view from the by-phrase. *Linguistic Inquiry*, *29*(2), 311–332.
Freudenthal, D., Pine, J., & Gobet, F. (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, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, *31*(2), 311–341.
Freudenthal, D., Pine, J. M., & Gobet, F. (2009). Simulating the referential properties of Dutch, German, and English root infinitives in MOSAIC. *Language Learning and Development*, *5*(1), 1–29.
Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2015). Defaulting effects contribute to the simulation of cross-linguistic differences in Optional Infinitive errors. In *Proceedings of the 37th annual meeting of the Cognitive Science Society* (p. 746-751). Pasadena.
Gagliardi, A., Feldman, N. H., & Lidz, J. (2017). Modeling statistical insensitivity: Sources of suboptimal behavior. *Cognitive Science*, *41*(1), 188–217.
Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.
Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.
Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362–366.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(4), 407–454.
Gordon, P., & Chafetz, J. (1990). Verb-based versus class-based accounts of actionality effects in children's comprehension of passives. *Cognition*, *36*(3), 227–254.
Hirsch, C., & Wexler, K. (2006). Children's passives and their resulting interpretation. In *The proceedings of the inaugural conference on Generative Approaches to Language Acquisition–North America, University of Connecticut Occasional Papers in Linguistics* (Vol. 4, pp. 125–136).
Hsu, A. S., & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, *34*(6), 972–1016.
Hsu, A. S., Chater, N., & Vitányi, P. (2013). Language Learning From Positive Evidence, Reconsidered: A Simplicity-Based Approach. *Topics in Cognitive Science*, *5*(1), 35–55.
Hsu, A. S., Chater, N., & Vitányi, P. M. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.
Hulsebus, R. C. (1974). Operant conditioning of infant behavior: A review. In *Advances in Child Development and Behavior* (Vol. 8, pp. 111–158). Elsevier.
Irani, A. (2019). How Children Learn to Disappear Causative Errors. In M. M. Brown & B. Dailey (Eds.), *Proceedings of the 43rd Boston University Conference on Language Developmen* (pp. 298–310). Somerville, MA: Cascadilla Press.
Jackendoff, R. S. (1994). *Patterns in the mind: Language and human nature*. Basic Books.
Kemp, N., Lieven, E., & Tomasello, M. (2005). Young children's knowledge of the "determiner" and "adjective" categories. *Journal of Speech, Language, and Hearing Research*.
Kirby, S. (2009a). Do what you know:"semantic scaffolding" in biclausal raising and control. In *Annual meeting of the berkeley linguistics society* (Vol. 35, pp. 190–201).
Kirby, S. (2009b). *Semantic scaffolding in first language acquisition: The acquisition of raising-to-object and object control* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill, Chapel Hill, NC.
Kirby, S. (2010). Semantic scaffolding in l1a syntax: Learning raising-to-object and object control. In *Proceedings of the 2009 mind-context divide workshop* (pp. 52–59).
Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42.
Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, *52*(2), 217–276.
Legate, J., & Yang, C. (2007). Morphosyntactic Learning and the Development of Tense. *Linguistic Acquisition*, *14*(3), 315–344.
Legate, J., & Yang, C. (2013). Assessing Child and Adult Grammar. In R. Berwick & M. Piatelli-Palmarini (Eds.), *Rich Languages from Poor Inputs* (pp. 168–182). Oxford, UK: Oxford University Press.
Lipsitt, L. P., Pederson, L. J., & Delucia, C. A. (1966). Conjugate reinforcement of operant responding in infants. *Psychonomic Science*, *4*(1), 67–68.
Liter, A., Huelskamp, T., Weerakoon, S., & Munn, A. (2015). What drives the Maratsos Effect, agentivity or eventivity? In *Boston University Conference on Language Development (BUCLD).* Boston University, Boston, MA..

Maratsos, M. P. (1974). Children who get worse at understanding the passive: A replication of bever. *Journal of Psycholinguistic Research*, *3*(1), 65–74.

Maratsos, M. P., & Abramovitch, R. (1975). How children understand full, truncated, and anomalous passives. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 145–157.

Messenger, K., Branigan, H., McLean, J., & Sorace, A. (2009). Semantic factors in young children's comprehension and production of passives. In *Proceedings of the 33rd Boston University Conference on Language Development* (pp. 355–366).

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568–587.

Meylan, S. C., Frank, M. C., Roy, B. C., & Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, *28*(2), 181–192.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Mintz, T. (2006). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). Oxford: Oxford University Press.

Mitchener, W. G., & Becker, M. (2010). Computational models of learning the raising-control distinction. *Research on Language and Computation*, *8*(2-3), 169–207.

Nguyen, E., Lillo-Martin, D., & Snyder, W. (2016). Actionality speaks louder than felicity: children's comprehension of long passives. In *Proceedings from the generative approaches to language acquisition 2015* (pp. xxx–xxx).

Nguyen, E., & Pearl, L. (2018). Do You Really Mean It? Linking Lexical Semantic Profiles and the Age of Acquisition for the English Passive. In *Proceedings of the 35th West Coast Conference on Formal Linguistics* (pp. 288-295–xxx). Somerville, MA: Cascadilla Press.

Nguyen, E., & Pearl, L. (2019). Using Developmental Modeling to Specify Learning and Representation of the Passive in English Children. In *Proceedings of the Boston University Conference on Language Development 43* (pp. xxx–xxx). Somerville, MA: Cascadilla Press.

Niyogi, P., & Berwick, R. C. (1996). A language learning model or finite parameter spaces. *Cognition*, *61*, 161–193.

O'Brien, K., Grolla, E., & Lillo-Martin, D. (2006). Long passives are understood by young children. In *Proceedings from the 30th boston university conference on language development* (pp. 441–451).

Orfitelli, R. M. (2012). *Argument intervention in the acquisition of A-movement*. University of California, Los Angeles.

Orita, N., McKeown, R., Feldman, N. H., Lidz, J., & Boyd-Graber, J. (2013). Discovering pronoun categories using discourse information. In *Proceedings of the 35th annual conference of the cognitive science society.*

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

Pearl, L. (2019). *Poverty of the stimulus without tears.* Retrieved from https://ling.auf.net/lingbuzz/004646 (University of California, Irvine)

Pearl, L. (in press). Modeling syntactic acquisition. In J. Sprouse (Ed.), *Oxford handbook of experimental syntax.* Oxford University Press.

Pearl, L. (in press). How statistical learning can play well with Universal Grammar. In Nicholas Allott and Terje Lohndal and Georges Rey (Ed.), *Wiley-Blackwell Companion to Chomsky.* Wiley. Retrieved from https://ling.auf.net/lingbuzz/004772

Pearl, L., & Lidz, J. (2013). Parameters in Language Acquisition. In K. Grohmann & C. Boeckx (Eds.), *The Cambridge Handbook of Biolinguistics* (pp. 129–159). Cambridge, UK: Cambridge University Press.

Pearl, L., Lu, K., & Haghighi, A. (2017). The character in the letter: Epistolary attribution in Samuel Richardson's *Clarissa*. *Digital Scholarship in the Humanities*, *32*(2), 355–376.

Pearl, L., & Sprouse, J. (2013a). Computational Models of Acquisition for Islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Islands Effects* (pp. 109–131). Cambridge: Cambridge University Press.

Pearl, L., & Sprouse, J. (2013b). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*, 19–64.

Pearl, L., & Sprouse, J. (2015). Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research*, *58*, 740–753.

Pearl, L., & Sprouse, J. (2019). *The acquisition of linking theories: A Tolerance and Sufficiency Principle approach to learning UTAH and rUTAH.* University of California, Irvine and University of Connecticut. Retrieved from https://ling.auf.net/lingbuzz/004088

Pearl, L., & Sprouse, J. (in press). Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language*. Retrieved from https://ling.auf.net/lingbuzz/003913

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

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

Pine, J. M., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, *127*(3), 345–360.

Pine, J. M., & Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Pine, J. M., & Martindale, H. (1996). Syntactic categories in the speech of young children: The case of the determiner. *Journal of Child Language*, *23*(2), 369–395.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 399–441). New Jersey: Lawrence.

Ross, J. (1967). *Constraints on variables in syntax* (Unpublished doctoral dissertation). MIT, Cambridge, MA.

Rovee, C. K., & Rovee, D. T. (1969). Conjugate reinforcement of infant exploratory behavior. *Journal of Experimental Child Psychology*, *8*(1), 33–39.

Rovee-Collier, C. K., & Capatides, J. B. (1979). Positive behavioral contrast in 3-month-old infants on multiple conjugate reinforcement schedules. *Journal of the Experimental Analysis of Behavior*, *32*(1), 15–27.

Rowland, C. F., & Theakston, A. L. (2009). The acquisition of auxiliary syntax: A longitudinal elicitation study. Part 2: The modals and auxiliary DO. *Journal of Speech, Language, and Hearing Research*.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Sakas, W. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 415–422). Sapporo, Japan: Association for Computational Linguistics.

Sakas, W. (2016). Computational approaches to parameter setting in generative linguistics. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The oxford handbook of developmental linguistics* (pp. 696–724). Oxford, UK: Oxford University Press.

Sakas, W., & Fodor, J. (2012). Disambiguating Syntactic Triggers. *Language Acquisition*, *19*(2), 83–143.

Sakas, W., & Fodor, J. D. (2001). The Structural Triggers Learner. In S. Bertolo (Ed.), *Language Acquisition and Learnability* (pp. 172–233). Cambridge, UK: Cambridge University Press.

Sakas, W., & Nishimoto, E. (2002). *Search, Structure or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition.* City University of New York, NY. (Manuscript)

Schuler, K., Yang, C., & Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *The 38th Cognitive Society Annual Meeting, Philadelphia, PA*.

Shin, Y. K. (2012). A new look at determiners in early grammar: Phrasal quantifiers. *Language Research*, *48*(3), 573–608.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language*, *88*(1), 82–124.

Stahl, A. E., Romberg, A. R., Roseberry, S., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Infants segment continuous events using transitional probabilities. *Child Development*, *85*(5), 1821–1826.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Theakston, A. L., Ibbotson, P., Freudenthal, D., Lieven, E. V., & Tomasello, M. (2015). Productivity of noun slots in verb frames. *Cognitive Science*, *39*(6), 1369–1395.

Theakston, A. L., & Rowland, C. F. (2009). The acquisition of auxiliary syntax: A longitudinal elicitation study. Part 1: Auxiliary BE. *Journal of Speech, Language, and Hearing Research*.

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

Tomasello, M. (2004). What kind of evidence could refute the ug hypothesis? *Studies in Language*, *28*(3), 642–645.

Tomasello, M., & Abbot-Smith, K. (2002). A tale of two theories: Response to Fisher. *Cognition*,

83(2), 207–214.

Tomasello, M., & Brandt, S. (2009). Flexibility in the semantics and syntax of children's early verb use. *Monographs of the Society for Research in Child Development*, *74*(2), 113–126.

Ud Deen, K., & Timyam, N. (2018). Condition C in adult and child Thai. *Language*, *94*(1), 157–190.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*(4), 562.

Valian, V. (2009). Innateness and learnability. In *Handbook of Child Language* (pp. 15–34). Cambridge University Press Cambridge, England.

Valian, V. (2014). Arguing about innateness. *Journal of Child Language*, *41*(S1), 78–92.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, *36*(4), 743–778.

Wang, H., Höhle, B., Ketrez, N., Küntay, A. C., Mintz, T. H., Danis, N., ... Sung, H. (2011). Cross-linguistic distributional analyses with Frequent Frames: The cases of German and Turkish. In *Proceedings of 35th Annual Boston University Conference on Language Development* (pp. 628–640).

Wang, H., & Mintz, T. (2008). A dynamic learning model for categorizing words using frames. In H. Chan, H. Jacob, & E. Kapia (Eds.), *Proceedings of the 32nd annual Boston University Conference on Language Development [BUCLD 32]* (pp. 525–536). Somerville, MA: Cascadilla Press.

Watson, J. S. (1969). Operant conditioning of visual fixation in infants under visual and auditory reinforcement. *Developmental Psychology*, *1*(5), 508.

Weisleder, A., & Waxman, S. R. (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.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, *47*(5), 1220.

Xiao, L., Cai, X., & Lee, T. (2006). The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. In Y. Otsu (Ed.), *Proceedings of the Seventh Tokyo Conference on Psycholinguistics* (pp. 299–322). Tokyo: Hitizi Syobo.

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*(2), 245-272.

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford, UK: Oxford University Press.

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

Yang, C. (2005). On productivity. *Yearbook of Language Variation*, *5*, 333–370.

Yang, C. (2011). A statistical test for grammar. In *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics* (pp. 30–38).

Yang, C. (2012). Computational models of syntactic acquisition. *WIREs Cognitive Science*, *3*, 205–213.

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

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.
Yang, C. (2017). How to wake up irregular (and speechless). *On looking into words (and beyond)*, 211–233.