Chapter 1

How statistical learning can play well with Universal Grammar

Lisa S. Pearl[1]

[1] *Language Science & Cognitive Sciences, University of California, Irvine, 3151 Social Science Plaza A, Irvine, California, 92697, USA, lpearl@uci.edu*

Abstract: One motivation for Universal Grammar (**UG**) is that it allows children to acquire the linguistic knowledge that they do as quickly as they do from the data that's available to them. A key legacy of Chomsky's work in language acquisition is highlighting the separation between children's hypothesis space of possible representations (typically defined by UG) and how children might navigate that hypothesis space (sometimes defined by UG). While statistical learning is sometimes thought to be at odds with UG, I review how statistical learning can both complement UG and help us refine our ideas about the contents of UG. I first review some cognitively-plausible statistical learning mechanisms that can operate over a predefined hypothesis space, like those UG creates. I then review two sets of examples: (i) those where statistical learning allows more efficient navigation through a UG-defined hypothesis space, and (ii) those where statistical learning can replace prior UG proposals for navigating a hypothesis space, and in turn lead to new ideas about how UG might construct the hypothesis space in the first place. I conclude with a brief discussion of how we might make progress on understanding language acquisition by incorporating statistical learning into our UG-based acquisition theories.

Keywords: Universal Grammar, statistical learning, linguistic parameters, parameter setting, linking theories, Subset Principle, syntactic islands

1.1. Introduction

A key motivation for Universal Grammar (**UG**) is developmental: UG can 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 and Margolis, 2001, Crain and Pietroski, 2002]. That is, UG allows children to solve what Chomsky [Chomsky, 1965, 1975, 1980a,b] and many subsequent others (see Pearl [2019] for a recent review) have called the *Poverty of the Stimulus*, where the available data often seem woefully inadequate for pinpointing the right linguistic knowledge as efficiently as children seem to. So, without some internal bias, children would be lost when it comes to language acquisition. Enter UG: a concrete proposal for what that internal bias could be that enables language acquisition to succeed.

Among other things, UG provides a way to structure a child's hypothesis space about the linguistic system – what explicit hypotheses are considered, and (perhaps more importantly) what building blocks allow children to construct those explicit hypotheses for consideration. For example, traditional linguistic macro-parameters [Chomsky, 1981, 1986] are building blocks that children can construct language grammars from – that is, a language's grammar would be a specific collection of parameter values for these linguistic parameters. Having these building blocks then allows a child to construct and consider explicit hypotheses about a language's grammar as language data are encountered. But, setting these parameters from the available language data is a *hard* problem, which is why a lot of energy has been devoted to figuring out how that process could work (e.g., see the more detailed discussion of this body of work in Pearl [in press]). In particular, parameter setting is hard even if UG allows the child to already know which parameters there are and which parameter values are possible for each parameter. The problem is simply that the hypothesis space is vast (e.g., n binary parameters yield 2^n possible grammars – this number gets very big very quickly); also, the child's input data are often ambiguous between different grammars.

One plausible solution to the parameter-setting problem relies on reinforcement learning – a domain-general statistical learning mechanism – applied to linguistic parameter values (see Pearl [in press] for more discussion on this approach implemented by Yang [2002, 2004], Legate and Yang [2007], Yang [2012] and Legate and Yang [2013]). Relatedly, some of the most fruitful recent work in language acquisition has combined ideas about different hypothesis space building blocks with domain-general statistical learning (e.g., Piantadosi et al. [2012], Orita et al. [2013], Pearl and Sprouse [2013a], Abend et al. [2017], Pearl and Sprouse [in press]).

What all these approaches to language acquisition demonstrate is that statistical learning only works in tandem with a predefined hypothesis space. To me, this is one key legacy of Chomsky's work in the area of language acquisition: statistical learning by itself can't be an answer to how children learn the linguistic representations they do. More specifically, Chomsky highlighted that children's hypothesis space of representations is a crucial component in the language acquisition process – and therefore, understanding the nature of that hypothesis space is a concrete way to make progress in understanding how language acquisition proceeds. Statistical learning can then provide a way to help navigate the hypothesis space (e.g., of possible grammars) in order to converge on the correct hypothesis (e.g., a specific language's grammar). In this way, statistical learning can complement the prior knowledge that defines the hypothesis space – but statistical learning does *not* itself define the hypothesis space and so could never replace prior knowledge that does in fact define the hypothesis space. So, any components of UG that define the child's linguistic hypothesis space are not replaceable by statistical learning. However, any components of UG that involve navigating the hypothesis space *may* be replaceable by statistical learning. And, perhaps by considering other ways of navigating a predefined hypothesis space, we may be able to refine our ideas about the UG knowledge that constructs that hypothesis space in the first place.

Below, I explore these ideas more concretely – cognitively-plausible statistical learning mechanisms, how statistical learning can complement UG, and how statistical learning can help us refine our ideas about what's in UG. I'll first give a brief overview of statistical learning mechanisms we have evidence for in small humans, along with some illustrative examples. This evidence is the foundation for talking about statistical learning at all for language acquisition – if these mechanisms aren't there in small humans, there's no need to spend energy thinking about how they might be used by small humans learning language. But, if statistical learning mechanisms are in fact there and useable by small humans, we have motivation for thinking about how children might leverage them during language acquisition.

I'll then discuss some examples where statistical learning complements UG, which happens when statistical learning offers a way to efficiently navigate the hypothesis space defined by UG. I'll then turn to some examples where statistical learning can replace prior UG proposals for navigating a hypothesis space; this can lead to new ideas for how the hypothesis space might be constructed in the first place. I'll conclude briefly with what I think is a concrete way to make progress on understanding the process of language acquisition, integrating what we know about statistical learning.

1.2. Statistical learning mechanisms in small humans

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). So, to effectively use statistical learning, a child has to know what to count. 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.

For language acquisition, counting things often means counting linguistic things. Importantly, what the child might be counting often depends on the representational and learning theories we currently have for the linguistic knowledge to be acquired and children's developing representation and processing capacities [Pearl, in press]. For instance, depending on the theories involved, a child acquiring syntactic knowledge might be doing inference over counts of of lexical items [Yang, 2005, Freudenthal et al., 2007, 2009, 2010, 2015, Yang, 2016, 2017], syntactic category sequences [Perfors et al., 2010, 2011a, Pearl and Sprouse, 2013a], syntactic signals realized in certain overt structures [Sakas and Fodor, 2001, Yang, 2004, Legate and Yang, 2007, Mitchener and Becker, 2010, Pearl and Sprouse, 2013a, Becker, 2014, Sakas, 2016, Pearl and Sprouse, 2019], or something else entirely. 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.

Also importantly, we have good reason to think that small humans are good at both counting things and doing inference over those counts. There's a lot of evidence that infants under a year old are capable of exactly this (9-month-olds: Fiser and Aslin [2002], Wu et al. [2011]; 8-month-olds: Saffran et al. [1996], Aslin et al. [1998], Saffran et al. [1999], Stahl et al. [2014], Aslin [2017]; 2-month-olds: Kirkham et al. [2002]; newborns: Ferry et al. [2016], Fló et al. [2019]; among many other studies). So, statistical learning isn't unreasonable to think about as something children *could* use during language acquisition. Happily, we have several current proposals for useful ways children could harness statistical learning for language acquisition, which I'll briefly review below.[1]

[1] In the interest of space, I'll only provide abbreviated descriptions of each statistical learning mechanism below, but see Pearl [in press] for a more comprehensive overview of each one.

1.2.1. Some ways of doing statistical learning

1.2.1.1. Reinforcement learning

Reinforcement learning (see Sutton and Barto [2018] for a recent overview) is a principled way to update the probability of a categorical option which is in competition with other categorical options. So, if we're thinking in terms of linguistic parameters, perhaps a child might consider a parameter that controls whether *wh*-movement is required as the default in questions (such as in English, with the *wh*-word *what* in *What's Jack climbing* $\_\__{what}$*?* = +wh-movement), or whether the *wh*-word can remain in-situ (such as in Japanese = -wh-movement). As mentioned above, Yang & Legate's approach relies on reinforcement learning, and uses an implementation by Bush and Mosteller [1951] called the linear reward-penalty scheme. As the name suggests, there are two choices when a data point is processed – either the categorical option under consideration is rewarded or it's penalized. This translates to the option's current probability being increased (rewarded) or decreased (penalized). 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 to linguistic parameters 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 et al. [1966]; 10 months: de Sousa et al. [2015]; 3 months: Rovee-Collier and Capatides [1979]; 10 weeks: Rovee and Rovee [1969], Watson [1969]; among many others). So, it's plausible that young children could use reinforcement learning for language acquisition.

1.2.1.2. The Tolerance and 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, is there a general rule for forming the past tense in English from a verb's root form (e.g., *kiss* → *kissed*), a generalization about how thematic roles connect to their syntactic positions (e.g., is the AGENT often linked to the *subject* syntactic position), or a default stress pattern for three-syllable English words (e.g., is the stress pattern in *óctopus* typical)?

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 examples above, there are certainly exceptions in the child's input: past tense forms like *drank* (rather than *drinked*), non-AGENTs appearing in the *subject* position (*Jack*$_{\text{PATIENT}}$ *got kissed*), and three-syllable words with different stress patterns like *horízon*.

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 (i) a past tense rule to retrieve a regular past tense form, (ii) a linking rule to retrieve a thematic role reliably associated with a syntactic position, or a syntactic position reliably associated with a thematic role, and (iii) a metrical stress rule to retrieve a predictable metrical stress pattern. However, if the past tense is too irregular, the link too unreliable, or the metrical stress too unpredictable, it's not useful to have the rule: retrieving the target information 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 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, including English past tense morphology [Yang, 2005, 2016], English noun pluralization [Yang, 2016], German noun pluralization [Yang, 2005], English nominalization [Yang, 2016], English metrical stress [Legate and Yang, 2013, Yang, 2015, Pearl et al., 2017], English *a*-adjective morphosyntax [Yang, 2015, 2016], English dative alternations [Yang, 2016, 2017], noun morphology in an artificial language [Schuler et al., 2016], linking theories in English [Pearl and Sprouse, 2019], and the development of causative use in English [Irani, 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.

1.2.1.3. Bayesian inference

Bayesian inference operates over probabilities, and involves both prior assumptions about the probability of different options (typically referred to as hypotheses) and an estimation of how well a given hypothesis fits the data. 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 child might consider both a +wh-movement option and a -wh-movement option as two hypotheses ($\{$+wh-movement, -wh-movement$\} \in H$), while 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), which involves two key components: the *likelihood* of the data, given the hypothesis ($P(D|h)$), and the *prior* of the hypothesis ($P(h)$).

(1) $P(h|D) \propto P(D|h) * P(h)$

The likelihood $P(D|h)$ captures how compatible hypothesis h is with the data D: hypotheses with a poor fit to the data (e.g., the -wh-movement hypothesis for a dataset with 30% of the data compatible only with +wh-movement) have a lower likelihood; hypotheses with a good fit to the data have a higher likelihood. The prior $P(h)$ captures how plausible the hypothesis is, irrespective of its fit to the 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 et al. [2011b] and Piantadosi et al. [2012]). For example, more complex hypotheses will typically have lower prior probabilities.

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

1.2.2. Why statistical learning in small humans shouldn't alarm people who like UG

What all these statistical learning approaches demonstrate is that there are in fact several cognitively-plausible statistical learning mechanisms that children could harness for language acquisition. (Of course, how children would know which one to use for which language acquisition problem is another question.) Perhaps more importantly, there are linguistically-savvy ways to harness those mechanisms. That is, being able to count things and do inference over those counts doesn't negate the need to figure out what to count. This key point about what to count is echoed outside the UG-supporting community as well – for example, see the discussion on this point in Scholz and Pullum [2006].

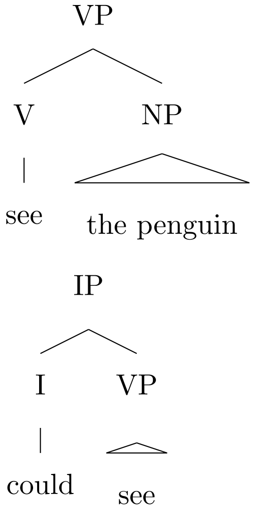
Within the generative linguistics community, there are several examples of researchers who take both statistical learning and theories of linguistic representation very seriously; their studies demonstrate how linguistically-savvy statistical learning can yield informative results about both language acquisition and the linguistic representations being acquired (e.g., parameter setting in syntax [Yang, 2002, 2004, 2012], parameter setting in metrical phonology [Legate and Yang, 2013, Pearl et al., 2017], phonetic category acquisition [Feldman et al., 2013], morphosyntactic acquisition [Yang, 2005, 2015, 2016], pronoun acquisition [Orita et al., 2013, Pearl and Mis, 2016], syntactic island acquisition [Pearl and Sprouse, 2013a], linking theory acquisition [Pearl and Sprouse, in press, 2019], as well as many other examples of syntactic acquisition discussed in Pearl [in press]).

So, these studies allow us to see how statistical learning can complement the innate language-specific structure that UG can provide. I also want to highlight a strong connection between one statistical learning mechanism, Bayesian inference, and the intuition of linguistic macro-parameters [Chomsky, 1981]. I think this connection underscores the potential that statistical learning has to aid theories of language acquisition that rely on UG – and this connection is a ripe avenue for future investigation.

The key idea is that Bayesian inference can be implemented over a hierarchical hypothesis space, such that there are different levels of abstraction for the hypothesis space [Goodman, 1955, Kemp et al., 2007, Kemp and Tenenbaum, 2008]. In particular, there may be narrower hypotheses about certain observable properties, and more abstract hypotheses (called *overhypotheses*) about the nature of those narrower hypotheses. A concrete example taken from Pearl and Lidz [2013] helps demonstrate this intuition, specifically linking it to the concept of linguistic macro-parameters.

Suppose an English-learning child has a collection of utterances in her data D. Some utterances contain verbs and objects, and whenever there's an object, suppose it appears after the verb, e.g., *see the penguin* rather than *the penguin see*. Other utterances contain modal verbs and nonfinite main verbs, and whenever both occur, the modal verb precedes the main verb, e.g., *could see* rather than *see could*. A child could be aware of the shared structure of these data – specifically that these observable forms can be characterized as the phrasal head appearing before its complements, as shown in (2).

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



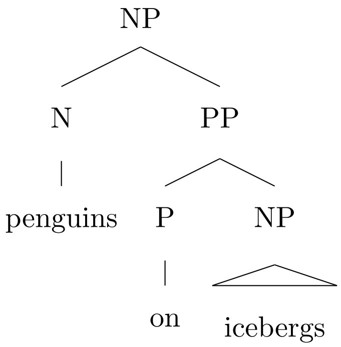
a. *see the penguin*

b. *could see*

This "head-first" idea can be encoded at a level that describes utterances in general, and would represent an overhypothesis about structure which is similar to the classical headedness linguistic macro-parameter: namely, phrases have their heads before their complements. Each example from (2) instantiates this overhypothesis for particular phrases, VPs and IPs, and so predicts the narrower hypotheses these data support: both VPs and IPs have their heads before their complements. Each narrower hypothesis strengthens the more abstract head-first overhypothesis, and allows the child to make predictions about phrases not yet encountered. For example, if this child encountered an utterance with a preposition surrounded by NPs, such as *penguins on icebergs*, there are (at least) two structural hypotheses possible, shown in (3).

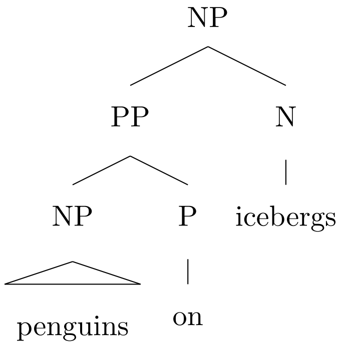
(3) Possible structures for *penguins on icebergs*

a. head-first:



Meaning: penguins who are on icebergs

b. head-last:



Meaning: icebergs that are on penguins

Using the head-first overhypothesis for both the NP and the PP, the child would prefer the structure in (3a), whose meaning indicates this phrase is about penguins who are on icebergs (rather than icebergs that are on penguins).

Importantly, the child could have this head-first preference about NP and PP structure, even *without* having encountered a single NP or PP data point. This is the power of overhypotheses, and is the same powerful intuition that made the classical notion of linguistic macro-parameters attractive to developmental linguists. Children can learn how to make generalizations for data they've never seen before on the basis of the data they do see precisely because of the abstract system underlying the generation of the observable data.

In short, the classical notion of a linguistic macro-parameter is an abstract structural property that constrains the hypothesis space of the child, and is defined by UG. So, this type of parameter is an overhypothesis about the narrower structural hypotheses available. Bayesian inference can easily operate over hypothesis spaces that involve overhypotheses (and even more abstract levels – see Perfors et al. [2011b] for a concrete example in the syntactic domain). Because of this, there may be a very intuitive connection between hierarchical Bayesian inference and acquisition with linguistic macro-parameters.

This connection in turn highlights the main strength of statistical learning that I mentioned in the introduction: statistical learning can help a child efficiently search a large hypothesis space. For instance, Bayesian inference can help a child efficiently search vast hypothesis spaces defined by overhypotheses and related narrower hypotheses; I discussed this above for linguistic macro-parameters, and it's also the approach taken by Pearl and Sprouse [in press] for linking theory acquisition.

What this then means is that children may not need their hypothesis spaces as constrained as we thought. That is, because children can harness statistical learning to more efficiently navigate a predefined hypothesis space, that hypothesis space could be larger or messier than we thought; this in turn may mean that we don't need as much specific innate knowledge to define children's hypothesis spaces as we thought. In other words, statistical learning could allow children to get away with hypothesis spaces that are more loosely defined – and yet children could still manage to successfully acquire their target linguistic knowledge.

So, statistical learning may either (i) replace language-specific knowledge that's meant to help a child navigate a predefined hypothesis space, or (ii) obviate the need for language-specific knowledge that constrains the child's hypothesis space so tightly. Both of these results help us refine our ideas about what's in UG, and I'll discuss a concrete example of each below in section 1.4.

An upshot of refining our ideas about what's in UG (in particular, reducing the amount of innate knowledge required – a goal at the heart of the Minimalist Program [Chomsky, 2014]) is that we save ourselves some neurobiological explanatory work. In particular, everything that's part of UG must be present in a human's neurobiological endowment somehow. So, the more we build into UG, the more we have to explain about how that neurobiological endowment works. This generally involves explanations from both developmental neurobiology and evolutionary biology. So, it seems like a good idea to be very careful about what we propose to be in UG; if we consider the contributions of statistical learning in order to refine what's in UG, we may save ourselves some difficult explanatory work on the biological front.

1.3. When statistical learning complements UG

I'll now briefly review the highlights of two recent studies [Pearl and Sprouse, 2019, in press] that use statistical learning to investigate the acquisition of linking theories. (See Pearl [in press] for a more extensive review of both studies.) In each case, children's hypothesis space for linking theory knowledge is constrained by UG, and so statistical learning is operating over a linguistically-predefined hypothesis space.

To get an idea about what linking theories are for, consider the following sentence: *The little girl blicked the kitten on the stairs.* Even if we don't know what *blick* means, we still have preferences about how to interpret this sentence. In particular, out of all the logically possible interpretations involving the little girl, the kitten, and the stairs, we prefer an interpretation where the little girl is doing something (blicking) to the kitten, and that event is happening on the stairs. The reason we as adults have this preferred interpretation is because we have *linking theories* that link the thematic roles specified by a verb's lexical semantics to the syntactic argument positions specified by that verb's syntactic frame. Moreover, our linking theories are so well-developed that they can impose these links even when we don't know a verb's specific lexical semantics (as we see here with *blick*). Developing linking theories is a fundamental component of verb learning for children.

There are two prominent options for linking theory theoretical representations: the Uniformity of Theta Assignment Hypothesis (**UTAH**) and the relativized form of that theory, relativized UTAH (**rUTAH**). The key difference between these two approaches has to do with how they deal with individual thematic roles. UTAH [Fillmore, 1968, Perlmutter and Postal, 1984, Jackendoff, 1987, Baker, 1988, Grimshaw, 1990, Speas, 1990, Dowty, 1991, Baker, 1997] groups these roles into larger thematic categories (like AGENT-ish, PATIENT-ish, or OTHER-ish) that then always map to a specific syntactic position (e.g., AGENT-ish ↔ *subject*). In contrast, rUTAH assumes there's a relative ordering of thematic roles (e.g., AGENT>PATIENT), and whichever thematic roles are present are sorted according to this ordering. Then, the highest role available maps to the highest syntactic position (e.g., *subject*). This gives rUTAH more flexibility than UTAH, for example allowing it to directly handle unaccusative constructions like *The ice*$_{\text{PATIENT}}$ *broke*. In this case, UTAH would expect the *subject* to have an AGENT-ish role, which PATIENT certainly isn't. In contrast, rUTAH would note there was only one role available (PATIENT) and expect that role to appear in the highest syntactic position available (*subject*), which it does. Notably, both UTAH and rUTAH seem to be compatible with current cross-linguistic data, and so cross-linguistic coverage can't easily decide which one is likely to be the correct one.

There are also two main approaches to how children develop linking theory knowledge: (i) it's innate knowledge of the specific links between thematic roles and syntactic positions that comes fully-formed in the child's mind [Fillmore, 1968, Perlmutter and Postal, 1984, Baker, 1988, Larson, 1990, Speas, 1990, Grimshaw, 1990], or (ii) the specific linking knowledge develops over time via an interplay between the input that children receive and the mechanisms (both statistical and otherwise) that underlie verb learning [Bowerman, 1988, Goldberg, 1995, 2006, Boyd and Goldberg, 2011, Goldberg, 2013]. As with the theoretical representations, both developmental options seem viable at face value.

This is where models using linguistically-savvy statistical learning may help. In particular, Pearl and Sprouse [in press] offers insight into the development of linking theory knowledge, suggesting that derivation of specific linking theory knowledge may occur over time from children's input. More specifically, their results argue against early maturation of innate linking theory knowledge. So, deriving that linking theory knowledge from the input becomes a plausible option to explore. An important takeaway from Pearl and Sprouse [in press] is that these results come from combining the statistical learning mechanism of hierarchical Bayesian inference with linguistic representations that define a child's hypothesis space (here, UTAH and rUTAH); by doing so, we gain insight about the process of acquisition that needs to accompany either linking theory representation.

Pearl and Sprouse [2019] offers insight into which theoretical option is more compatible with deriving linking theory knowledge from children's input. In particular, their results suggest two significant advantages for rUTAH over UTAH, if children derive their linking theories from their input using the Tolerance and Sufficiency Principles. First, a rUTAH-learning child will have a much easier time generating the complex linking patterns that comprise the rUTAH linking theory from the available input. Second, only rUTAH – and not UTAH – can be successfully generalized from English children's input using the Tolerance and Sufficiency Principles. So, by combining statistical learning with a linguistically-defined hypothesis space, Pearl and Sprouse [2019] find support for a relativized linking theory like rUTAH over a fixed linking theory like UTAH, when coupled with a derivational theory of acquisition.

1.4. When statistical learning refines what's in UG

We'll now turn to some examples where statistical learning can help refine our ideas of what's in UG. The first example demonstrates how statistical learning can replace a language-specific way to navigate a certain type of linguistic hypothesis space. The second example demonstrates how statistical learning can allow us to replace very specific UG syntactic knowledge with more general-purpose UG syntactic knowledge, and thereby lead to a hypothesis space that's less tightly constrained than previously thought.

1.4.1. A bias for the subset hypothesis

One potential issue, known as the *Subset Problem* [Berwick, 1985, Manzini and Wexler, 1987], can arise when a child is navigating her hypothesis space. This problem occurs when the correct hypothesis for the language is a subset of a competing hypothesis for the language and the observed data are ambiguous between these hypotheses (see Figure 1.1). Here we see a two-dimensional space, where some data points are observed (each indicated with a d). All the observed data are compatible with both correct hypothesis C and incorrect hypothesis N.

The problem is that there are no data that can unambiguously indicate C is correct – every d in Figure 1.1 that can be accounted for by C can also be accounted for by N. Importantly, this isn't just an issue with the observed data in this particular scenario – it's a problem for C in general. There are no data a child could possibly encounter that would be compatible with C but

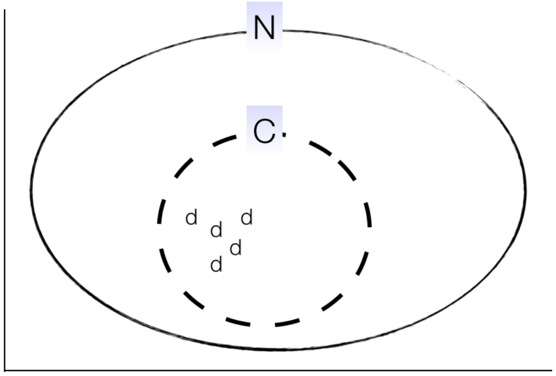


Figure 1.1: A visual demonstration of the Subset Problem: a two-dimensional hypothesis space where one hypothesis (the correct one, C, in dashed lines) is a subset of a competing hypothesis (a incorrect one, N). Each d corresponds to an observed data point, and both hypotheses are compatible with the observed data.

not with N – this is what it means for C to be a subset of N (and for N to be a superset of C). So, because every data point in C is also in N, all "C" data points have unresolvable ambiguity between C and N (i.e., there's a Subset Problem caused by N if the child needs to learn C).

Importantly, this is a problem for the type of learning mechanisms that learning theorists previously assumed children used: mechanisms (i) relying on unambiguous data to trigger the correct hypothesis [Lightfoot, 1989, Fodor, 1998, Dresher, 1999, Sakas and Fodor, 2001, 2012], or (ii) involving error-driven learning that rely on some data being incompatible with the incorrect hypotheses [Gibson and Wexler, 1994, Niyogi and Berwick, 1996, Sakas and Fodor, 2001]. In our example, a child looking for data that unambiguously signal C won't find any – all data compatible with C are also compatible with N; a child looking for data that signal N is incorrect also won't find any – again, all data she'll encounter (which come from speakers using C) will be compatible with N, too.

To tackle this problem, the Subset Principle [Berwick, 1985, Manzini and Wexler, 1987] was proposed; this principle causes the child to prefer the subset hypothesis, even if the data are ambiguous (e.g., preferring C over N, despite all the *d*s being compatible with both hypotheses). This principle was language-specific (and so part of UG), and intended to help children successfully navigate linguistic hypothesis spaces that had these subset problems.

However, this same preference can be derived from the mechanics of Bayesian inference, which doesn't have to be built into UG. Recall the two components used to calculate the posterior probabilities of different hypotheses: the prior of a hypothesis, $P(h)$, and the likelihood of the data, given a hypothesis, $P(D|h)$. The prior captures how much the child favors the hypothesis before any data have been encountered yet. Let's suppose that the priors for both C and N are the same (say, $P(C) = P(N) = 0.5$), so there's no reason to prefer one over the other a priori.

The likelihood can be calculated by considering the probability of a particular hypothesis generating the observed data. This is where children's expectations about the data each hypothesis could generate are really important. In particular, because C is a subset of N, C is by definition capable of generating fewer data than N is. To be concrete, let's suppose C is capable of generating 10 data points (including the 5 we saw in Figure 1.1) while N is capable of generating 20 data points (including the 10 that C can generate, which include the 5 we saw in Figure 1.1), as shown in Figure 1.2.

Let's now consider the probability of each hypothesis generating a data point like d_1, which is compatible with both C and N. Because C can generate 10 data points, the probability of C generating d_1 is $\frac{1}{10}$ ($P(d_1|C) = 0.10$). Because N can generate 20 data points, the probability of N generating d_1 is

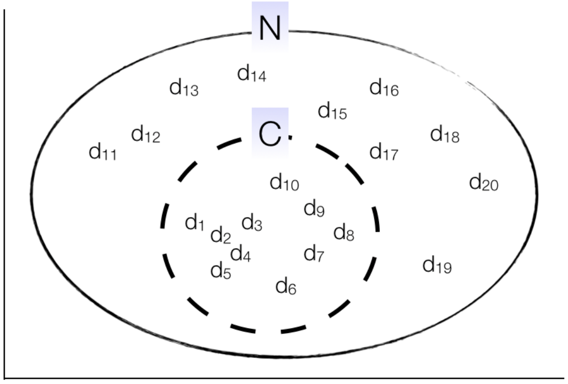


Figure 1.2: A visual demonstration of the data each hypothesis (correct subset hypothesis C and incorrect superset hypothesis N) can generate.

$\frac{1}{20}$ ($P(d_1|N) = 0.05$).

Because the posterior probabilities of C and N, given the data, are proportional to their prior * likelihood, this means that a child seeing only ambiguous data point d_1 would assess the posterior probabilities of C and N as in (4):

(4) a. $P(C|d_1) \propto P(C) * P(d_1|C) = 0.5 * 0.1 = 0.05$

b. $P(N|d_1) \propto P(N) * P(d_1|N) = 0.5 * 0.05 = 0.025$

So, just from seeing d_1, a child would believe correct subset hypothesis C is twice as likely as incorrect superset hypothesis N ($\frac{P(C|d_1)}{P(N|d_1)} = \frac{0.05}{0.025} = 2$). That is, the child has a preference for the subset hypothesis C, even without having encountered either unambiguous data for C or error data indicating N is incorrect. This is exactly what the Subset Principle was meant to instill.

Moreover, the more data the child sees like d_1 (i.e., compatible with both the subset and superset hypotheses), the stronger this preference becomes. For instance, if the child saw 5 data points compatible with both C and N (i.e., $\{d_1, d_2, d_3, d_4, d_5\} = Data$, as in Figure 1.1), the likelihood for C would be $(\frac{1}{10})^5 = 0.00001$, while the likelihood for N would be $(\frac{1}{20})^5 = 0.0000003125$. Because the priors for C and N are still the same (0.5), these ambiguous data would cause the child to believe correct subset hypothesis C is 32 times more likely than incorrect superset hypothesis N ($\frac{P(C|Data)}{P(N|Data)} = \frac{0.5*0.00001}{0.5*0.0000003125} = 32$). This is an even stronger bias for the subset. Notably, Bayesian inference allows us to quantify how much the child ought to prefer the subset hypothesis, given the data encountered; this is in contrast to the Subset Principle, which specifies that the child ought to prefer the subset hypothesis, but not *how much* she ought to prefer it.

More generally, this example is meant to highlight how statistical learning (in this case, Bayesian inference) can offer an alternative solution for navigating a tricky language learning scenario (i.e., when the linguistic hypotheses overlap in a particular way and all the available data are ambiguous). So, a child could successfully navigate this linguistic hypothesis space without relying on a language-specific mechanism to do so; instead, Bayesian inference would allow her the same navigation ability and also not require us to build that navigation mechanism into UG.

1.4.2. Linguistic knowledge for syntactic islands

Relationships between two words in a sentence that aren't adjacent to each other, such as the long-distance dependency between *what* and *stole* in (5a), can be potentially infinite in length. However, there are specific syntactic structures that long-distance dependencies can't cross, known as *syntactic islands*, with four examples shown in (5b)-(5e) [Chomsky, 1965, Ross, 1967, Chomsky, 1973]. During acquisition, children must infer the constraints on long-distance dependencies (i.e., the syntactic islands) that allow them to recognize that the *wh*-dependencies in (5b)-(5e) aren't allowed, while the *wh*-dependency in (5a) is.

(5) a. What does Jack think that Lily said that the goblins stole t_{what}?

b. *whether* island

*What do you wonder [whether Jack bought t_{what}]?

c. complex NP island

*What did you make [the claim that Jack bought t_{what}]?

d. subject island

*What do you think [the joke about t_{what}] was hilarious?

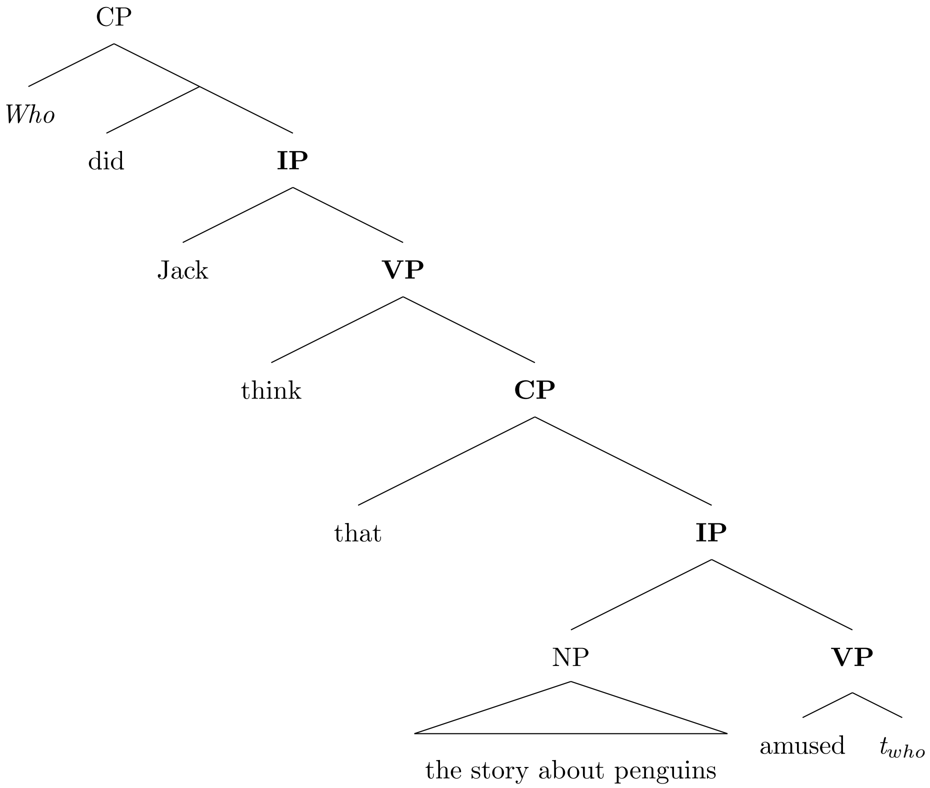
e. adjunct island

*What do you worry [if Jack buys t_{what}]?

Pearl and Sprouse [2013b,a, 2015] investigated a statistical learning strategy for learning about syntactic islands that probabilistically leveraged linguistic knowledge; more specifically, this strategy relied on children characterizing a long-distance dependency as a path from the head of the dependency (e.g., *Who* in (6)) through the phrasal nodes that contain the tail of the dependency, as shown in (6a)-(6b). Under this view, children simply need to learn which long-distance dependencies have licit syntactic paths and which don't. The probabilistic learning strategy of Pearl & Sprouse tracks local pieces of these syntactic paths, which we can think of as the building blocks of the syntactic paths. Here, I'll briefly review here the highlights of how we can interpret the results of this linguistically-savvy statistical learning strategy, but see Pearl [in press] for a more detailed overview of both the strategy itself and how to interpret it.

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

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



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

b. Syntactic path of *wh*-dependency:

start-IP-VP-CP$_{that}$-IP-VP-*end*

Pearl & Sprouse's results suggested that their proposed learning strategy was a reasonable way for English children to acquire knowledge of the four islands in (5b)-(5e). Importantly, this strategy offers an alternative to a previous acquisition theory for syntactic islands from the Government and Binding framework of the 1980s, which involved a constraint called the *Subjacency Condition*. This constraint basically says that dependencies can't cross two or more *bounding nodes* (Chomsky 1973, Huang 1982, Lasnik and Saito 1984, among others); if a dependency crosses two or more bounding nodes, a syntactic island occurs. What counts as a bounding node varies cross-linguistically, though bounding nodes are always drawn from the set {NP, IP, CP}. So, when using this representation, children needed to learn which of these are bounding nodes in their language, though they already know (via UG) about the 2-bounding-node restriction and the set of possible bounding nodes. That is, children's hypothesis space is highly restricted.

The learning strategy of Pearl and Sprouse [2013b,a, 2015] shares the intuition with Subjacency that there's a local structural anomaly when syntactic islands occur. However, instead of characterizing this anomaly with bounding nodes, Pearl & Sprouse suggested that it could be described as a low probability region with respect to the phrase structure nodes that characterize the syntactic path of the *wh*-dependency. There's no need for the child to know about a hard constraint prohibiting the crossing of two bounding nodes, and in fact no need for the child to identify bounding nodes at all. Instead, children would need to identify the phrase structure nodes containing the *wh*-dependency and break that syntactic path into smaller building blocks based on phrase structure nodes – certainly no trivial matter, but one that involves building blocks that are more general-purpose than bounding nodes.

And this is the point: the knowledge of building blocks based on phrase structure nodes and how to use them doesn't require children to have a highly-constrained hypothesis space defined in terms of UG-specified bounding nodes. That is, because children could use statistical learning to navigate a hypothesis space comprised of syntactic path building blocks, their linguistic hypothesis space doesn't have to be as constrained by UG as the original Subjacency approach assumed it was.

1.5. Concluding thoughts

I hope to have shown here how statistical learning can play well with UG, by both complementing and refining our ideas about UG. This is why I believe researchers who like the idea of UG should be interested in incorporating statistical learning into their theories of language acquisition. Of course, there's much work to be done if we continue down this avenue: we need to provide concrete, testable theories of how acquisition could proceed using both statistical learning and whatever innate linguistic knowledge we think may be appropriate (e.g., see Pearl [2014, in press] for specific examples of how to do this with mathematical and computational developmental modeling). By doing so, we can see how statistical learning can allow children to get far more mileage out of the innate, language-specific knowledge we believe is in UG.

Bibliography

Omri Abend, Tom Kwiatkowski, Nathaniel J Smith, Sharon Goldwater, and Mark Steedman. Bootstrapping language acquisition. *Cognition*, 164:116–143, 2017.

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

Richard N Aslin, Jenny R Saffran, and Elissa L Newport. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321–324, 1998.

Mark C Baker. *Incorporation: A theory of grammatical function changing.* University of Chicago Press, Chicago, IL, 1988.

Mark C Baker. Thematic roles and syntactic structure. In *Elements of grammar*, pages 73–137. Springer, 1997.

Misha Becker. *The acquisition of syntactic structure: Animacy and thematic alignment*, volume 141. Cambridge University Press, 2014.

Robert C Berwick. *The acquisition of syntactic knowledge*, volume 16. MIT press, 1985.

Melissa Bowerman. The 'No Negative Evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins, editor, *Explaining language universals*, pages 73–101. Blackwell, Oxford, England, 1988.

Jeremy Boyd and Adele Goldberg. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, 87(1):55–83, 2011.

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

Noam Chomsky. *Aspects of the Theory of Syntax.* The MIT Press, Cambridge, 1965.

Noam Chomsky. Conditions on transformations. In S. Anderson and

P. Kiparsky, editors, *A Festschrift for Morris Halle*, pages 237–286. Holt, Rinehart, and Winston, New York, 1973.

Noam Chomsky. *Reflections on language*. Pantheon Books, New York, 1975.

Noam Chomsky. Rules and representations. *Behavioral and Brain Sciences*, 3: 1–61, 1980a.

Noam Chomsky. *Rules and Representations*. Basil Blackwell, Oxford, 1980b.

Noam Chomsky. *Lectures on Government and Binding*. Foris, Dordrecht, 1981.

Noam Chomsky. *Barriers*, volume 13. MIT press, 1986.

Noam Chomsky. *The minimalist program*. MIT press, 2014.

Stephen Crain and Paul Pietroski. Why language acquisition is a snap. *The Linguistic Review*, 19:163–183, 2002.

Naiara Minto de Sousa, Lucas Tadeu Garcia, and Maria Stella Coutinho de Alcantara Gil. Differential reinforcement in simple discrimination learning in 10-to 20-month-old toddlers. *The Psychological Record*, 65(1):31–40, 2015.

Stephanie Denison, Christie Reed, and Fei Xu. The emergence of probabilistic reasoning in very young infants. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

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

David Dowty. Thematic proto-roles and argument selection. *Language*, pages 547–619, 1991.

B. Elan Dresher. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30(1):27–67, 1999.

Naomi Feldman, Thomas Griffiths, Sharon Goldwater, and James Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778, 2013.

Alissa L Ferry, Ana Fló, Perrine Brusini, Luigi Cattarossi, Francesco Macagno, Marina Nespor, and Jacques Mehler. On the edge of language acquisition: Inherent constraints on encoding multisyllabic sequences in the neonate brain. *Developmental Science*, 19(3):488–503, 2016.

Charles J Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart, & Winston, Inc, New York, 1968.

József Fiser and Richard N Aslin. Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, 99(24):15822–15826, 2002.

Ana Fló, Perrine Brusini, Francesco Macagno, Marina Nespor, Jacques Mehler, and Alissa L Ferry. Newborns are sensitive to multiple cues for word segmentation in continuous speech. *Developmental Science*, page e12802, 2019.

Janet D. Fodor. Unambiguous Triggers. *Linguistic Inquiry*, 29:1–36, 1998.

Daniel Freudenthal, Julian M Pine, Javier Aguado-Orea, and Fernand Gobet. Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, 31(2):311–341, 2007.

Daniel Freudenthal, Julian M Pine, and Fernand Gobet. Simulating the referential properties of Dutch, German, and English root infinitives in MOSAIC. *Language Learning and Development*, 5(1):1–29, 2009.

Daniel Freudenthal, Julian Pine, and Fernand Gobet. 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, 2010.

Daniel Freudenthal, Julian M Pine, Gary Jones, and Fernand Gobet. 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*, pages 746–751, Pasadena, 2015.

LouAnn Gerken. Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67–B74, 2006.

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

Edward Gibson and Kenneth Wexler. Triggers. *Linguistic Inquiry*, 25(4):407–454, 1994.

Adele E Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

Adele E Goldberg. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand, 2006.

Adele E Goldberg. Argument structure constructions versus lexical rules or derivational verb templates. *Mind & Language*, 28(4):435–465, 2013.

Nelson Goodman. *Fact, fiction and forecast*, volume 74. JSTOR, 1955.

Jane Grimshaw. *Argument structure.* The MIT Press, Cambridge, MA, 1990.

C.-T. James Huang. *Logical relations in Chinese and the theory of grammar.* PhD thesis, MIT, Cambridge, MA, 1982.

Robert C Hulsebus. Operant conditioning of infant behavior: A review. In *Advances in Child Development and Behavior*, volume 8, pages 111–158. Elsevier, 1974.

Ava Irani. How Children Learn to Disappear Causative Errors. In Megan M. Brown and Brady Dailey, editors, *Proceedings of the 43rd Boston University Conference on Language Developmen*, pages 298–310. Cascadilla Press, Somerville, MA, 2019.

Ray Jackendoff. The status of thematic relations in linguistic theory. *Linguistic inquiry*, 18(3):369–411, 1987.

Ray S Jackendoff. *Patterns in the mind: Language and human nature.* Basic Books, 1994.

Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

Charles Kemp, Amy Perfors, and Joshua Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 10(3):307–321, 2007.

Natasha Z Kirkham, Jonathan A Slemmer, and Scott P Johnson. Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2):B35–B42, 2002.

Richard K Larson. Double objects revisited: Reply to Jackendoff. *Linguistic Inquiry*, 21(4):589–632, 1990.

Howard Lasnik and Mamuro Saito. On the nature of proper government. *Linguistic Inquiry*, 15:235–289, 1984.

Stephen Laurence and Eric Margolis. The poverty of the stimulus argument. *The British Journal for the Philosophy of Science*, 52(2):217–276, 2001.

Julie Legate and Charles Yang. Morphosyntactic Learning and the Development of Tense. *Linguistic Acquisition*, 14(3):315–344, 2007.

Julie Legate and Charles Yang. Assessing Child and Adult Grammar. In Robert Berwick and Massimo Piatelli-Palmarini, editors, *Rich Languages from Poor Inputs*, pages 168–182. Oxford University Press, Oxford, UK, 2013.

David Lightfoot. The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences*, 12:321–334, 1989.

Lewis P Lipsitt, Linda Johnson Pederson, and Clement A Delucia. Conjugate reinforcement of operant responding in infants. *Psychonomic Science*, 4(1): 67–68, 1966.

M Manzini and K Wexler. Parameters, binding theory, and learnability. *Linguistic Inquiry*, 18:413–444, 1987.

William Garrett Mitchener and Misha Becker. Computational models of learning the raising-control distinction. *Research on Language and Computation*, 8(2-3):169–207, 2010.

Partha Niyogi and Robert C. Berwick. A language learning model or finite parameter spaces. *Cognition*, 61:161–193, 1996.

Naho Orita, Rebecca McKeown, Naomi H Feldman, Jeffrey Lidz, and Jordan Boyd-Graber. Discovering pronoun categories using discourse information. In *Proceedings of the 35th annual conference of the cognitive science society*, 2013.

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

Lisa Pearl. Poverty of the stimulus without tears. University of California, Irvine, 2019. URL https://ling.auf.net/lingbuzz/004646.

Lisa Pearl. Modeling syntactic acquisition. In Jon Sprouse, editor, *Oxford Handbook of Experimental Syntax*. Oxford University Press, in press.

Lisa Pearl and Jeffrey Lidz. Parameters in Language Acquisition. In Kleanthes Grohmann and Cedric Boeckx, editors, *The Cambridge Handbook of Biolinguistics*, pages 129–159. Cambridge University Press, Cambridge, UK, 2013.

Lisa Pearl and Benjamin Mis. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, 92(1):1–30, 2016.

Lisa Pearl and Jon Sprouse. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20:19–64, 2013a.

Lisa Pearl and Jon Sprouse. Computational Models of Acquisition for Islands. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Islands Effects*, pages 109–131. Cambridge University Press, Cambridge, 2013b.

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

Lisa Pearl and Jon Sprouse. The acquisition of linking theories: A Tolerance and Sufficiency Principle approach to learning UTAH and rUTAH. 2019. URL https://ling.auf.net/lingbuzz/004088.

Lisa Pearl and Jon Sprouse. Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language*, in press. URL https://ling.auf.net/lingbuzz/003913.

Lisa Pearl, Kristine Lu, and Anousheh Haghighi. The character in the letter: Epistolary attribution in Samuel Richardson's *Clarissa*. *Digital Scholarship in the Humanities*, 32(2):355–376, 2017.

Amy Perfors, Joshua Tenenbaum, and Elizabeth Wonnacott. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(3):607 – 642, 2010.

Amy Perfors, Joshua Tenenbaum, Thomas Griffiths, and Fei Xu. A tutorial introduction ot bayesian models of cognitive development. *Cognition*, 120 (3):302–321, 2011a.

Amy Perfors, Joshua Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118:306–338, 2011b.

David M Perlmutter and Paul M Postal. Impersonal passives and some relational laws. *Studies in Relational Grammar*, 2:126–170, 1984.

Steven Piantadosi, Joshua Tenenbaum, and Noah Goodman. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217, 2012.

John Ross. *Constraints on variables in syntax*. PhD thesis, MIT, Cambridge, MA, 1967.

Carolyn Kent Rovee and David T Rovee. Conjugate reinforcement of infant exploratory behavior. *Journal of Experimental Child Psychology*, 8(1):33–39, 1969.

Carolyn Kent Rovee-Collier and Joanne Bitetti Capatides. Positive behavioral contrast in 3-month-old infants on multiple conjugate reinforcement schedules. *Journal of the Experimental Analysis of Behavior*, 32(1):15–27, 1979.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

Jenny R Saffran, Elizabeth K Johnson, Richard N Aslin, and Elissa L Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.

William Sakas. Computational approaches to parameter setting in generative linguistics. In Jeffrey Lidz, William Snyder, and Joe Pater, editors, *The Oxford Handbook of Developmental Linguistics*, pages 696–724. Oxford University Press, Oxford, UK, 2016.

William Sakas and Janet Fodor. Disambiguating Syntactic Triggers. *Language Acquisition*, 19(2):83–143, 2012.

William Sakas and Janet D. Fodor. The Structural Triggers Learner. In Stefano Bertolo, editor, *Language Acquisition and Learnability*, pages 172–233. Cambridge University Press, Cambridge, UK, 2001.

Barbara Scholz and Geoffrey Pullum. Irrational nativist exuberance. In Robert Stainton, editor, *Contemporary Debates in Cognitive Science*. 2006.

Kathryn Schuler, Charles Yang, and Elissa Newport. 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*, 2016.

Margaret Speas. *Phrase structure in natural language*, volume 21. Springer Science & Business Media, 1990.

Aimee E Stahl, Alexa R Romberg, Sarah Roseberry, Roberta Michnick Golinkoff, and Kathryn Hirsh-Pasek. Infants segment continuous events using transitional probabilities. *Child Development*, 85(5):1821–1826, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Rachel Wu, Alison Gopnik, Daniel C Richardson, and Natasha Z Kirkham. Infants learn about objects from statistics and people. *Developmental Psychology*, 47(5):1220, 2011.

Fei Xu and Joshua Tenenbaum. Word Learning as Bayesian Inference. *Psychological Review*, 114(2):245–272, 2007.

Charles Yang. *Knowledge and Learning in Natural Language.* Oxford University Press, Oxford, UK, 2002.

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

Charles Yang. On productivity. *Yearbook of Language Variation*, 5:333–370, 2005.

Charles Yang. Computational models of syntactic acquisition. *WIREs Cognitive Science*, 3:205–213, 2012.

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

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

Charles Yang. How to wake up irregular (and speechless). *On looking into words (and beyond)*, pages 211–233, 2017.