Poverty of the Stimulus Without Tears

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

Abstract

Poverty of the stimulus has been at the heart of ferocious and tear-filled debates at the nexus of psychology, linguistics, and philosophy for decades. This review is intended as a guide for readers without a formal linguistics or philosophy background, focusing on what poverty of the stimulus is and how it's been interpreted, which is traditionally where the tears have come in. I discuss poverty of the stimulus from the perspective of language development, highlighting how poverty of the stimulus relates to expectations about learning and the data available to learn from. I describe common interpretations of what poverty of the stimulus means when it occurs, and approaches for determining when poverty of the stimulus is in fact occurring. I close with illustrative examples of poverty of the stimulus in the domains of syntax, lexical semantics, and phonology, and discuss the value of identifying instances of poverty of the stimulus when it comes to understanding language development.

Keywords: poverty of the stimulus, nativism, quantitative approaches, syntax, lexical semantics, phonology

1 Introduction

"Poverty of the stimulus" is essentially a claim about the data available to children when they're trying to learn certain pieces of knowledge (more on this in the next section). Notably, poverty of the stimulus has been at the heart of debates in cognitive science – in particular, the intersection of psychology, linguistics, and philosophy – for decades. Here are just some of the papers focusing on it over the years: Chomsky (1965); Kimball (1973); Chomsky (1975); Baker (1978); Stich (1978); Pinker (1979); Chomsky (1980a, 1980b); Baker and McCarthy (1981); Hornstein and Lightfoot (1981); Gordon (1986); Chomsky (1988); Lasnik (1989); Lightfoot (1989); Pinker (1989); Sampson (1989); Crain (1991); Ramsey and Stich (1991); Wexler (1991); Marcus (1993); Garfield (1994); Jackendoff (1994); Morgan, Bonamo, and Travis (1995); Pullum (1996); Seidenberg (1997); Ambridge, Rowland, and Pine (2008); Lightfoot (1998); Cowie (1998); Laurence and Margolis (2001); Matthews (2001); Crain and Pietroski (2002); Fodor and Crowther (2002); Legate and Yang (2002); Pullum and Scholz (2002); Scholz and Pullum (2002); Chouinard and Clark (2003); Dresher (2003); Collins (2004); Pinker (2004); Idsardi (2005); Scholz and Pullum (2006); Ambridge et al. (2008); Valian (2009); Hsu and Chater (2010); Perfors, Tenenbaum, and

Wonnacott (2010); Clark and Lappin (2011); Hsu, Chater, and Vitányi (2011); Perfors, Tenenbaum, and Regier (2011); Hsu, Chater, and Vitányi (2013); Pearl and Sprouse (2013b, 2013a); Rey (2014); Chater, Clark, Goldsmith, and Perfors (2015); King (2015); Han, Musolino, and Lidz (2016); Pearl and Mis (2016); Skidelsky (2016); Abend, Kwiatkowski, Smith, Goldwater, and Steedman (2017); Belletti (2017); Fodor and Sakas (2017); Lasnik and Lidz (2017); Piattelli-Palmarini (2017); Zenker and Schwartz (2017); Fitz and Chang (2017); Lidz (2018); McCoy, Frank, and Linzen (2018); Pearl and Sprouse (2018); and Rawski and Heinz (2019). Interestingly, people who have written on poverty of the stimulus are often quite exercised by it. Much of the ink spilled above has been spent expressing (in my opinion) rather strong feelings, in addition to actual debate. So, why has poverty of the stimulus gotten so much continuous – and often quite feisty – attention? I believe that this is because of the implications people take poverty of the stimulus to have, which I'll discuss in more detail throughout this article.

More generally, I'll try to give an overview of what poverty of the stimulus is and how it's been interpreted, which is really where all the tears have traditionally come in. My goal is to accomplish this overview without the tears, as much as possible, particularly for readers without a formal linguistics or philosophy background. To do this, I approach poverty of the stimulus from the perspective of language development, and focus on how poverty of the stimulus relates to expectations about learning and the data available to learn from. I then describe how people interpret poverty of the stimulus when it occurs. Because a lot of debate surrounding poverty of the stimulus is actually about *when* poverty of the stimulus occurs, I'll discuss some approaches people have taken for determining if poverty of the stimulus is in fact occurring. These approaches often focus on determining how much data is required for children to learn what they do when they do; if the quantity of data required is more than the quantity of data children seem to get, then poverty of the stimulus is occurring. I'll then discuss some potential examples of poverty of the stimulus and how I think it best to interpret them, given our current understanding. I'll conclude with what seem to me to be the best ways to make progress on determining (i) when poverty of the stimulus occurs, and (ii) what it means when it does, particularly if we're interested in understanding language development better.[1]

1.1 Alright, what is it?

At its heart, poverty of the stimulus is a developmental claim, i.e., a claim about whether something is in fact learnable (and more practically, learnable by typically-developing children). Poverty of the stimulus occurs when there's unresolvable ambiguity in the data – in this way, the data are insufficient on their own (i.e., the stimulus is too "impoverished") for children to infer the correct answer (Laurence & Margolis, 2001; Scholz & Pullum, 2002; Clark & Lappin, 2011; Chater et al., 2015; Skidelsky, 2016; Lasnik & Lidz, 2017). So, unresolvable ambiguity causes an induction problem; this means that "poverty of the stimulus" is just another way of saying that there's an induction problem for the child. Researchers are then rather impressed when children nonetheless

[1] I recognize that this review is rather long, as I attempted to draw together information on many different aspects of poverty of the stimulus. Given this, some readers may find it helpful to skip to sections targeting aspects of poverty of the stimulus that they're particularly interested in, after reading the introductory section.

seem to reliably resolve that ambiguity, and end up with the right answer despite the induction problem (Chomsky, 1975; Ramsey & Stich, 1991; Garfield, 1994; Laurence & Margolis, 2001; Matthews, 2001; Collins, 2004).

I think it's helpful to see this kind of unresolvable ambiguity more concretely, as in Figure 1. Here we see a two-dimensional space, where some data points are observed (each indicated with an X). Now, let's assume learners believe that observable data are generated by speakers using some internal representation (e.g., an observed vowel pronunciation generated from an underlying vowel definition of some kind, or an observed utterance generated from an underlying syntactic rule set of some kind). So, for this kind of learner, the data in Figure 1 were generated from some internal representation the speakers used. The learner is then trying to figure out what that representation was, on the basis of these observed data. In a two-dimensional space like what's shown in Figure 1, a representation might correspond to a mean value for each dimension, which has some variance in each dimension (leading to ellipses like those shown in Figure 1). So, speakers generate data points that fall within the bounds of that internally-defined ellipse. This means the child's job is to figure out which ellipse, corresponding to a particular internal representation, is the one speakers used to generate those data. Some sample hypotheses are shown (A-F), with the correct one (C) indicated with a dashed line. The problem here is that *all* these hypotheses are compatible with the data. That is, these data could be generated by a speaker who had representations corresponding to A, B, C, D, E, or F. So, these data are ambiguous about which hypothesis is correct, and this ambiguity is unresolvable without some kind of additional guidance about which one to pick. Thus, an induction problem, i.e., poverty of the stimulus.

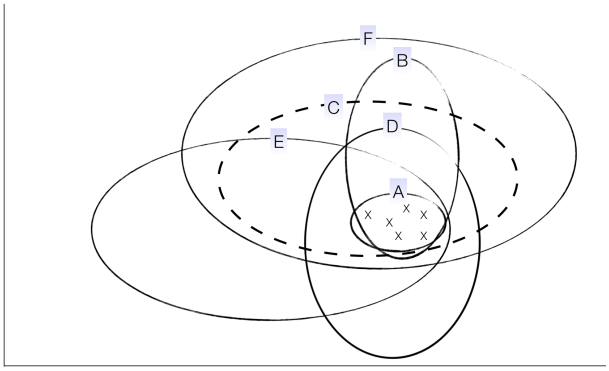


Figure 1: A visual demonstration of poverty of the stimulus for two-dimensional data. Each X corresponds to an observed data point, and A-F correspond to potential representations that speakers could use to generate the observed data. The correct representation (C) is in dashed lines. Poverty of the stimulus occurs because all these representations are compatible with the data.

Now, imagine if all the children we observe always end up with hypothesis C. Maybe they explore some other hypotheses along the way (like A or B or F), but they all still converge on C as the underlying representation for these data. This would be very surprising. Put simply, why C? What does C have going for it that the other hypotheses don't? Why not A or B or F as the final answer, given these data? What's causing this *constrained generalization* to C? This constrained generalization becomes even more surprising when the hypothesis space has more than these 6 hypotheses in it. Generally, language scientists think of children's hypothesis spaces being quite vast. So, when we observe poverty of the stimulus and we see constrained generalization – that is, children *not* picking hypotheses that are perfectly compatible with the available data and so avoiding "tempting errors" (Laurence & Margolis, 2001) – we get very interested in why children are doing this.

This is really where all the interest in poverty of the stimulus comes from. Poverty of the stimulus means the data are insufficient for explaining how children make the generalizations they do. This is because the data are compatible with multiple hypotheses that children seem to ignore (Laurence & Margolis, 2001; Belletti, 2017; Piattelli-Palmarini, 2017). So, children must be using something else to help them decide among the possible hypotheses – and that something else doesn't come from the data themselves. It has to come from somewhere else. But, where else *is* there, if the data are the only external signal the child has to work with? The answer is that the something else is internal – it comes from the child herself. Prior to learning from these data, the child already knows something (or is able to do something), and that something helps her navigate through these hypotheses that are equally compatible with the data. That is, the prior something (sometimes called the child's "inductive bias") allows the child to make constrained generalizations.[2]

So, the argument from poverty of the stimulus can be summed up as in (1), which shows why poverty of the stimulus *and* children making constrained generalizations should lead us to believe children have prior knowledge or abilities. I want to note that tears often arise about the nature of the prior knowledge or abilities required (more discussion of this in section 1.4), though sometimes tears do come in at other points in the reasoning process described in (1), as discussed in other parts of this review.

(1) The argument from poverty of the stimulus: Poverty of the stimulus and constrained generalizations together imply prior knowledge or abilities
   a. **Data**: There are data external to the child that are available for learning about some underlying representation humans have (e.g., about language).
   b. **Poverty of the stimulus**: These data are compatible with more than one hypothesis about the underlying representation. This is why they're considered impoverished or insufficient – they don't pinpoint the correct hypothesis on their own.
   c. **Constrained generalization**: Children figure out the correct representation anyway.
   d. **Prior knowledge or abilities**: Therefore, children have prior knowledge or abilities that cause them to make the constrained generalization.

1.2 How do we tell that data are insufficient?

One key component to this chain of reasoning is the actual poverty of the stimulus part itself, i.e., that the data are insufficient on their own to guide the child to the correct answer. This is why children have to somehow go beyond the information provided in those data (Ramsey & Stich, 1991; Wexler, 1991; Pullum & Scholz, 2002; Skidelsky, 2016; Lasnik & Lidz, 2017). Above, I described this insufficiency as the data being compatible with more than one hypothesis (with the visual example in Figure 1). However, there's been considerable discussion (sometimes with tears) of precisely how data can be insufficient, which I'll briefly summarize now. Where possible, I'll connect each idea to the poverty of the stimulus definition used above, where the data are

[2]It's important to underscore that "prior" isn't the same as "innate", something which will be discussed more thoroughly in section 1.3.

compatible with more than one hypothesis. I also want to note that only the first idea mentioned below (where the data just aren't there) is in line with the poverty of the stimulus definition here. The other ideas are often related, but don't necessarily mean the data are compatible with more than one hypothesis. (This has been a definite source of confusion in the debate surrounding poverty of the stimulus, and so has caused some tears.) For each of those other ideas, I'll note why those ideas have been connected to poverty of the stimulus, even though they themselves may not be poverty of the stimulus.

The data just aren't there. One way poverty of the stimulus is discussed is that the required data simply aren't available (Chomsky, 1965; Kimball, 1973; Baker, 1978; Hornstein & Lightfoot, 1981; Gordon, 1986; Lasnik, 1989; Ramsey & Stich, 1991; Lightfoot, 1998; Legate & Yang, 2002; Pullum & Scholz, 2002; Lasnik & Lidz, 2017; Zenker & Schwartz, 2017; Lidz, 2018). What does it mean for data to be required? Required data are those that pinpoint the correct hypothesis. For example, there are no data points in the hypothesis space in Figure 1 that would uniquely identify hypothesis C as the correct hypothesis. This is because every data point in C is also in F (i.e., C is a subset of F), so even the least ambiguous "C" data points still have unresolvable ambiguity between C and F (i.e., there's a "Subset Problem" caused by F if the child needs to learn C. In particular, there is no obvious C-but-not-F data point available – and this is exactly what's caused both significant worry and theories for how children deal with it (Brown & Hanlon, 1970; Bowerman, 1988; Marcus, 1993; Hsu et al., 2013; Chater et al., 2015; Lasnik & Lidz, 2017; Pearl, in press-b).[3] This situation aligns exactly with our current definition of poverty of the stimulus – multiple hypotheses are compatible with the available data. In this particular case, multiple hypotheses will *always* be compatible with data points covered by the correct hypothesis C.

As a concrete example of this situation, suppose we consider *wh*-words like *what* in English, and where they appear.[4] Let's suppose hypothesis C is that *wh*-words must always appear at the front of the clause (this is the default for English); so, C predicts a simple question (which is a main clause) to have a form like (2a) and embedded clauses (indicated with [...] in (2)) to have forms like (2b) and (2c). In each case, the *what* is at the front of the clause. Suppose that hypothesis F is that *wh*-words can either appear at the front of the clause or instead remain in place where they're understood ("in-situ"), as in (2d)-(2f). So, F includes data that aren't the default for English (though they may be acceptable in certain situations, discussed more below) – I've indicated these with a * in (2).

[3] One modern solution to the Subset Problem involves a domain-general ability to leverage ambiguous C-and-F data by considering how likely each hypothesis is to generate those kind of data (here, C is more likely than F to generate C-and-F data). This has been called the Size Principle (Tenenbaum & Griffiths, 2001), and is discussed recently in more detail for the Subset Problem in Lasnik and Lidz (2017) and Pearl (in press-a). This ability to leverage ambiguous data would allow children to use ambiguous data as *indirect evidence* for C (and against F). Indirect evidence is an evidence type discussed later on in this section when considering the data available to children.

[4] I want to note that this is a hypothetical example, rather than one motivated by a particular empirical debate. However, I think this example sketches out a potential hypothesis space for the child that makes the distinctions I discuss later on easier to follow.

(2) *wh*-word position examples

a. Main clause, fronted: "*What did this penguin do?*"
b. Embedded clause, fronted: "*I saw [what this penguin did].*"
c. Doubly embedded clause, fronted: "*I thought [I saw [what this penguin did]].*"
d. *Main clause, in-situ: "*This penguin did what?*"
e. *Embedded clause, in-situ: "*I saw [this penguin did what].*"
f. *Doubly embedded clause, in-situ: "*I thought [I saw [this penguin did what]].*"

This means that hypothesis C allows only the first three examples in (2) to occur, while hypothesis F allows all the examples in (2) to occur. So, if hypothesis C is correct (and an English child has to learn that from this hypothesis space), there's no obvious C-but-not-F data point available: the three data points compatible with C ((2a)-(2c)) are also compatible with F.

A lot of debate (and tears) then happens about exactly what kind of data are required to pinpoint the correct hypothesis. This in turn depends on quite a number of things: (i) what the correct hypothesis is, (ii) what the hypothesis space looks like, (iii) what data are available, and (iv) how children are able to leverage those available data. In particular, there can often be significant disagreement about any of these components, with specific ideas for each component changing over time as we learn more about how humans represent linguistic knowledge (component (i)), how children represent linguistic knowledge (component (ii)), what children's input looks like (component (iii)), and how children's information processing abilities develop (component (iv)).[5]

To begin with, it matters quite a lot what the hypothesis space looks like. Imagine if the hypothesis space was like the one shown in Figure 2 – there are certainly possible data points that would uniquely pinpoint C in this hypothesis space (some are shown as Xs in Figure 2). That is, if those unambiguous C data were available, no poverty of the stimulus would happen.

As a concrete example with the position of *wh*-words, suppose hypothesis E is that *wh*-words must always appear at the front of main clauses (like (2a)) but must always remain in place in embedded clauses (like (2e)-(2f)). (Remember that hypothesis C is that *wh*-words must always appear at the front of both main and embedded clauses, corresponding to (2a)-(2c).) So, while both C and E allow (2a), only C allows (2b)-(2c). This means that data points like (2b)-(2c) would correspond to Xs in Figure 2 and signal unambiguously that hypothesis C is correct.

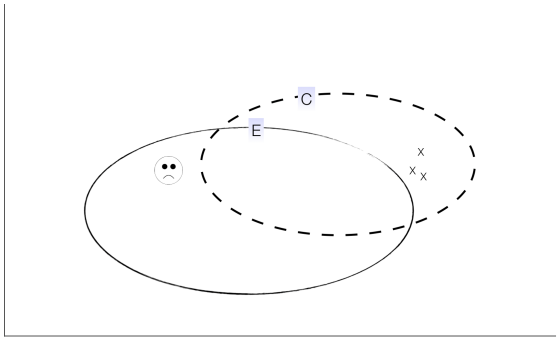
Let's turn now to the availability of data to children, which is also a place where a lot of tears have historically occurred. For instance, researchers have often described children as learning primarily (or only) from what's called *direct positive evidence* (Pinker, 1989; Seidenberg, 1997; Marcus, 1993; Laurence & Margolis, 2001; Fodor & Crowther, 2002; Fodor & Sakas, 2017). In Figure 2, this means that to learn C is correct, children need to encounter data that are only compatible with C (like the Xs shown). These data are then positive examples of hypothesis C that children directly observe.

Direct positive evidence contrasts with both *negative* evidence and *indirect* evidence. Negative evidence is evidence about what's *not* the correct hypothesis. An example would be the frowny face in Figure 2. Continuing with the *wh*-word example, this might be a data point where the *wh*-word is in-situ in an embedded clause, like (2e)-(2f) – these data points are compatible with E but not C.

[5]For a concrete example of changing ideas, see the discussion in section 3.1 about structure dependence.

One way to think about negative evidence is that someone directly tells the child that this data point isn't in fact right (i.e., it's not covered by the correct hypothesis C - something like "You can't say *I saw this penguin did what*."). However, we don't often go around telling people what isn't so unless they've somehow indicated they think it might be so. Here, this might correspond to children indicating they think that data point is part of the correct hypothesis, and then being corrected (i.e., the child says "I saw this penguin did what" and her caretaker replies, "No, honey, you can't say that."). But, it turns out that children aren't often explicitly corrected (Bowerman, 1988); even when they are, they seem to ignore those explicit corrections (McNeill, 1996; Brown & Hanlon, 1970). Still, there are other kinds of corrective feedback they receive (Quine, 1960; Morgan et al., 1995; Saxton, 1997; Saxton, Kulcsar, Marshall, & Rupra, 1998; Saxton, 2000; Chouinard & Clark, 2003; Saxton, 2005; King, 2015), and this feedback potentially highlights the "negative" status of the data point (i.e., that this data point isn't covered by the correct hypothesis).

Figure 2: A hypothesis space where there are unambiguous data available for the correct hypothesis C, shown with dashed lines. Unambiguous (direct positive evidence) data points for C are shown with Xs. A negative evidence data point for C is shown with a frowny face.



As a concrete example of corrective feedback, let's consider recasts (Saxton, 1997; Saxton et al., 1998; Saxton, 2000, 2005), where the child says something incorrect and the caretaker responds by recasting that incorrect thing into something correct. In Figure 2, this might occur if the child said the frowny face data point, and the caretaker responded by recasting it as an X data point. A language example might be something like the child saying "I saw this penguin did what", and the caretaker recasting that as "Right, you saw what this penguin did": the in-situ *what* is recast as fronted *what* in the embedded clause. In this way, the caretaker has indirectly communicated the negative status of the frowny face data point – that is, if the child is astute enough to notice the recast. This particular issue of the child noticing the recast as negative evidence highlights why even this kind of corrective feedback may also have limited effectiveness (Morgan et al., 1995).

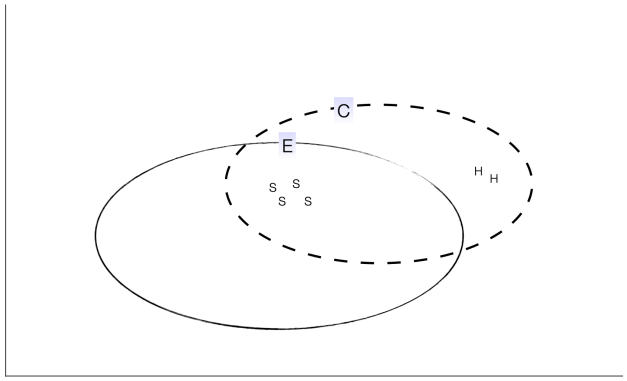
This relates to the second distinction I mentioned above: indirect evidence. Indirect evidence means the data point isn't directly about the correct hypothesis – instead, the child has to infer what information from that data point is relevant for pinpointing the correct hypothesis. For recasts, the indirect part was recognizing that what the caretaker said was a replacement to be used instead of the frowny face data point. More specifically, in the case of the fronted *what* recast for in-situ *what*, the child has to infer that the caretaker wasn't saying fronted *what* simply because she felt like it; instead, the child must infer that fronted *what* is the correct way to express that meaning; this makes the recast, if reasoned about correctly by the child, indirect negative evidence for in-situ *what*.[6] This train of required reasoning also highlights how harnessing indirect evidence is likely to take more effort from the child.

[6]When indirect evidence is also negative evidence like this, it's sometimes referred to as "implicit negative evidence" (Rohde & Plaut, 1999) rather than "indirect negative evidence".

So, the availability of indirect evidence in children's input very much depends on the child's ability to make (potentially sophisticated) inferences. In theory, many types of data could function as indirect evidence (typically indirect negative evidence) for a particular hypothesis. For example, we saw that the recast of fronted *what* for in-situ *what* can function as indirect negative evidence for hypotheses that front *wh*-words. It could also be that *echo questions*, which have distinctive prosody (e.g, *You saw that this penguin did WHAT?"*), function as indirect negative evidence in favor of *wh*-word fronting; however, this is only true if children infer from this distinctive prosody that in-situ *wh*-words are unusual – and therefore, fronted *wh*-words are the norm. But, can children in practice harness one or both of these types of indirect negative evidence for *wh*-word fronting? It really hinges on children's ability to make these inferences.

This then relates to the final point: we need to be clear about how we think children are able to leverage the available data. If data are available in the input, but children can't use them (perhaps because these data are just too indirect), we might well have poverty of the stimulus after all. For instance, imagine a scenario like Figure 3, where C is the correct hypothesis and there are in fact unambiguous data available, but they're hard for the child to use for whatever reason (these are labeled H in Figure 3).

Figure 3: A hypothesis space with different types of data available for the correct hypothesis C, shown with dashed lines. Data points that are hard to use are shown with an H, while data points that are simpler to use are shown with an S.

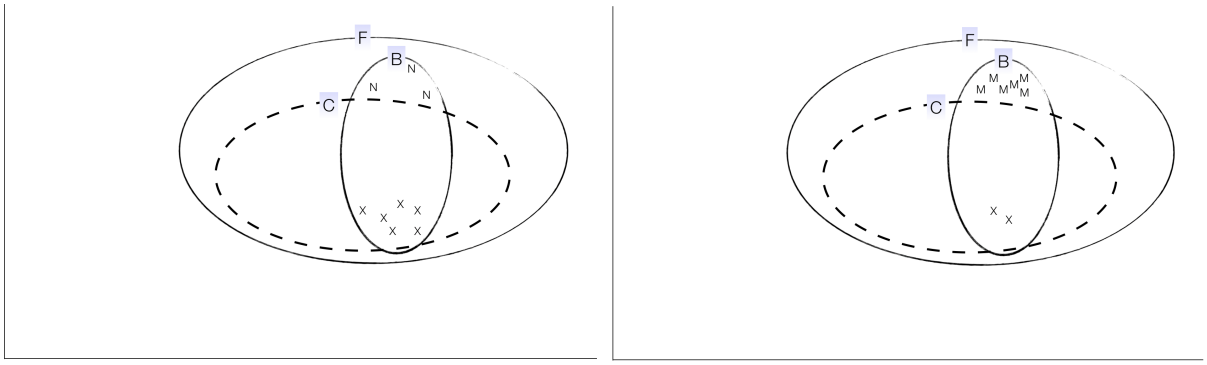


In our *wh*-word example, these might correspond to data points like (2c) where the *wh*-word is in an embedded clause inside a second embedded clause (i.e., a doubly embedded clause). These data points could be hard for a child to use because her developing processing abilities might not yet let her accurately process a doubly embedded clause. If a particular child is in fact unable to use the hard data points, this leaves that child with a smaller set of (simpler) useable data (these are labeled S in Figure 3 – in our *wh*-word example, they would correspond to data points like (2a) that involve main clauses). Then, these data do in fact cause poverty of the stimulus because they're compatible with multiple hypotheses (correct hypothesis C, and incorrect hypothesis E).

The data are noisy. Because of speech errors on the part of the speakers or processing errors on the part of the child listeners, the data may be "noisy" (Chomsky, 1965; Hornstein & Lightfoot, 1981; Ramsey & Stich, 1991; Garfield, 1994; Seidenberg, 1997; Laurence & Margolis, 2001; Piattelli-Palmarini, 2017). I want to emphasize that noisy data don't automatically cause poverty of the stimulus. In fact, Clark and Lappin (2011) note that noise in the input signal is something children have to deal with in many areas of cognitive development (e.g., auditory perception and vision). The issue related to poverty of the stimulus is that the data are noisy in a way that leads to unresolvable ambiguity.

For example, let's consider the left panel of Figure 4, where C is the correct hypothesis and potential noisy data are shown as Ns. For our *wh*-word example, suppose hypothesis B is that *wh*-words can be either fronted or in-situ in main clauses, but must be fronted in embedded clauses. So, the N data points could be like (2d), with the *wh*-word in-situ in the main clause (*This penguin did what?*). This situation might occur if someone is speaking English as a second language, and her first language is one where *wh*-words aren't fronted in main clauses, such as Japanese, Hungarian, Igbo, or Zulu. That is, by accident, the speaker transfers her use of *wh*-words from her native language to English, and creates these kind of noisy data points in English children's input.

Figure 4: A visual demonstration of noisy data (left panel) and misleading data (right panel). Each X corresponds to an observed non-noisy data point while each N is a noisy data point in the left panel and each M is a misleading data point in the right panel. The correct representation (C) is in dashed lines.



These N noisy data points are compatible with B and F, but not compatible with C. So, these noisy data would cause poverty of the stimulus because all the data (that is, the data points shown as Xs as well as the additional noisy N data points) are *still* compatible with multiple hypotheses. In this case, those noisy data points would make the entire available data compatible with both B and F.

However, separate from the poverty of the stimulus issue, these data unfortunately would rule out the correct hypothesis C. This is the (other) main problem with noisy data, and the one that seems to get more discussion: noisy data throw the child off the right track. This isn't poverty of the stimulus necessarily, because there may be only one hypothesis left after the noisy data are considered. However, that one hypothesis isn't the outcome we want (remember: children should end up with C in this scenario). So, children have to ignore the noisy data somehow. But again, this is a separate issue from poverty of the stimulus, which is when noisy data create unresolvable ambiguity.

The data are misleading. This is related to the idea of noisy data, where perhaps due to the vagaries of what people choose to talk about or how children interpret the data, the data are misleading. That is, it's not that speakers are making errors and so generating noise; rather, something internal to children (e.g., their developing knowledge or their developing processing capabilities) causes children to perceive these perfectly correct data incorrectly. Belletti (2017) describes an

instance where an incorrect hypothesis seems to be the one the data suggest, if the child is viewing the data in a particular (unhelpful) way. So, the child is then thrown off the right track.

In our *wh*-word example, we might again have data points where the *wh*-word is in-situ in the main clause (as in (2d): *The penguin did what?*), but these are produced by native English speakers. Remember that in English this form is called an *echo question*; echo questions are often used in casual conversation (perhaps the speaker is surprised by what the penguin did, with the *what* emphasized) and teaching scenarios ("Now, do you see the penguin? The penguin did something. Tell me: The penguin did what?"). English children have to be savvy that these are special uses of *wh*-words in main clauses, and there's typically some kind of clue that they're special. Perhaps the speaker's prosody isn't the default (as when a speaker emphasizes the *what*) or the context is a special teaching context. But, suppose children miss the cue that these data are special uses; then, these data can be misleading data that seem to signal it's perfectly okay to just leave the *wh*-word in-situ in main clauses as a default word order for English (i.e., without special prosody or a special context).

In the right panel of Figure 4, we see this situation: many misleading data (shown with Ms) appear, and these data can't be captured by the correct hypothesis C. So, these data would mislead the child, in this case away from generalization C. As before with noisy data, this situation is actually separate from poverty of the stimulus: while misleading data might cause poverty of the stimulus, they don't have to. In the right panel of Figure 4, they do cause poverty of the stimulus – the child again sees data that are compatible with more than one hypothesis (here: B and F). But the real problem, and the one that causes much discussion, is the fact that the correct hypothesis (C) isn't one of the compatible hypotheses. Again, as with noisy data, this is a separate issue from poverty of the stimulus.

Children learn so quickly and uniformly. Another point that often comes up is how quickly typically-developing children learn language, given those children's immature cognitive abilities (Jackendoff, 1994; Laurence & Margolis, 2001; Crain & Pietroski, 2002). This is related to the concept of computational complexity or computational efficiency in formal learnability studies (Clark & Lappin, 2011) – in essence, how complex is it to navigate to the right answer and how efficiently can that navigation be done, given the data? If children compute something complex very fast, this is pretty surprising.

The surprise at children's quickness can link to poverty of the stimulus, though it does so implicitly. In particular, if we have poverty of the stimulus, as in Figures 1, 3, or 4, the fact that children converge on the right answer at all is already surprising. The fact that children do it despite not being as cognitively adept as adults is really a bonus surprise. This surprise at children's quickness may also relate more directly to the idea of noisy or misleading data; when those are present in the input, it's surprising that children can so quickly navigate to the right answer.

An associated idea is that children learn uniformly – that is, typically-developing children all seem to end up with the same answer (Chomsky, 1965; Crain & Pietroski, 2002; Lasnik & Lidz, 2017). This is again a surprise if we assume poverty of the stimulus is happening. In fact, recall from (1), where I summarized the argument from poverty of the stimulus, that children ending up with the same (correct) answer is the "constrained generalization" part that comes *after* we see

that poverty of the stimulus exists. That is, children showing constrained generalization is part of the argument *from* poverty of the stimulus that leads to the idea that children must have prior knowledge (as the papers cited above are careful to note); this contrasts with constrained generalization being an argument *in favor of* poverty of the stimulus. In particular, children showing "constrained generalization" depends on poverty of the stimulus *already* being established – it's only really surprising that children all end up with the same answer if there was more than one answer possible.

1.3 So what does it mean if the data are insufficient?

Wait, does that really happen? There's definitely debate (and tears) about when the data do in fact seem insufficient (e.g., see Morgan, 1986; Pinker, 1989; Pullum & Scholz, 2002; Scholz & Pullum, 2006; Ambridge et al., 2008; Chater et al., 2015). This has a lot to do with what goes into figuring out all those components I mentioned above: what the correct hypothesis is, what the hypothesis space looks like, what data really are available in children's input, and how children are able to leverage those available data. In order to claim poverty of the stimulus is occurring, we need to assume something about each of these components. Only then can we concretely specify the data that are available (and useable), which was the "Data" in (1). Sometimes, people aren't as explicit about the assumptions that underlie a particular definition of the available data. But, if someone declares some dataset is the relevant data for some learning problem, this person must have in mind (even if implicitly) a correct hypothesis the child is trying to identify from those data, a hypothesis space with competing hypotheses to choose from, a notion of what data children actually encounter, and ideas about which data children are able to use. This is why it's helpful to state these components explicitly – that way, we can at least identify where the disagreement is occurring, if and when it occurs. That is, is there disagreement because we don't think the child is trying to identify that particular hypothesis? Or because we don't think those are the correct set of competing hypotheses? Or because we think other data are actually available, and those other data are useable by children?

Okay, but if it *does* happen. If we agree about all the components that go into defining the data, we can then agree when poverty of the stimulus is occurring. Then, if we agree that children show constrained generalization despite that poverty of the stimulus, we can agree that children have some kind of prior knowledge or abilities.

It seems to me that this is where another round of tears happens – there's often significant disagreement about the nature of that prior knowledge or those prior abilities. Children must have the necessary knowledge or abilities available internally if they're not derivable from the external information currently available. So, one option is that children derived the necessary knowledge or acquired the necessary abilities from external information that was *previously* available. Because they completed that knowledge derivation or ability acquisition before the current time, the knowledge or abilities they gained are therefore prior knowledge or abilities. If that's what's happening, then we need an explanation for how children could derive that necessary knowledge or those necessary abilities beforehand from the data that was available beforehand. That is, this "derived prior

knowledge or abilities" is just a placeholder for another learning story that we need to fill in (see Lidz (2018) for this same idea, applied to a specific linguistic knowledge piece about interpreting pronouns like *herself*). That learning story may in fact rely on knowledge or abilities that were derived previously, so then we need a learning story about how *that* knowledge or *those* abilities were derived from the child's previous data. And so on.

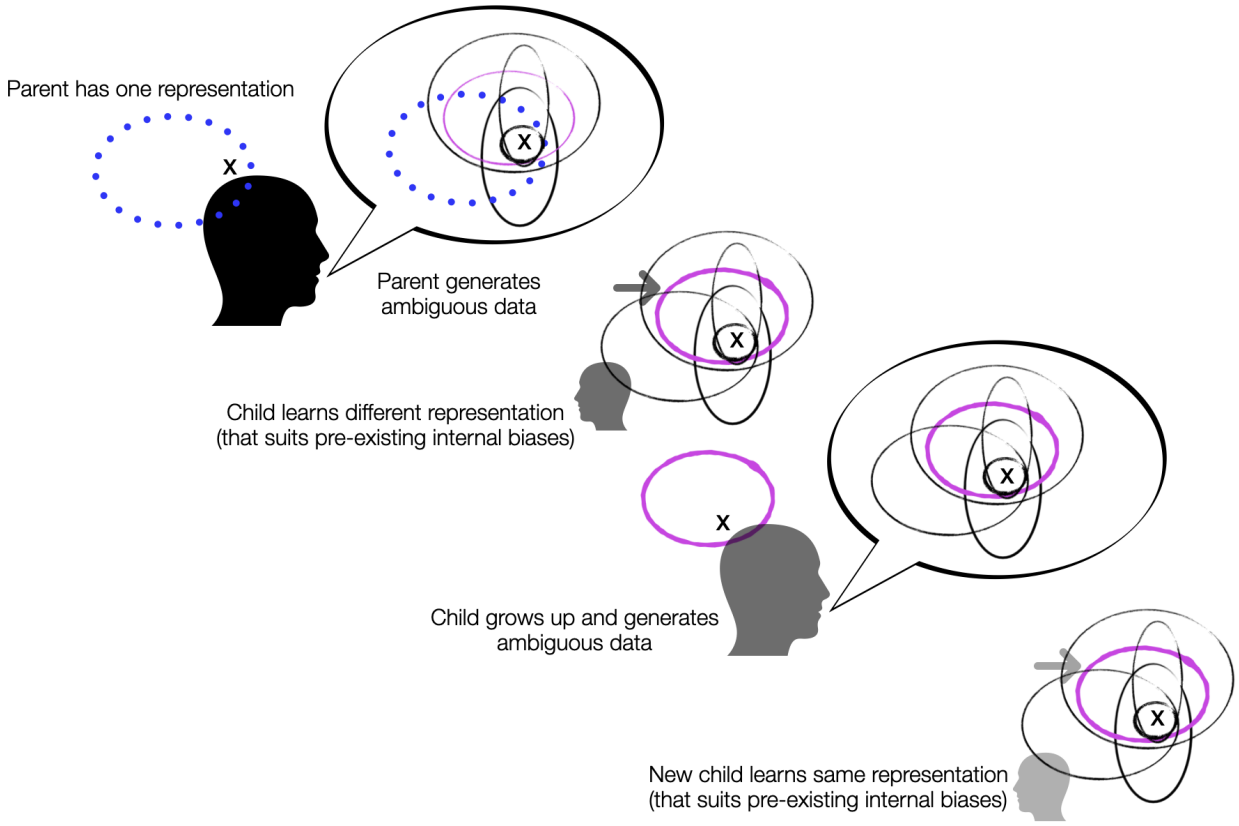
At some point, however, the learning story we come up with will have to involve something about knowledge or abilities that were already built into the child. That is, any learning story (especially a learning story about language) eventually comes back to some kind of innate knowledge or abilities the child has. Why? Because something internal about typically-developing children causes them to develop language when they're exposed to linguistic data, and something about kittens and rocks doesn't (Chomsky, 2000). This internal something is sometimes referred to as children's "biological endowment" (Lasnik & Lidz, 2017). The idea that typically-developing children have innate knowledge or abilities that allow them to learn has been called the *nativist* position.[7]

So, this line of argument is how the prior knowledge or abilities implied by poverty of the stimulus and constrained generalization leads to the assumption of innate knowledge or abilities. Many people have noted that even once we recognize the need for innate knowledge or abilities, nowhere have we specified the nature of that innate knowledge or those innate abilities (Stich, 1978; Laurence & Margolis, 2001; Scholz & Pullum, 2006; Chater et al., 2015; Skidelsky, 2016). A range of nativist positions are compatible with the necessary prior knowledge or abilities. Importantly, once we've decided that innate knowledge or abilities are required, we then have to explain how they got into human biology. This requires explanations from both developmental neurobiology and evolutionary biology. So, it's often a good idea to be very careful about what we want to claim regarding that innate knowledge or those innate abilities that enable children to make their constrained generalizations in the face of poverty of the stimulus. Once we decide on the nature of the innate something, we've just created explanatory work to do on the biological front.

Nativism isn't crazy from a language evolution perspective. Related to the explanatory work I noted above, I want to briefly mention some evolutionary modeling work by Kirby (2017) that takes on the explanatory challenge of innate knowledge or abilities for learning language. In particular, Kirby's work ties directly into poverty of the stimulus; this computational modeling work demonstrates why we could find ourselves with children who just happen to have the right innate knowledge or abilities to converge on the correct language answers in the face of the unresolvable ambiguity that poverty of the stimulus presents. I'll summarize the essence of those modeling results below, following along with the caricatures in Figure 5, where we consider the transmission of language over time from one generation to the next. The idea is that parents generate language data from their underlying representation, and children try to figure out that underlying representation from those data. Then, the children grow up and become the ones producing language data for the next generation of children. Language thus survives in a specific form because that form can be successfully transmitted from parents to children.

[7] More on this terminology in the next section, especially since it's been taken to mean different things by different people. That is, the use of the term "nativist" has itself caused some tears.

Figure 5: A visual demonstration of how poverty of the stimulus can cause children to have just the right innate biases for learning language from ambiguous data, given parents transmitting language to their children over time.



More specifically, at some initial stage, parents have an underlying representation (the blue dotted ellipse over the parent's head at the top of Figure 5) that they use to generate the observable data for their children (an example data point is the X inside the speech bubble at the top of Figure 5). These data are compatible with multiple hypotheses about the underlying representation (e.g., our classic poverty of the stimulus, as shown in the first adult speech bubble in Figure 5). The children still have to learn an underlying representation. However, the children can't do it on the basis of the external data, because we have that unresolvable ambiguity. So, what do they use? Something internal, of course – their prior (or innate) knowledge and abilities, shown with an arrow in the first child's mind in Figure 5. That is, the children converge on whatever answer *already* best suits their own internal biases (the thicker purple ellipse in Figure 5).

These children then grow up and are the ones producing data for the next generation of children. What representation are our grown-up children using to generate the data? The one their internal biases guided them towards, as shown by the thicker purple ellipse in the grown-up child (which is the second adult) in Figure 5. So, here they are, generating the same ambiguous data they themselves learned from (the X in Figure 5), but from an underlying representation they're naturally inclined towards. Now, which representation will their children pick out when given

those same ambiguous data? The one their children's internal biases steer the children towards, shown with an arrow in the second child's mind in Figure 5, which leads children to that underlying representation they're naturally inclined towards (the thicker purple ellipse in Figure 5).

All this is to say that, over time, language may well have been shaped to capitalize on the internal biases children have (i.e., the innate knowledge and abilities they have), and it may have done this precisely because the data were ambiguous. Put simply, when the data are ambiguous, children have the freedom to pick a representation suited to their internal biases; this in turn allows these children, once grown up, to continue generating ambiguous data for their children to learn from. Importantly, *those* children will converge on the same answer, which aligns with their internal biases, despite the unresolvable ambiguity in the data. There's no need to make the data unambiguous – children's innate biases will guide them safely to the right answer anyway, just as it guided their parents. So, poverty of the stimulus causes children to rely on their innate biases to shape the language; future generations can successfully learn that same-shaped language knowledge from ambiguous data by relying on those innate biases. This then is how we end up with typically-developing children who have remarkably helpful innate biases for learning language as it's shaped today.

1.4 So what *is* that special prior something?

As I mentioned above, once researchers agree that poverty of the stimulus and constrained generalization are happening, they generally can agree that at some point in development, innate knowledge or abilities are required (Pullum & Scholz, 2002; Clark & Lappin, 2011; Rawski & Heinz, 2019). After that, opinions often sharply diverge (with tears). I'll describe the two most prominent viewpoints I'm aware of.

Linguistic nativism and Universal Grammar. One form of nativism is *linguistic nativism*, a term used (among others) to describe the view that the innate knowledge or abilities are language-specific (Pullum & Scholz, 2002; Scholz & Pullum, 2002, 2006; Clark & Lappin, 2011). That is, the innate knowledge or abilities that children are equipped with is something that's only useful for learning language, and not for other aspects of cognition – some proponents of this view include Chomsky (1971), Hornstein and Lightfoot (1981), Baker and McCarthy (1981), Crain and Pietroski (2002), and Valian (2009). Innate, language-specific knowledge or abilities often goes by the name *Universal Grammar* (e.g., Pearl and Sprouse (2013b) and Skidelsky (2016) use this term for it). However, sometimes Universal Grammar is used to describe plain nativism (Clark & Lappin, 2011) – that is, whatever innate knowledge or abilities (language-specific or not) are needed for children to learn language successfully from ambiguous data. So, if someone says they believe in Universal Grammar because poverty of the stimulus exists and children show constrained generalization, it's useful to determine what kind of Universal Grammar that person is supporting.

More generally, proponents of linguistic nativism traditionally have been interested in showing that poverty of the stimulus and constrained generalization both occur; they then draw the conclusion that the prior knowledge or abilities needed ultimately come from innate language-specific knowledge or abilities. However, this causes tears, because it's not the only conclusion to draw.

Non-linguistic nativism and empiricism. In particular, other people see both poverty of the stimulus and constrained generalization occurring, and conclude the required innate knowledge or abilities aren't specific to language (e.g., see Clark & Lappin, 2011; Chater et al., 2015; Piantadosi & Kidd, 2016).[8] We might reasonably think of this viewpoint as *non-linguistic nativism*, because the innate knowledge or abilities required are claimed not to be language-specific So, in this way, it contrasts with linguistic nativism, though both viewpoints are perfectly fine with innate knowledge or abilities being required (i.e., plain nativism).

However, sometimes linguistic nativism is described simply as "nativism" – this leads to yet more tears, as people then conflate nativism (innate knowledge or abilities required) with linguistic nativism (innate, linguistic knowledge or abilities required). This conflation causes non-linguistic nativists to prefer terms that contrast with nativism: *non-nativism* or *empiricism* (Chater et al., 2015; Piantadosi & Kidd, 2016).[9] But, non-nativism and empiricism might sound odd when describing child language development, because a non-nativist/empiricist child wouldn't have any innate knowledge or abilities; this seems strange when we think about children learning language, where everyone's pretty happy with children having innate knowledge or abilities of some kind (Quine, 1969; Pinker, 2004; Scholz & Pullum, 2006; Heinz & Rawski, to appear). As I said though, what's happened is that the non-linguistic nativists have co-opted these terms to make clear their viewpoint is distinct from the viewpoint of the linguistic nativists. That is, those who call themselves non-nativists or empiricists are in fact nativists, as I've defined here.

For example, Clark and Lappin (2011) note that "No empiricist that we are aware of denies that humans have innate cognitive abilities through which they acquire language". In fact, Quine (1969) noted over 50 years ago that a behaviorist (the contrasting position to linguistic nativist at that time) "is knowingly and cheerfully up to his neck in innate mechanisms". The non-linguistic nativists have gravitated recently towards the term "empiricist" in particular to highlight their belief that children harness the data much more effectively than linguistic nativists have traditionally thought (Chater et al., 2015). That is, today's empiricists believe there's a larger role for the available empirical data, even though they still believe in innate knowledge or abilities.

Linguistic vs. non-linguistic nativists. So, it seems that everyone who agrees that poverty of the stimulus and constrained generalization occur then believes in some kind of innate knowledge or abilities (that is, we're all nativists). People simply split on the nature of that innate knowledge or those innate abilities. Linguistic nativists believe that at some point sometime in children's language development, there's required innate knowledge or abilities that are specific to learning language; non-linguistic nativists don't believe this. Instead, non-linguistic nativists believe all required prior knowledge and abilities are derivable from the child's experience coupled with non-linguistic innate knowledge and abilities.

The reason this split causes so many tears is that it's currently quite hard to prove which perspective is right, even for a specific poverty of the stimulus and constrained generalization pairing. The good news for linguistic nativists is that it only takes one piece of innate linguistic knowledge or one innate linguistic ability for them to be right. That is, linguistic nativism doesn't specify how

[8] See sections 2.4 and 3.2 for concrete examples of this kind of proposed innate knowledge.

[9] Special thanks to Vic Ferreira for first pointing this out to me.

much innate linguistic knowledge or how many innate linguistic abilities are needed – just that *some* are. Of course, it may be that many people who identify as linguistic nativists think quite a lot of the required innate knowledge or abilities are linguistic – but technically, a linguistic nativist is someone who believes at least one piece of required innate knowledge or ability is linguistic. This contrasts with the non-linguistic nativists, who believe that absolutely no innate linguistic knowledge or ability is required, ever. So, for them to be right, they have to demonstrate that for every poverty of the stimulus and constrained generalization pairing, children *never* need innate linguistic knowledge or abilities. That's a tall order.

However, the good news for all the nativists is that we can make progress on this debate about the nature of the built-in knowledge or abilities by showing concretely which prior knowledge and abilities will allow children to show constrained generalization for specific poverty of the stimulus instances. Computational modeling is a particularly effective way to do this (Clark & Lappin, 2011; Chater et al., 2015; Pearl, in press-b), and this approach is discussed more in the next section, with some examples reviewed in section 3. Once we know which prior knowledge and ability options will allow children to show constrained generalization, we can argue about which options, linguistic or non-linguistic, we think are best among the options that do in fact work.

2 How to tell if we have poverty of the stimulus

A first step is to identify if poverty of the stimulus is occurring for a particular piece of language knowledge. This means we need to be clear about what the correct hypothesis is, what the hypothesis space is, what the available data are, and how children leverage those data. If all of those align to show unresolvable ambiguity in the child's input, we have poverty of the stimulus. Below I'll review two general approaches for doing exactly this: demonstrations of poverty of the stimulus in principle and demonstrations in practice, discussing how each can leverage specific case studies and why doing so is helpful for understanding when poverty of the stimulus is occurring. I'll then connect this to children's constrained generalization, where the idea is that the ambiguity is unresolvable given the time constraints children have to resolve it and the quantity of data they encounter. This means we need to determine how much data *would* be enough to resolve that ambiguity as quickly as children do, and I'll review some current approaches for doing that.

2.1 Show it in principle

Formal learnability approaches (interpret with caution). One way to show the data are insufficient is to use formal learnability approaches to show it in principle (Gold, 1967; Horning, 1969; Wharton, 1974; Angluin, 1980; Valiant, 1984; Angluin, 1988; Jain, Osherson, Royer, Sharma, et al., 1999; Niyogi, 2006; Clark & Eyraud, 2007; Heinz, 2016). The reason these approaches have the caveat "in principle" is because they attempt to idealize away from aspects of the learning problem in order to be able to draw very general conclusions that should hold, no matter what the specifics are of the particular learning problem. So, when formal learnability approaches yield a learning result suggesting the data are insufficient, it holds in principle, given the assumptions that were made.

This is often where the trouble comes in. Formal learnability results are hard to interpret out of context – see Johnson (2004), Clark and Lappin (2011), and Heinz (2016) for how results from Gold (1967) have notoriously been misinterpreted (a lot). In particular, Gold's results were used to assume that poverty of the stimulus exists for "language" as a whole, where language is something that generates sequences of units – so, we can think about this as sequences of sounds or morphemes or words or phrases or sentences, or whatever else. That is, because linguistic knowledge is implemented as allowed vs. unallowed sequences of units, Gold's results have the potential to inform us about many different kinds of linguistic knowledge, as long as the linguistic knowledge can be realized as a sequence of units. On the flip side, this means that these results are disconnected from any particular piece of linguistic knowledge (e.g., the case studies discussed in section 3). This is especially true if the linguistic knowledge isn't about sequences of units (e.g., the lexical semantic knowledge discussed in section 3.2). Still, to be able to have such a general result about any kind of language knowledge involving sequences of units was very exciting, as many types of linguistic knowledge can be realized as allowed vs. unallowed sequences. For example, the *wh*-word position example from before can be thought of as sequences of words that are allowed (i.e., word sequences with the *wh*-word fronted for English's default) vs. unallowed (i.e., word sequences with the *wh*-word in-situ for English's default).

Importantly, however, Gold's result was based on assumptions about the nature of the learning process that are unlikely to be true for children. So, drawing parallels between what happens for those learners and what happens for children isn't recommended. That is, asserting that poverty of the stimulus holds for children learning language on the basis of these results isn't safe to do. Why not? In essence, the idealizing assumptions idealized away too much from the learning situation children actually face. I encourage interested readers to see Heinz (2016) for a recent detailed discussion about these assumptions. But, the main problematic assumption was that children need to succeed for all possible presentations of the input – whether these are realistic input samples or strategically difficult ones concocted to purposefully lead a learner astray. In real life, children only need to succeed on the particular presentations of the input they actually encounter, i.e., the realistic ones. So, because of these idealizing assumptions, the results weren't applicable to people interested in the learning problems children face. (Again, see Johnson (2004), Clark and Lappin (2011), and Heinz (2016) for more detailed discussion about why these learning results tend not to bear on poverty of the stimulus for children. In this particular case, the idealizations made the learning problem far harder than the one children actually face.)

However, other formal learnability results relied on assumptions that made the learning problem being investigated more similar to the learning problem children face (see Clark and Lappin (2011) for more detailed discussion of how those learning problems were more similar). For instance, Clark and Lappin (2011) discuss learnability results where the learning scenario is much more similar to a child's learning scenario in several key ways.

First, the learner has to converge on target knowledge that's "close enough" to the true target in some measurable way (rather than being identical). We might think of this as children generating adult-like behavior in an experimental task. When we see children behaving in adult-like ways, we typically interpret them as having adult-like knowledge. In point of fact, these children might not have identical knowledge to that of adults – but the knowledge they've acquired is "close enough"

to lead to adult-like observable behavior.

Second, the input examples aren't labeled as to whether they're positive examples of what's grammatical in the language or instead noise, and there is in fact noise in the input. We might think of this as a child being around a non-native speaker who says something ungrammatical in the language, or a native speaker who has a slip of the tongue and says something ungrammatical. These ungrammatical items aren't explicitly tagged in children's input (e.g., "Whoops, just kidding – I can't say that form!"). Instead, the child has to figure out (perhaps using other cues and some sophisticated inference) that these items aren't grammatical items in the language, but instead noise that should be ignored when trying to determine the correct generalizations for the language.

Third, the learner has to learn in a "reasonable" amount of time instead of getting an infinite amount of time to learn. We might think of this as learning occurring within a defined learning period. For instance, this learning period could correspond to the age when children seem to acquire adult-like knowledge of a particular linguistic representation.

When all these restrictions on the learning scenario are in place, Clark and Lappin (2011) note that the learners can succeed as long as they leverage indirect negative evidence in a particular way: the learners must interpret an item with relatively low probability as ungrammatical. That is, direct evidence that an item occurs in the language very rarely must be interpreted as indirect evidence that the item in fact doesn't really occur in the language (it's noise). With respect to the *wh*-word position example, if the learner rarely hears a *wh*-word in situ, then this learner will assume that the default rule for English doesn't involve allowing the *wh*-word to be in situ. So, these learners must be able to leverage the data in a certain way in order to succeed – i.e., they need a specific type of learning process in order to show constrained generalization in the face of poverty of the stimulus.

Clark and Lappin (2011) demonstrate both that (i) poverty of the stimulus exists in the hypothesis space they define, given the hypothesis the learners must identify and the available data, and (ii) what learners need to do (and thus know how to do) to solve this poverty of the stimulus. Here, the necessary knowledge is to leverage indirect negative evidence (and have the innate ability to do so). To the extent that the assumptions that Clark and Lappin (2011) made are translatable to children's learning, we can translate the insight about relying on indirect negative evidence to children as well. More specifically, for children to succeed in a learning scenario where they need to identify a close-enough generalization from noisy input in a limited amount of time, these learnability results suggest children must rely on indirect negative evidence in this particular way: they need to believe that sufficiently low probability indicates ungrammaticality.

A reasonable person might wonder why someone would use formal learnability approaches, given how difficult it is to idealize away details of children's learning scenarios without also losing relevance to those learning scenarios. Clark and Lappin (2011), referencing Pinker (1979), offer this explanation: it can be useful to adopt idealizations that are demonstrably false when those idealizations make the learning problem harder than the one children face. If the harder learning problem can be solved, then the easier real problem that children face should also be solvable. However, just because we can't solve the harder problem doesn't mean children can't solve the easier problem – this was a key misinterpretation of the Gold (1967) results, where the idealized learning problem was much harder. This potential for misinterpretation underscores how careful

we have to be when interpreting formal learnability results, as they apply to children's potential poverty of the stimulus.

Ideal learner approaches that use realistic hypothesis spaces and available data. A related "in principle" approach is to focus on a set of learning scenarios that map more directly to children's learning scenarios. For example, Chater and colleagues (Hsu & Chater, 2010; Hsu et al., 2011, 2013; Chater et al., 2015) have developed a "rigorous formal framework for understanding what can, in principle, be learned from positive linguistic evidence alone" (ch.5, p.150 of Chater et al., 2015). This contrasts with additionally learning from indirect negative evidence, as Clark and Lappin's (2011) ideal learner did. That is, this learning approach only pays attention to examples that are in the input, and doesn't try to make inferences based on examples that either aren't in the input or aren't in the input often enough.

This approach also assumes an *ideal learner* who can extract the maximal possible information from the available input; the available input is restricted to the input we believe is available to children (this is the positive linguistic evidence mentioned by Chater and colleagues). For this reason, the target hypothesis and hypothesis space are also more directly tied to specific linguistic phenomena where we know what these look like for children. That is, this ideal learner approach focuses on specific case studies where we know information about the data available to children, and have concrete proposals for both the correct hypothesis and the hypothesis space that children have to navigate. So, this ideal learner approach defines the correct hypothesis, the hypothesis space, and the available data to be what they are for children, and then idealizes how children leverage those data.

Idealizing how children leverage the available data allows us to define an upper limit on what could be learned from the available data – that is, what's possible in principle for children to learn from the available data when they don't have limitations on relevant cognitive abilities (like memory or parsing) that allow them to harness all the information in the data. So, if poverty of the stimulus exists even for an ideal learner able to extract as much information as possible from the data, it likely exists for children too (remember: this is because children aren't ideal learners – they'll have an even harder time leveraging the data). This is a related line of logic to the formal learnability approaches described above: there, if a learner could solve a harder learning problem than the one children face, we might assume that children could solve their easier problem; here, if we find that an ideal learner *can't* solve the problem when it has the best possible information extraction abilities (i.e., making it an easier learning problem than what children face), then we might assume that children can't solve their harder problem either. That is, if an ideal learner with the optimal ability to leverage the information in children's input fails, then it's likely that children, who have sub-optimal abilities to leverage the information in their input, would fail too (without some prior knowledge or abilities helping them along).

I should note that this is the general approach of *computational-level* models of development, in the sense of Marr's (1982) framework. Marr's framework is often used to specify different types of explanations for problems in cognition, and includes characterizing the mental "computation" that's being computed (computational-level), the sequence of steps (or algorithm) used to accomplish that mental computation (algorithmic-level), and how that computation is implemented phys-

ically (implementational-level). (See Pearl (in press-b) for more detailed discussion about how this framework can be applied to models of language development.) For our purposes here, the important idea is that a computational-level explanation involves characterizing the mental computation we think children are accomplishing for a particular acquisition task, and seeing if that's a mental computation that leads to success. More specifically, the mental computation we characterize in an ideal learner model is the task of identifying a particular linguistic generalization from a particular linguistic hypothesis space, given realistic child-directed input and using whatever prior knowledge or abilities are built into the ideal learner. If that ideal learner is successful, then that mental computation, which involves relying on specific prior knowledge or abilities built into the ideal learner, could be a good description of the mental computation that children are doing.

So, this approach investigates what an ideal learner can do, when given the same learning computation to accomplish as children (i.e., the same correct hypothesis, the same hypothesis space, the same available data, and a specific combination of prior knowledge and abilities). Then, if we find something is possible to learn in principle with an ideal learner relying on specific prior knowledge and abilities, we can see if it's possible to learn in practice with child-like learners who have constraints on how they leverage their data. (Maybe it's not, and we still have poverty of the stimulus after all.)

Here, I'd like to briefly note the difference between constraints and biases in learning models, as I think this is another point where confusion (and tears) can arise. In general, I see these terms used to describe different aspects of a modeled learner. A constraint typically separates an ideal learner from a less-ideal (constrained) learner, with the ideal learner having unlimited abilities to accomplish inference and the constrained learner having limited abilities. For instance, an ideal learner may have unlimited memory and adult-like parsing abilities, while a constrained learner may have limited memory and immature parsing abilities. In contrast, a bias is simply some prior knowledge that the learner has (e.g., prior knowledge about hypotheses in the hypothesis space, prior knowledge about how to leverage the available data, prior knowledge about what kind of inference to use on the data, and so on). Notably, a constrained learner's limitations can give rise to biases (e.g., limited memory and immature parsing leading to a bias to only extract information from the first clause uttered). However, a bias that can be made explicit (e.g., only extract information from the first clause uttered) could just as easily be implemented in an ideal learner that doesn't have limited abilities – for instance, an ideal learner with unlimited memory and parsing abilities could have prior knowledge built in (i.e., a bias) to only extract information from the first clause.[10] More generally, an ideal learner can incorporate explicit biases (and so be a biased learner) while a constrained learner can be unbiased in various ways (e.g., intending to heed all information available in the input – even if limited memory and parsing prevent this from occurring in practice). This is why "constraint" is different from "bias" in the learning model literature.

Returning to our discussion of how to interpret ideal learner results, suppose we find something isn't possible to learn in principle with an ideal learner that has specific prior knowledge

[10]Interestingly, it's not always true that constraints lead to biases that we can easily describe and therefore implement in an ideal learner; in Pearl and Phillips (2018), a constraint on memory led to better acquisition performance, but it wasn't from anything easy to explicitly describe.

and abilities. Then, we might plausibly assume that children, who have far more constraints on how they leverage those data, also couldn't learn the correct hypothesis from those data even if they had that same specific prior knowledge and those same specific abilities. For instance, see Pearl (2011) for a case study in learning English metrical stress (i.e., which syllables are stressed in words such as *elephant*); there, modeled learners were implemented that had some of the constraints children have when learning. To my surprise, I discovered that every single one of those constrained learners failed to learn the correct hypothesis from plausible English data. I then analyzed the acquisition problem from an ideal learner perspective and discovered that an ideal learner would fail too. The issue was this: learning that hypothesis from that hypothesis space as the best fit for those available data wasn't possible because that hypothesis *wasn't* the best fit for those data, given the other hypotheses available, without some additional bias that would privilege the correct hypothesis. This conception of the acquisition problem (i.e., my assumption that the mental computation children had to solve was learning that hypothesis in that hypothesis space from those data with that specific combination of prior knowledge and abilities) meant the learner – idealized or constrained – was doomed to failure. What was needed was a way of changing the acquisition problem the learner faced by changing one of those pieces.

One promising piece to change is the data children are learning from, by using some sort of data filter. More specifically, the idea is that children don't in fact learn from everything available. Instead, they have helpful biases that filter the data down to a helpfully-skewed subset; that subset makes learning the right hypothesis easier. For the English metrical stress case above, Pearl, Lu, and Haghighi (2017) and Pearl (2017) discuss how a certain kind of data filter, which keys into how reliable the data seem to be, can yield a subset of available data that are far more amenable for learning the correct hypothesis.

More generally, the idea of children using a subset of the available data for whatever reason is in line with the "less is more" hypothesis (Newport, 1990) and other data filtering approaches (e.g., Pinker, 1984; Lidz & Gleitman, 2004; Pearl, 2007; Pearl & Weinberg, 2007; Pearl & Lidz, 2009; Adriaans & Swingley, 2012; Perkins, Feldman, & Lidz, 2017). In particular, children would be using "less" data to achieve "more" language learning success. Another way to view this is "quality over quantity": having a helpfully-skewed subset of data to learn from yields higher quality data, even though the overall quantity of data being learned from is smaller.

So, let's return to our discussion of ideal learner approaches. Suppose we have the right conception of the acquisition problem (i.e., the right description of the mental computation the child is trying to accomplish): a learner with specific prior knowledge and abilities is trying to learn a particular hypothesis from a particular hypothesis space, given a particular – perhaps filtered – set of data. Then, the failure of a learner capable of idealized inference should mean actual children would fail too.

But, the key is getting the right conception of the acquisition problem. Very often, investigations necessarily simplify the acquisition problem to make it tractable, and this can involve simplifying the data in the input. That can then mean that constrained learners, whose constraints make them less able to leverage the information available in the data, may do better than idealized learners, whose lack of constraints make them perfectly able to leverage information available in the data. (See Pearl and Phillips (2018) for an example of this in speech segmentation.)

So, if a constrained learner is succeeding more than an ideal learner, the reason is that the acquisition problem that the ideal learner is solving isn't the right one. That is, if constrained learners are doing better, something about their constrained inference is causing them to change one of the key pieces of the problem (either the data considered relevant and how relevant to consider it, the other hypotheses available for consideration, or the specific combination of prior knowledge and abilities that's built in). If these changes could be made explicit, an ideal learner working with those data in that hypothesis space should also then be able to come up with the correct hypothesis.

However, suppose we do manage to get exactly the right conception of the acquisition problem that children face (hypothesis, hypothesis space, relevant data, and specific combination of prior knowledge and abilities), and we find that learners capable of ideal inference fail. We can then assume poverty of the stimulus exists for children learning that hypothesis, given that hypothesis space, those available data, that specific combination of prior knowledge and abilities, and children's constrained ability to leverage the available data. The work to do at that point is to identify a different combination of prior knowledge and abilities that would allow the learner to succeed, whether it's an ideal learner or a constrained learner – and so offer a concrete solution to that instance of poverty of the stimulus.

2.2 Show it in practice for particular case studies

A very related idea is to demonstrate that poverty of the stimulus occurs in practice for specific case studies where a constrained learner – rather than an ideal learner – is trying to identify a specific correct hypothesis within a specific hypothesis space, given specific available data and a specific combination of prior knowledge and abilities. The main difference for "in practice" investigations is using use a modeled learner that's more child-like in its constraints on leveraging the available data (e.g., Pearl & Sprouse, 2013a; Pearl & Mis, 2016). Such child-inspired constraints are typically tied to current knowledge about how children of various ages are known to leverage available data (e.g., statistical learning abilities at different ages: Saffran, Aslin, & Newport, 1996; Xu & Tenenbaum, 2007; Smith & Yu, 2008; Perfors, Tenenbaum, Griffiths, & Xu, 2011; Aslin, 2017).

2.3 The usefulness of case studies for identifying poverty of the stimulus

As noted above, both "in principle" idealized learning approaches and "in practice" constrained learning approaches for identifying poverty of the stimulus can focus on specific case studies in language development. To me, this is valuable because researchers can be explicit about what they think the available data are that children can leverage; disagreements about the available data are a notorious source of tears about poverty of the stimulus. It may well be that poverty of the stimulus exists when children can only leverage certain available data, and doesn't exist if children leverage other available data. One example of this is if researchers believe children only leverage direct positive evidence for a particular piece of linguistic knowledge, as opposed to also leveraging indirect evidence, which I discussed above in section 1.2. Another example would be if researchers believe children only leverage positive evidence of the same kind as the target knowledge, such as only relying on syntactic information to learn about syntactic knowledge; this

contrasts with children relying on semantic, pragmatic, prosodic, or extralinguistic information for learning about syntactic knowledge (Clark & Lappin, 2011).

More generally, a lot of responses to claimed poverty of the stimulus scenarios take this form: what previous researchers assumed was the entire set of available data children can leverage in fact isn't. Instead, the data are richer than assumed (particularly if the child is learning a system, rather than just an isolated piece of knowledge: Perfors, Tenenbaum, & Regier, 2011; Pearl & Mis, 2016; Abend et al., 2017; Pearl, in press-b). So, perhaps poverty of the stimulus isn't occurring. For me, this is a real benefit of focusing on particular case studies: I find it easier to engage with available empirical data on linguistic representation and child development when there's a specific piece of knowledge we're looking at as a possible poverty of the stimulus scenario. Then, we can examine linguistic theory to tell us about (i) the correct hypothesis, and (ii) the hypothesis space children are navigating; we can also examine developmental data to tell us about (iii) realistic input data for children trying to learn that piece of knowledge, and (iv) information that children of the appropriate age are able to leverage from their data. With this in hand, we can more explicitly characterize the acquisition problem children are facing and determine if poverty of the stimulus seems to be occurring.

2.4 Specifying how much data is enough

One thorny issue for identifying poverty of the stimulus scenarios, especially for particular case studies, is that povery of the stimulus is intricately tied to children's constrained generalization. More specifically, it's often tied to how quickly children make these constrained generalizations, given the available data (Jackendoff, 1994; Laurence & Margolis, 2001; Crain & Pietroski, 2002). Recall from the discussion in section 1.2 that children's rapid learning has been used as a signal that poverty of the stimulus may be occurring because it seems that children are "computing" something complex very fast (Clark & Lappin, 2011) – in fact, faster than we would expect, given the available data. So, in practice, it seems the actual data aren't often unresolvably ambiguous; rather, the ambiguity is unresolvable, given those data, as *fast* as children resolve it. More specifically, there isn't "logical" poverty of the stimulus where the data are forever ambiguous; rather there seems to be practical poverty of the stimulus where the data are ambiguous enough to slow down learning, and yet children aren't slowed down as much as expected. That is, children are resolving the ambiguity faster than expected.

How can we tell that children are resolving the ambiguity faster than expected? What's expected? It seems that the expected time to resolve the data ambiguity must be tied to how many informative data are present. If this amount is less than what children would need to resolve the ambiguity as fast as they do, then we have poverty of the stimulus after all. That is, informative data may exist, but there simply aren't enough of them for children not to also need some prior knowledge or abilities. So, it seems a crucial component for identifying poverty of the stimulus is identifying how much data is enough.

Developmental benchmarking with unambiguous data. In one lively debate, Pullum and Scholz (2002) noted that no one was saying how much data was enough, even though many people were quite convinced poverty of the stimulus was occurring for various case studies. Legate and Yang

(2002) took up that challenge and offered one approach to specifying how much data was enough for children to make the constrained generalizations they did. Legate and Yang's approach relied on children primarily leveraging only direct positive evidence that was unambiguous (so, something like the data in Figure 2 for hypothesis C).[11] Note that to determine data are unambiguous, we need to have a clear idea of both the correct hypothesis and the hypothesis space children are navigating, such that those data are in fact unambiguous for the correct hypothesis.

However, once we know what the unambiguous data would be, Legate and Yang (2002) suggested that the quantity of unambiguous data in children's input for any piece of linguistic knowledge should correlate with the age children figure out that piece of linguistic knowledge (also known as children's *age of acquisition*). In particular, ages of acquisition that were correlated with certain amounts of unambiguous data could serve as developmental benchmarks. The reasoning was that, all else being equal,[12] linguistic knowledge with less unambiguous data available should be learned later while linguistic knowledge with more unambiguous data available should be learned earlier. If we find that something with less unambiguous data available (knowledge K_{less}) is nonetheless learned earlier – as early, say, as something with more unambiguous data available (knowledge K_{more}) – this could indicate poverty of the stimulus for knowledge K_{less}.

Legate and Yang (2002) then proceeded to explore what was then known about (i) different linguistic knowledge pieces in terms of age of acquisition, (ii) what could plausibly be the unambiguous data for a particular knowledge piece, and (iii) how frequently those unambiguous data appear in children's input. They found support for their suggested developmental correlation in two different linguistic knowledge pieces that seemed to have the same amount of unambiguous data available in children's input (about 1.2% of children's input) and also have the same age of acquisition (around 3 years old). Subsequent investigations (Yang, 2004, 2012) found many more examples of knowledge pieces that aligned with this idea, with unambiguous input frequencies ranging from 0.2%-25% of children's input and ages of acquisition from older than four down to younger than two. These examples then serve as developmental benchmarks for how fast we should expect children to show constrained generalizations. More specifically, if we assume unambiguous data drive children's acquisition and we can both define what those unambiguous data are and how frequently children encounter them in the input, we can determine if children are constraining their generalizations faster than expected.

Information-theoretic approaches. Another approach, which was more recently harnessed by Hsu, Chater, and colleagues (Hsu & Chater, 2010; Hsu et al., 2011, 2013; Chater et al., 2015), draws on information theory to specify how much data would be required for children to constrain their linguistic generalizations the way they do. The key idea for this approach is known as *Minimum Description Length* (**MDL**) (Li & Vitányi, 1994; Rissanen & Ristad, 1994), and focuses

[11] I say "primarily" because the variational learner of Legate and Yang (2002) actually learns from ambiguous data as well, but is really driven by unambiguous data. This is because over time, ambiguous data cancel out, leaving the unambiguous data to drive the learner's generalizations. See Pearl (in press-b) for more discussion about why this is.

[12] It's important to note that this "all else being equal" assumption does a lot of work. As just one example, we would need to assume that both knowledge pieces were equally complex from the perspective of the child. If not, it might be that the simpler one was learned more quickly than the more complex one, even with the same amount of unambiguous data.

on how compactly a particular representation would allow the child to represent the data she encounters. That is, an MDL-based learner is choosing representations based on storage capacity: representations that enable more compact storage are viewed as better. The idea that more compact representations should be favored has actually been around for quite some time – for instance, Chomsky (1957) argues for one representation over another because the first representation is more compact than the second one. In the context of formal learnability, Stabler (1998) uses an MDL approach to learn about certain types of languages that have properties similar to human languages (e.g., they allow words to move). This preference for compactness is in fact a type of prior knowledge – if two representations are each compatible with the available data (i.e., we have poverty of the stimulus), an MDL learner prefers the one that takes up less space.[13] Note that this prior knowledge is domain-general, since it could as easily apply to non-linguistic representations as it could to linguistic representations. So, an MDL learner relies on non-linguistic innate knowledge, and so is compatible with the non-linguistic nativist viewpoint. I should also note that the MDL approach is quite compatible with linguistic innate knowledge (and so with the linguistic nativist viewpoint) – for instance, the representations being compared using MDL might well be generated using linguistic innate knowledge. But, the fact that the MDL learner prefers a more compact representation is itself non-linguistic innate knowledge.

More specifically, as shown in the example in Figure 6, an MDL learner encodes both the representation itself and the available data on the basis of that representation. So, a more complex representation may take up more space, but allow more compact encoding of each data point; in contrast, a simpler representation may take up less space, but require more space for each data point. After enough data are encountered, this additional cost per data point can lead to the simpler representation being less compact than the more complex representation. This is the key insight of the MDL approach to diagnosing poverty of the stimulus: we can identify when the representation children select should overtake other competing options, on the basis of the data children encounter. If this occurs later than children actually seem to identify the correct representation, we have evidence that poverty of the stimulus is happening and *additional* prior knowledge (besides a preference for compactness) is required.

For example, let's look more closely at the learning scenario in Figure 6, where the child considers a hypothesis space consisting only of representations C and E. C is the correct representation (i.e., the one we observe children converging on by a certain age), and is more complex, as evidenced by its more complex shape – this complexity also shows up as C taking more space to encode in the top box, compared with the simpler representation E. Now, suppose we believe the available data are the Xs in Figure 6; they're clearly compatible with both C and E, and so we have classic poverty of the stimulus to begin with. However, remember that an MDL learner has prior knowledge that causes it to prefer more compact representations. So, initially, this learner will prefer E – E takes up less space, after all.

Then, some time passes (time 1) and the learner has encountered some quantity of those available data. When trying to encode those data with each representation, the overall storage required

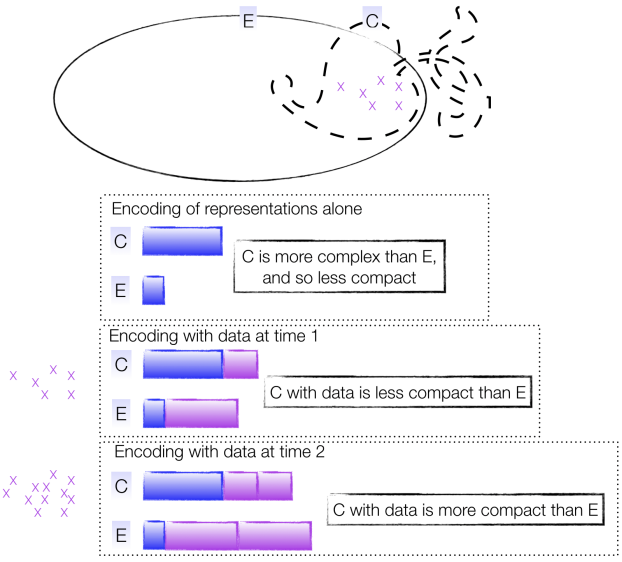
[13] I note that I'm using the term "knowledge" more broadly than it's traditionally been used. More specifically, I'm referring to any kind of bias or preference as knowledge (such as the preference for more compact representations). This contrasts with only using "knowledge" to refer to the representations in a child's hypothesis space.

is shown in the middle box of Figure 6.

E is still more compact overall than C, but C is definitely catching up. Still, an MDL learner prefers E at this point. However, by time 2, when the learner has encountered twice as much data, the encoding with C is more compact than the encoding with E. Therefore, an MDL learner now prefers C.

This demonstrates how a "break even" point can be calculated for an MDL learner – in Figure 6 terms, how many Xs have to be encountered before the encoding with C is more compact than the encoding with E? Time 1 was too soon, while time 2 was just a bit too late. Using this approach, we can figure out how many data are required before the encoding with C is the same size as the encoding with E. Just after that should be when an ideal MDL learner would switch from preferring E to preferring C. Let's call this amount of data $D_{breakeven}$.

Figure 6: An MDL encoding for hypotheses C and E, which are both compatible with the available data (shown as Xs). Over time, the more complex hypothesis C yields a more compact representation than the simpler hypothesis E.



Now, how do we determine how much time corresponds to $D_{breakeven}$? This is where we turn to developmental data about how often children hear these data in their input, as well as how much input children hear over time. We can calculate how often children hear these data by assessing realistic samples of child interactions: in those interactions, how often do these data show up? This corresponds to the input rate. Then, we just need to know how much total input children encounter by certain ages. For this input quantity estimate, we can turn to data like those of Hart and Risley (1995, 2003), which will tell us how many utterances children hear on average per hour (e.g., 487 in a professional household for children ages 13-36 months); we can then turn to data like those of Davis, Parker, and Montgomery (2004) for hours awake per day at each age (e.g., 11-12.5 hours between the ages of 2 and 5). Multiply utterances per hour by hours per day, and we have utterances per day at difference ages; multiply utterances per day by the input rate of our Xs, and we know how many Xs a child of a certain age hears per day. Then, we can multiply by days per year to estimate how many of these data a child of a certain age hears in a year. That is, with these input quantity data in hand, we can calculate how many of our Xs from Figure 6 children of different ages would have encountered. That's how we translate from a quantity of data to an estimated age of acquisition. More specifically, we can translate the quantity of data in $D_{breakeven}$ into an estimated age of acquisition, once we know the input rate of the Xs and how much input children hear over time.[14]

[14]Of course, this estimate of the input rate is also an idealization, in the sense that we only have samples of every component that goes into that estimated input rate; however, in contrast to formal learnability work, this view of the input derives from empirical estimates of children's input.

If the age of acquisition indicated by $D_{breakeven}$ is later than children's observed age of acquisition, then we've identified poverty of the stimulus. (Remember, of course, that this is poverty of the stimulus for a child using the MDL learner's preference for compactness. We don't know if it would be poverty of the stimulus for a child using different prior knowledge.) But suppose we don't know the exact age of acquisition for children – can we still use this MDL approach? The answer is yes, if it turns out that $D_{breakeven}$ translates to a time much longer than childhood. For example, Hsu and Chater (2010) estimate an age of acquisition that's thousands of years for learning one piece of linguistic knowledge.[15] This is clearly longer than childhood (and human lifespans more generally). Therefore, even if there's uncertainty about precisely when children converge on that knowledge, we can be pretty sure poverty of the stimulus is happening under this MDL approach because children do in fact converge on that knowledge before thousands of years have passed.

As with the developmental benchmarking discussed above, the assumptions in place are key: what the correct hypothesis is, what the space of competing hypotheses is, what the available data are, and how children leverage their data. For an MDL learner, another major decision is how to encode the representations under consideration (that is, the correct hypothesis and the competing hypotheses) – this is where prior linguistic knowledge and abilities (whether innate or derived from experience) come in. For the available data, an MDL learner is happy to encode both ambiguous and unambiguous data. What separates this from the developmental benchmarking approach in the previous section is that an MDL approach will function even if there aren't any unambiguous data (as is the case in Figure 6). This is because the MDL learner already assumes prior knowledge where the learner prefers compactness; so, compactness can decide between representations even when the data are ambiguous. This in turn relates to how an MDL learner leverages the available data: this learner is capable of encoding both the representation itself and the data via that representation as optimally as possible. The optimal encoding is what makes this approach an ideal learner analysis of poverty of the stimulus Then, if there's poverty of the stimulus even for an ideal learner, there's likely to be poverty of the stimulus for children, who aren't obviously ideal learners.

3 Some potential examples and how to interpret them

I now want to step through a few illustrative examples of linguistic knowledge that have been considered instances of poverty of the stimulus. For each one, I'll highlight what the correct hypothesis is, what the hypothesis space is, what the available data are, and how children leverage those data. Where available, I'll also include what we know about children's age of acquisition when they make their constrained generalizations. This makes it easier to see what the original assumptions were that caused people to think poverty of the stimulus was occurring, and how those assumptions may have changed across different investigations of the same linguistic knowledge piece.

[15] The particular knowledge was when to contract *want to* into *wanna* in English. That is, English speakers know they can contract *Who do you want to rescue?* into *Who do you wanna rescue?*, but not *Who do you want to do the rescuing?* into *Who do you wanna do the rescuing?*

I'll conclude the discussion of each example with what prior knowledge or abilities seem to be required, given the existing investigations, and how this relates to the linguistic vs. non-linguistic nativist perspective. The investigations I'll discuss are mathematical or computational approaches (similar to those discussed in section 2), as these are particularly helpful for investigating an acquisition problem that's been concretely specified in terms of the hypothesis space, the data available, and how children leverage the available data (Clark & Lappin, 2011; King, 2015; Chater et al., 2015; Fodor & Sakas, 2017; Pearl, in press-b). More specifically, these approaches are helpful because solutions that work in a concretely-specified investigation of poverty of the stimulus are an existence proof for how that specification of poverty of the stimulus could be solved. We can then look at the components of these solutions and consider their origin. When these components seem likely to be innate, we can think about whether they're language-specific (and so support the linguistic nativist view) or not (and so support the non-linguistic nativist view).

I also want to note that poverty of the stimulus is generally discussed for syntactic knowledge (e.g., structure dependence, English anaphoric *one*, syntactic islands, verb-raising, the linking problem, binding, verb alternations: Baker & McCarthy, 1981; Hornstein & Lightfoot, 1981; Lidz, Waxman, & Freedman, 2003; Regier & Gahl, 2004; Pearl & Lidz, 2009; Hsu & Chater, 2010; Perfors et al., 2010; Hsu et al., 2011; Pearl & Sprouse, 2013b; Chater et al., 2015; Han et al., 2016; Pearl & Mis, 2016; Cole, Hermon, & Yanti, 2017; Pearl, 2017; Pearl & Sprouse, 2018; Pearl, in press-b). However, poverty of the stimulus occurs all over linguistic development (and cognitive development more generally), whenever the available data are compatible with more than one hypothesis (Scholz & Pullum, 2006; Clark & Lappin, 2011). With this in mind, I'll review a syntactic example that's received a lot of attention, a lexical semantic example that's received some attention, and a phonological example that hasn't received much attention yet.

It's important to reiterate that there's far more discussion of specific examples of poverty of the stimulus than what I can reasonably cover in this review. The examples I selected here are meant to show the breadth of poverty of the stimulus in language development, though the lexical semantic and phonological examples aren't ones that linguistic and non-linguistic nativists have (yet) argued about. As I mentioned above, the syntactic example is only one of many discussed – I selected it because it's been discussed a lot and (in my view) caused a lot of tears due to misinterpretation. With that said, I should note that other syntactic cases (like the ones mentioned above after structure dependence) are often more impactful for linguistic nativists, as they concern specific theories of linguistic representation that are more likely to involve language-specific knowledge. So, any innate components needed to solve poverty of the stimulus for these syntactic knowledge pieces are likely to support the linguistic nativist viewpoint. Interestingly, there haven't been many computational or mathematical investigations of these syntactic knowledge pieces. So, that's an exciting area for future work on poverty of the stimulus that can offer concrete evidence for the debate between linguistic and non-linguistic nativists.

3.1 Syntactic: Structure dependence

The correct hypothesis and the hypothesis space. Linguistic representations are generally agreed to be hierarchical, with units inside other units (e.g., phrases inside other phrases). Moreover, rules for manipulating linguistic elements rely on this structure, rather than operating over

simpler representations such as linear word order. A traditional example of this is complex yes/no question formation in English, where we can think about a yes/no question like (3a) as being a *structure-dependent* transformation of the declarative utterance in (3b) that involves the same core contentful elements.

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

Why do we call this structure-dependent? One common way to describe the transformation in (3) is with the rule "Move the auxiliary in the main clause to a position at the front of the utterance". This rule relies on the structural notion of "main clause".

Importantly, there are other potential rules that *aren't* structure-dependent (i.e., they're *structure-independent*), such as "Move the last auxiliary", "Move the first auxiliary", "Move the auxiliary that's fourth from the end of the sentence", and so on. Structure-independent rules aren't able to account for the full range of English complex yes/no questions, though some (like the first two mentioned above) are often compatible with most yes/no questions in child-directed speech.

So again, for English yes/no question formation, the correct hypothesis is something like "Move the main clause auxiliary to a position at the front of the utterance." The hypothesis space consists of all possible rules for transforming (3b) into (3a), including this structure-dependent one, *other* structure-dependent ones (e.g., "Move the auxiliary that's in the same phrase as the main topic", where phrase is the structure-dependent component), and structure-independent ones.

Of course, structure-dependence is meant to apply to the linguistic system as a whole. So, we can translate this particular yes/no-question correct hypothesis and corresponding hypothesis space to any other linguistic knowledge piece as well. More generally, the idea would be to translate the correct hypothesis to something that applies to the entire linguistic system (Scholz & Pullum, 2006): an *overhypothesis* like "Use structure-dependent rules." Any specific linguistic knowledge piece then has a specific correct hypothesis and corresponding hypothesis space that is defined using this overhypothesis for that knowledge piece (Goodman, 1955; Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2008; Perfors, Tenenbaum, & Regier, 2011; Pearl, in press-b).[16]

The available data (or so we initially thought). Interestingly, when people first considered how children would learn that their linguistic system uses structure-dependent rules, the focus was on the individual knowledge pieces (like complex yes/no questions) rather than the system as a whole. This initial focus was relaxed later on, as discussed in more detail below. However, this focus meant that the available data initially considered were only those related to English yes/no questions specifically (e.g., see the discussion in Pullum and Scholz (2002) and Legate and Yang (2002)). Various analyses of children's possible input suggested that the vast majority of their data were ambiguous between the correct structure-dependent rule and other competing rules, including structure-independent ones.

[16]Note that overhypotheses are very similar to the traditional notion of linguistic parameters often used by linguistic nativists, where a linguistic parameter is a piece of abstract structural knowledge that can be applied to may different linguistic knowledge pieces. See Pearl and Lidz (2013) and Pearl (in press-b) for more discussion on this point.

How children leverage the available data. Initial discussions of structure-dependence assumed children could only use direct positive evidence (i.e., for this knowledge piece, examples of yes/no questions); moreover, only unambiguous data were informative (e.g., see the review of prior discussion in Pullum & Scholz, 2002). Given these assumptions, the lack of sufficient unambiguous data for structure-dependence when forming complex yes/no questions in English seemed like it could lead to poverty of the stimulus.

Age of acquisition for constrained generalizations. Based on behavioral work by Crain and Nakayama (1987), English children as young as three years old seem to know that rules controlling complex yes/no question formation in English ought to be structure-dependent. From an input standpoint, it turns out that English children are likely to encounter very few examples of complex yes/no questions that would unambiguously indicate that a structure-dependent rule is required (Pullum & Scholz, 2002; Legate & Yang, 2002). So, English children's relatively rapid development of this structure-dependent knowledge seemed very much to be an instance of overcoming poverty of the stimulus.

The investigations. One early computational investigation by Reali and Christiansen (2005) rejected the assumption that children were trying to explicitly learn a particular structure-dependent rule, given a hypothesis space of both structure-dependent and structure-independent rules. Instead, this modeled learner was tuned to children's observable behavior, and learned to distinguish grammatical complex yes/no questions like *Is the boy who is watching Mickey Mouse happy?* from ungrammatical ones like **Is the boy who watching Mickey Mouse is happy?* In addition, this investigation rejected the assumption that the only available data were those related to yes/no questions. Instead, this modeled learner learned from all the available utterances, and leveraged the relative frequencies of 2-word and 3-word sequences (e.g., 2-word: *Is-the*, *the-boy*, *boy-who*; 3-word: *Is-the-boy*, *the-boy-who*). These relative frequencies were then used to successfully distinguish grammatical from ungrammatical complex yes/no questions. This result suggested that there wasn't poverty of the stimulus in principle, once the hypothesis space and available data were redefined.

However, Kam, Stoyneshka, Tornyova, Fodor, and Sakas (2008) demonstrated that the modeled learner of Reali and Christiansen (2005) benefited from a "lucky fluke" between the particular sample of input it learned from and the particular set of complex yes/no questions it had to judge. When this modeled learner judged a wider range of complex yes/no questions, it failed to make the distinctions that children were observed to. Therefore, it seemed that poverty of the stimulus did still exist, even with the redefined hypothesis space this modeled learner assumed.

A later investigation by Perfors, Tenenbaum, and Regier (2011) broadened the focus back to structure-dependence in the linguistic system. Here, the modeled learner considered a hypothesis space of different representation types (some structure-independent and some structure-dependent, including the correct structure-dependent one). Of course, these representational hypotheses connected to many individual phenomena, like complex yes/no questions – so, there would be a representation of complex yes/no questions for each hypothesized representation type. Importantly, because the hypotheses weren't *only* tied to complex yes/no questions, the modeled learner could

learn from any data that might be informative for choosing among representation types. This meant the available data for the modeled learner was the set of all utterances children encounter, not just the set of complex yes/no questions they encounter. Children would then need to be able to identify the sequences of syntactic categories that comprise each utterance (e.g., viewing *the kitty slept* as *determiner noun verb*). However, the available data were still compatible with all the hypotheses under consideration (i.e., there was unresolvable ambiguity) – so, poverty of the stimulus persisted, despite a wider base of data to learn from.

Perfors, Tenenbaum, and Regier (2011) offered a concrete solution by having their modeled learner leverage the available data with a simplicity-based approach similar to that of information-theoretic approaches (Chater et al., 2015). In particular, the learner used Bayesian inference to identify the representation that balanced the complexity of the representation with the ability of that representation to encode the data. With this prior preference in hand, the modeled learner successfully identified the correct structure-dependent representation from those available in the hypothesis space, on the basis of the data children encounter. So, one solution in principle to the poverty of the stimulus presented by structure-dependence is for children to have this simplicity bias, and then they can deploy that bias successfully over their available data. This is non-linguistic knowledge, and so would be compatible with the non-linguistic nativist perspective. Of course, children also need the ability to define the hypothesis space so that it includes the correct structure-dependent one. Whether that hypothesis space creation requires innate linguistic knowledge and/or abilities is still an open question.

A subsequent investigation by Abend et al. (2017) also approached the learning problem as a question of the representation that was appropriate for the larger linguistic system, rather than for an individual knowledge piece like complex yes/no questions. An interesting difference from the Perfors, Tenenbaum, and Regier (2011) investigation was the nature of the hypothesis space: rather than explicitly defining several structure-dependent and structure-independent representations, Abend et al. (2017) implicitly defined a hypothesis space of infinite size via a set of pre-defined structure-dependent building blocks and constraints on how those building blocks can combine. In this way, the modeled learner assumed structure-dependence already,[17] but didn't know *which* of infinitely many constructable structure-dependent representations was the correct one. During learning, the modeled learner would generate explicit structure-dependent hypotheses on the basis of these building blocks and constraints, and then would evaluate those hypotheses on the basis of the available data.

Importantly, this modeled learner saw syntactic structure as part of a larger system of linguistic representation, with syntax intimately connected to meaning (in the form of precise logical representations). Because of this, the learner of Abend et al. (2017) not only used all the available utterances, but also both the syntactic and meaning information in those utterances, as well as the meaning information from the surrounding non-linguistic context.[18]

[17]We might imagine this prior knowledge about structure-dependence could result from a Perfors, Tenenbaum, and Regier (2011) learner who figured out that structure-dependence was the right overhypothesis.

[18]Note that this meant the modeled learner had to be able to reliably extract both the syntactic and meaning information from the input. This is of course an idealization – perhaps especially for the meaning information – though see Pearl (in press-b) for more discussion about why the particular meaning information that the learner had to extract isn't implausible for young children to be able to extract.

Still, with only this to work with, there seems to be poverty of the stimulus, because more than just the correct structure-dependent representation is compatible with the available data. Abend et al. (2017)'s learner offers a concrete solution that's similar to that of Perfors, Tenenbaum, and Regier (2011): the learner has a prior bias to use Bayesian inference, balancing the complexity of a representation with how well that representation encodes the data. Like the learner of Perfors, Tenenbaum, and Regier (2011), this learner assessed how well the representation encodes the data; however, when calculating how costly the representation itself was to encode (i.e., its complexity), the learner had an additional bias to prefer representations built from building blocks that were more frequently used. The intuition is that reusing building blocks that have been used before is efficient, so representations that reuse more building blocks are more efficient (i.e., less costly) than representations that don't reuse building blocks. These combined biases for "simple" representations and representations that reuse common building blocks is sometimes referred to as a bias for "rational rules" (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), where "rational" means optimal. So, a learner with a bias for rational rules looks for "rules"' (i.e., representations) that are optimal – here, that means representations that are both as simple as possible and also as efficient as possible (in the sense of reusing common building blocks).

With these biases, the modeled learner was able to infer the representation (both syntactic and meaning) that adults use far more often than chance, given just a few thousand utterances of realistic child input.[19] So, Abend et al. (2017) offer a promising solution in principle to the poverty of the stimulus occurring over a much larger hypothesis space; this solution relies on both linguistic and non-linguistic prior knowledge and abilities. In particular, we might classify as linguistic knowledge (i) the structure-dependent building blocks for linguistic representations, (ii) the constraints on how structure-dependent building blocks can combine, and (iii) the tight connection between syntactic representations and logical representations. We might classify as non-linguistic knowledge or abilities (i) the ability to perceive meaning from non-linguistic context, (ii) using Bayesian inference, and (iii) preferring rational rules that maximize simplicity and reuse in potential representations.

Another recent approach by Fitz and Chang (2017) also capitalized on the connection between syntactic representations and meaning. Like Abend et al. (2017), this meant Fitz and Chang (2017) had their modeled learner use both syntactic and meaning information available in children's utterances, though they focused on a subset of all possible utterances. However, like Reali and Christiansen (2005), Fitz and Chang (2017) rejected the assumption that children were trying to learn a particular structure-dependent rule for complex yes/no questions; instead, this modeled learner was tuned to observable behavior. In particular, if the modeled learner was given a meaning that adults would express using a complex yes/no question, would it generate the correct complex yes/no question form (i.e., the correct sequence of words that make up a complex yes/no question)?

Fitz and Chang (2017) provided a concrete solution to this learning problem by creating a learner that (i) knew that meaning representations were hierarchically structured, and (ii) observed that sequences of words were generated from those underlying meaning representations. A neural

[19]This investigation was restricted to a small input set due to how long it takes to generate accurate annotations of available syntactic and logical information for child input utterances. We might reasonably infer that the modeled learner would show even more impressive performance if it was given a quantity of data more in line with what children truly learn from.

network was used to navigate the space of possible ways to generate observable sequences from underlying meaning representations, given the available data.[20] After training, this modeled learner was indeed able to generate the appropriate sequence of words for a complex yes/no question, given the intended question's meaning representation. This suggests that a learner who knows that (i) meaning representations are hierarchical, and (ii) meaning representations are used to generate the observable word sequences can learn to produce specific observable sequences that correspond to complex yes/no questions. Notably, it's unclear what the syntactic representation looks like that mediates between the hierarchical meaning representation and the observed sequence of words – perhaps it's the one English adults are thought to use, or perhaps it isn't. This is one of the tricky things about using non-symbolic approaches like neural networks to search a hypothesis space – they're notoriously difficult to interpret (Dunbar, 2019; Pearl, 2019).

McCoy et al. (2018) also use a neural network approach, but instead keep the assumption that the learner is trying to identify a particular structure-dependent representation from the hypothesis space of possible representations. Like the learner of Fitz and Chang (2017), this learner generates an appropriate sequence of words corresponding to a complex yes/no question; unlike the learner of Fitz and Chang (2017) however, this learner is given the declarative version of the complex yes/no question (rather than the meaning representation of the question). The modeled learner was also given access to a subset of declarative utterances of the form that children might encounter in their input (both without modifiers like *the penguin can laugh* and with modifiers like *the penguin that the seals saw can laugh*). Given the restricted dataset, the available data were compatible with many different representations, including both structure-independent and structure-dependent ones, thereby leading to poverty of the stimulus.

McCoy et al. (2018) investigated several neural-network-based learners using a variety of architectures, and discovered that one modeled learner solved the task of transforming a declarative utterance into an equivalent complex yes/no question by relying on a structure-dependent representation. (Other learners relied on structure-independent representations.) This result suggests two main implications. First, it's not necessary for the learner to already have an explicit preference for structure-dependent representations in order to learn them from the available data.[21] This is particularly salient given the restricted input set McCoy et al. (2018) had their learners learn from, and aligns with the previous results of Perfors, Tenenbaum, and Regier (2011) that relied on Bayesian inference. Second, because only one of the modeled learners achieved structure-dependent representations, this underscores the poverty of the stimulus. To infer structure-dependent representations from the available data, children need some kind of internal bias. Here, it's whatever was encoded implicitly in the neural network architecture; for Perfors, Tenenbaum, and Regier (2011),

[20] Note that a neural network's non-linear learning process is another way to navigate a large hypothesis space, though it's hard to interpret exactly *how* that space is being navigated, in contrast with symbolic approaches like Bayesian inference (Pearl, 2019). Still, neural network results do represent an "in principle" solution to learning problems, confirming what it's possible to learn, given certain input and learning constraints (as encoded in the neural network's architecture).

[21] However, the caveat is that the innards of neural networks are hard to interpret, so it's possible (though perhaps not highly probable) that an explicit preference for structure-dependent representations was in fact present in the specific numbers in the vectors inside the neural network. It's just that we as humans can't yet easily interpret vectorized representations, and so it's difficult to tell for sure.

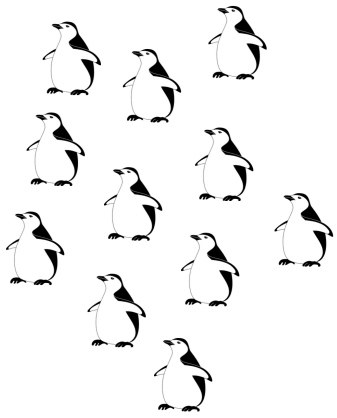
it's a bias for overall compactness of the representation and the ability of that representation to encode the data.

How to interpret these investigations. From this collection of structure dependence investigations, it seems to me that there are several key takeaways. First, as mentioned above with the McCoy et al. (2018) investigation, restricting the relevant input easily leads to poverty of the stimulus that requires child-internal biases to solve. So, one useful bias on the learner's part is to consider any individual linguistic knowledge piece as part of a larger linguistic system (e.g., the syntactic system as a whole, or the syntactic and semantic system together). Then, many more data are able to provide information with respect to the preferred underlying representation. A second useful bias is a non-linguistic one: to prefer "rational rules" that prioritize compact representations, the ability of those representations to encode the available data compactly, and representations that rely on reusable building blocks. Additional useful biases might be construed as linguistic: (i) which structural building blocks to use (whether syntactic or logical/conceptual), and (ii) knowing about the tight relationship between syntax and meaning. To the extent these latter biases are truly linguistic (i.e., they're not applicable outside the domain of language), they would represent innate, linguistic knowledge. However, it remains to be seen if they can perhaps be derived from other non-linguistic biases children might have. (For example, perhaps the knowledge of the tight relationship between syntax and meaning is some kind of simplicity bias that assumes maximal similarity between representational systems, unless shown otherwise. But that's a question for future work.) More generally, because some of the useful innate biases in current solutions to this poverty of the stimulus seem to be linguistic, the linguistic nativist viewpoint is supported. This of course comes with the caveat just mentioned above: if all the linguistic biases turn out to be derived from other non-linguistic biases, then the non-linguistic nativist viewpoint would be supported.

3.2 Lexical semantic: Exact cardinal number words

The correct hypothesis and hypothesis space. Exact cardinal number words are those like *one, two, seven, eleven, five hundred eighty three,* and *one million seven hundred sixty four thousand three hundred forty six* in English. They refer to specific quantities, and allow speakers whose languages use these kinds of terms to answer questions like "How many penguins are there?" more precisely than *several*, *some*, or *a lot*. (For instance, in Figure 7, there are eleven penguins.) Moreover, languages with exact cardinal number terms also have counting lists that begin with the equivalent of *one*, and then move through exact cardinal number words one by one, with the next number in the list differing by exactly one from

Figure 7: A scenario where exact cardinal number words like *eleven* and a counting list are helpful.



the previous number in the list: *two, three, four, five,..., eleven, twelve,...* and so on, in English. This is a tool speakers can use to answer the "How many" question. For example, in Figure 7, one strategy is to use the English counting list, where each word in the counting list is mapped to an individual penguin. When we run out of penguins, the exact cardinal number in the counting list that we stopped with is *eleven*, and so that's how many penguins we have.

Importantly, adult speakers don't explicitly know the entire counting list of their language (e.g., memorizing the lexical items *one* through *one million seven hundred sixty four thousand three hundred forty six*). Instead, adults explicitly know some key items, and then generate the counting list as needed for a specific situation (like counting lots of penguins). One key intuition is the *successor function* (Kripke, 1982; Carey, 2009, 2011; Rey, 2014; Sarnecka, 2016): the next exact number (i.e., the successor of the current exact number) differs from the current exact number by one. So, to generate the numerical meaning for a particular exact cardinal number word (like *eleven*), look to the previous exact cardinal number word on the list (i.e., *ten*) and add one. This relationship between exact cardinal number words in the counting list and exact cardinal number meaning, via the successor function, is also known as the *Cardinality Principle* (Wynn, 1990, 1992; Carey, 2009, 2011; Sarnecka, 2016).

So, a key element to understanding exact cardinal number words (and generating ones you haven't explicitly heard before on the fly) is the Cardinality Principle. The Cardinality Principle is thus a vital component of the adult representation of all exact cardinal number words, and what children need to converge on. However, it turns out that the hypothesis space is pretty vast for what exact cardinal number words could mean. For instance, one hypothesis is that exact cardinal number words are just related properties, the way color words are. Like number words, color words can be recited in a certain order in English: *red, orange, yellow, green, blue, purple*. But, this doesn't indicate that *red* somehow means one unit "more" than *orange* or that *green* means one unit "less" than *blue*; this situation contrasts with the exact cardinal number words in the English counting list. Put simply, why should a child assume a successor function for exact cardinal number words, which are in an ordered list, when color words, which are also in ordered list, don't use this function?

Even if the child knows the successor function is a key building block, there are still a variety of hypotheses using that building block that aren't right. For example, one hypothesis is to use the successor function if the number is less than 10, but after that, only use the ones digit to determine the exact number word meaning. So, *one* through *nine* mean what adults think those words mean; but, *two*, *twenty two*, and *five thousand six hundred twenty two* all mean the same thing, which is what adults think *two* means. (This is equivalent to a "mod 10" function.) This might seem like a strange hypothesis, but timekeeping systems work like this: there are 60 seconds in a minute, but then we loop back to 0 when counting seconds. So, seconds use a mod 60 function. The same is true for minutes in an hour, while hours in a day use a mod 24 function (24 hours in a day), and days in a month use a mod 28, mod 29, mod 30, or mod 31 function (28-31 days in a month), depending on the month.

The above hypotheses are just two of many. To get a sense of some others, imagine a "mod X" system, where X can be anything between two and whatever number you like. Then there are hypotheses which might seem even stranger: assume a successor function, except for the number

eleven, which gets its meaning by adding two to the previous word, instead of one. (So, *eleven* means what adults thinks *twelve* means.) Once again, substitute any single number word you like for *eleven*, and any quantity for two, and you can generate all kinds of different hypotheses.

The available data and how children leverage it. As far as I'm aware, researchers investigating the acquisition of exact cardinal number words assume children have access to direct positive examples, such as scenes where there are eleven penguins and someone using the word *eleven* to describe how many penguins there are. More generally, the relevant data points are assumed to be pairings of a specific number of countable things present and a cardinal number word labeling that set of countable things. Children are then assumed to be able to observe and leverage the correspondence between the set size and the number word.

Interestingly, Piantadosi, Tenenbaum, and Goodman (2012) found that actual child interactions involving these data points appear in a power-law distribution, where most instances children hear are of *one* or *two*, with *three* and above occurring in very, very few data points. As you might imagine, this means children are primarily inferring what exact cardinal number words mean on the basis of *one* and *two*, with only the occasional higher-number word appearing. So, this seems like a solid case of poverty of the stimulus. Moreover, even if we had a lot of examples of lots of different exact cardinal number words, we'd still have the problem of not encountering every cardinal number word that can be generated (like *one million seven hundred sixty four thousand three hundred forty six*). This could cause children to end up with mod 60 systems, or mod 1000 systems, and so on. But even worse in reality, it seems that children's input is highly skewed towards a very few cardinal number words that might well permit additional incorrect hypotheses. So, it seems like children must have some prior knowledge to help them constrain their generalizations appropriately for what exact cardinal number words mean.

Age of acquisition for constrained generalizations. The acquisition of exact cardinal number knowledge appears to follow a particular, staged trajectory: children settle on a variety of immature representations before converging on the adult representation that involves the Cardinality Principle (Carey, 2009, 2011). This is true not just in English but also in Arabic, Japanese, Mandarin Chinese, Russian, Slovenian, and Tsimane' (Sarnecka, 2016). Children initially learn the ordered counting list of their language (e.g., *one, two, three,...*) without knowing what these words mean (Fuson, 1988; Piantadosi et al., 2012; Sarnecka, 2016) – children at this stage are sometimes called "pre-number-knowers". So, when a pre-number-knower is asked to give someone three penguins, the child will simply pick out some number of penguins to give, even when told to use the counting list (Sarnecka, 2016).

Children then progress through several *subset-knower* stages: one-knower, two-knower, three-knower, and sometimes even four-knower (Wynn, 1990, 1992; Carey, 2009, 2011; Piantadosi et al., 2012; Sarnecka, 2016). A child at a particular knower stage (e.g., three-knower) seems to have the correct representation for each exact cardinal number up to and including that particular number (e.g., *one*, *two*, and *three* for a three-knower). For instance, if we asked a three-knower to give us three penguins, she would happily hand us exactly three penguins. However, any numbers beyond that number (e.g., four for a three-knower) seem to be unknown, despite the child knowing

the words in the counting list (e.g., *four*). For instance, if we asked a three-knower to give us four penguins, she'd give a random number greater than three (e.g., four, seven, or nine penguins).

After the three-knower or four-knower stage, children seem to make the leap to the Cardinality Principle, where they effectively learn all the rest of the exact number meanings at once (Wynn, 1990, 1992; Carey, 2009, 2011; Sarnecka, 2016). As mentioned above, this new adult-like representation involves recognizing the relationship of the successor function to the counting list (i.e., the next word on the counting list is one more than the previous word on the counting list); having this representation allows children (like adults) to generate the meaning of any particular exact cardinal number word. Interestingly, the progression from pre-number-knower to knowing the correct representation involving the Cardinality Principle takes a surprisingly long time – for example, it can take several months to go from being a one-knower to a two-knower in English. In English, children seem to settle on the Cardinality Principle around age four to five (Sarnecka, 2016).

The investigation. Piantadosi et al. (2012) used an ideal learner model to attempt to understand why children might have this staged trajectory that then leads to the correct representation involving the Cardinality Principle. That is, how could this very specific trajectory emerge, given children's input?

The modeled learner's hypothesis space was a set of building blocks that could be combined to generate explicit hypotheses about the meaning of exact cardinal number words in the counting list. These building blocks were domain-general, in that they weren't applicable only to language – in fact, they could be used for learning kinship relations, taxonomies, and theories of causality, in addition to meaning. The specific building blocks included what might be thought of as conceptual units (e.g., a function checking whether a set was size one), comparison units (e.g., a function identifying the intersection of two sets), combinatorial units (e.g., logical *and* and *if*), sequential relations (e.g., a function that identifies the next word in the counting list), and the ability to construct recursive relations. Notably, the successor function itself wasn't a building block – rather, it could be constructed from the building blocks in the hypothesis space. (Importantly, so could an infinite number of other hypotheses, including the mod systems mentioned before.) So, the correct representation was merely one of the explicit hypotheses that could be generated from these building blocks and evaluated; others were the immature representations of pre-number-knowers, one-knowers, two-knowers, three-knowers, and four-knowers; and still others were representations that would correspond to five-knowers, or representations that would correspond to a two-not-one-knower who recognizes the meaning of *two*, but not the meaning of *one* (or numbers higher than two).

The modeled learner was given a realistic sample of English children's exact cardinal number input, where a data point was an exact cardinal number word like *two* paired with two items. The modeled learner then used the "rational rules" approach (Goodman et al., 2008) to evaluate which representation was the most probable at different points in time (which corresponded to different amounts of data). In particular, the learner had built-in biases to prefer both simplicity (in terms of a compact representation) and reuse (in terms of using building blocks that had been used before). Moreover, representations involving recursion, like the correct one with the Cardinality Principle, received an extra penalty – that is, because they involved a more complex building block

(recursion), the modeled learner was biased against them. So, these representations involving recursion had to be even better at compactly encoding the data and reusing building blocks in order to be viewed as more probable than equivalent representations that didn't involve recursion.

With this setup, the ideal modeled learner was then able to reproduce the specific developmental trajectory for exact cardinal number knowledge. In particular, the learner went through progressive immature stages corresponding to the different knower levels before converging on the correct representation involving the Cardinality Principle. Thus, Piantadosi et al. (2012) provided a concrete strategy for potentially solving the poverty of the stimulus involved with mapping exact cardinal number words to their meanings, in the same way that children seem to. This strategy identified the child-like hypotheses at each age because those hypotheses, including the correct representation at the end of learning, were a better balance of compactness, reuse of building blocks, and capturing the input data than other hypotheses (e.g., the mod representations).

How to interpret this investigation. Based on this investigation, I think there are two built-in components that collectively offer a solution to this instance of poverty of the stimulus. First, the child has to have a set of useful building blocks for generating specific hypotheses about what exact cardinal number words mean. The specific collection that Piantadosi et al. (2012) used appears to be domain-general – that is, while the building blocks were clearly applied to language items (meaning and counting words), the building blocks themselves weren't obviously language-specific. (Remember: they could be used for non-language concept learning, and not just when mapping concepts to specific linguistic forms.) Second, the child has to have a preference for "rational rules": compact representations and representations that reuse building blocks. With this preference in hand, a child could converge on the correct representation involving recursion, even if the child views recursive representations as inherently more complex (and so less preferred, all other things being equal). Because both these components aren't specific to language, this solution to the poverty of the stimulus is in line with the non-linguistic nativist viewpoint.

I should note, however, that this poverty of the stimulus case isn't one that linguistic nativists have previously held up as one likely to involve innate linguistic knowledge. In fact, I don't think that lexical semantic development is generally argued to involve innate linguistic knowledge. Because of this, poverty of the stimulus in lexical semantic development tends to be less interesting for the linguistic vs. non-linguistic nativist debate. That said, poverty of the stimulus occurs a lot in lexical semantic development, including the specific case here of learning the meanings of exact cardinal number words. Just like in other areas of linguistic development, the goal is to understand how children can solve these poverty of the stimulus instances, wherever they occur.

3.3 Phonological: Word-final obstruent devoicing

The correct hypothesis and hypothesis space. In response to skepticism that poverty of the stimulus occurs in phonology, Idsardi (2005) notes an example involving particular kinds of consonants that appear at the end of words. The specific consonants are called *obstruents*, because they obstruct the airflow through the vocal tract; obstruents include voiced sounds like /d/, as well as voiceless sounds like /t/, which are made the same way in the vocal tract, except that the vocal

folds don't vibrate as early (or at all) for the voiceless sounds.

A certain process can occur when obstruents appear at the end of words (*word-finally*): if they're voiced like /d/, they become devoiced, like /t/. This word-final devoicing process occurs in a large number of the world's languages, including Dutch, Catalan, Bulgarian, Czech, Macedonian, Armenian, Maltese, Tok Pisin, and Wolof, among many others. So, for example, the Dutch word for the plural "dogs" is *honden*, while the singular "dog" is spelled *hond* but pronounced as if it had a /t/ on the end ("*hont*"). The reason *hond* is pronounced *hont* is due to word-final obstruent devoicing (i.e., the voiced /d/ becomes voiceless /t/ because it's word-final).

So, one way to describe the correct rule that children must learn if their language uses this process is this: if you have a word-final obstruent, devoice it. However, there are certainly many other plausible hypotheses that seem to be compatible with the available data, as Idsardi (2005) points out. For instance, a more specific hypothesis might be to devoice specific obstruents only (like /d/), rather than assuming all obstruents are devoiced. A more general hypothesis might be devoice the last consonant in a word, even if vowels come after. So, this more-general hypothesis would turn *hond* to *hont*, but would also turn the nonsense Dutch word *honde* into *honte*. A hypothesis that's not obviously more specific or more general than the correct representation might key into the singular vs. plural distinction rather than word-finality: perhaps the last obstruent in a singular like *hond* is devoiced to *hont*, while the last obstruent in a plural like *honden* isn't (note that /n/ is a non-obstruent consonant, so the /d/ would be the last obstruent in *honden*).

The available data and how children leverage the data. The relevant data for children would presumably be all words, with the correct rule covering words that show word-final devoicing (like *hond* pronounced as if it were *hont*) and not covering words where devoicing doesn't occur. To do this, children would need to recognize that something changed about the way the devoiced forms were pronounced – that is, they'd have to expect *hond* and then note that *hont* is what comes out instead. Children may be able to make this connection via the similarity in meaning to the plural *honden*, especially if they recognize the suffix *-en* as a plural marker. So, the child reasoning would be something like *honden* - *en* = *hond*. When *hont* comes out instead, this is a signal that devoicing happened. In terms of child capability, this doesn't seem like an unreasonable realization, depending on the age of the child. We know that children learn to recognize phonetic features like voicing before one year old (Maye, Weiss, & Aslin, 2008), and also seem capable of recognizing some regular morphology by two or three (Brown, 1973).

However, to truly know which specific word forms are available to children at certain ages for learning about word-final devoicing, a corpus analysis would probably be the best way to go. As far as I know, this has yet to be undertaken. Depending on what forms are found in any particular language, it's very possible poverty of the stimulus exists for learning about word-final obstruent devoicing.

Age of acquisition for constrained generalizations. As far as I'm aware, we don't currently know a precise age of acquisition for the correct devoicing rule; we're also unaware of intermediate stages children might go through when trying to decide what the correct devoicing rule is, though see Bergelson and Idsardi (2009) for some investigations about the generalizations both 8-month-

olds and adults make about final devoicing. But, there's some good news. First, Idsardi (2005) notes that adults who speak languages with this devoicing rule do in fact seem to use it, even on items they haven't heard before (and so they couldn't have just memorized the specific words where it happens). Additionally, Bergelson and Idsardi (2009) find that adults can learn a word-final devoicing rule in an artificial language where multiple generalizations are possible (i.e., a word-final devoicing rule and other possible devoicing rules). So, no matter what, we know two things: (i) from Idsardi (2005), that children eventually do figure out the correct devoicing rule because adults know it, and (ii) from Bergelson and Idsardi (2009), that some kind of bias is active that allows adults to pinpoint the right kind of devoicing rule even when the data are ambiguous.

Second, there are a lot of data about child pronunciations from a variety of languages available through the CHILDES PhonBank database (MacWhinney, 2000). So, we may be able to identify when children seem to produce the correct pronunciations that show word-final devoicing. Of course, even better would be more behavioral studies that test children of different ages on novel words – then we can be sure the children are relying on whatever internal rule they have, rather than just repeating what they've heard.

Investigations and how to (eventually) interpret them. Investigations of this potential instance of poverty of the stimulus are, of course, yet to be done. But, it may well be that a rational rules approach might work well here, especially if we can define useful building blocks from which children could generate explicit hypotheses to test. I do note that Idsardi (2005) cautions about how to implement the "simplicity" component of a rational rules approach; in particular, depending on what form the representations take, which representation is more compact may not be obvious. More generally, I think this phonology case is a concrete, potential instance of poverty of the stimulus that we have the tools to investigate. Depending on what solutions are found, we can see if linguistic or non-linguistic nativists are more enthusiastic.

4 Conclusion

Here, I've tried to give a tears-free overview of poverty of the stimulus, including what it is and how it relates to children's hypothesis spaces, the data available to children, and how children leverage those available data. Importantly, poverty of the stimulus is part of an argument (i.e., the "argument from poverty of the stimulus") that includes children making constrained generalizations, given the data available (and in practice, making those constrained generalizations faster than we might expect). This constrained generalization then implies that children have some kind of prior knowledge or abilities that guide them towards the correct answer. The prior knowledge or abilities for a particular instance of poverty of the stimulus are generally believed to involve some kind of innate knowledge or abilities, and the big divide is whether any of that innate knowledge or those innate abilities are language-specific. If so, the linguistic nativists are pleased; if not, the non-linguistic nativists are pleased.

I also reviewed some approaches to identifying when poverty of the stimulus is in fact happening, both in principle and in practice, both abstracting away from specific linguistic details and by zooming in on specific case studies. I then reviewed a few illustrative examples of poverty of the

stimulus, discussing why these are considered poverty of the stimulus, how different researchers have approached investigating them (or could in the future), and what we might interpret from current (or future) results. In my view, these examples highlight a way to make progress on poverty of the stimulus (and thus language development more generally) by (i) defining the learning problem as precisely as possible, (ii) assessing quantitatively if poverty of the stimulus is happening, and (iii) showing how specific learning proposals could solve that particular instance of poverty of the stimulus. As I mentioned in the introduction to the case studies, the ones I chose were meant to illustrate the breadth of poverty of the stimulus in language development, rather than focus on cases that have figured prominently (or at all) in the linguistic vs. non-linguistic nativist debate. This means that many cases that I didn't discuss (particularly in the realm of syntax) are ones that may offer more impactful data for the debate between linguistic and non-linguistic nativism. Making concrete progress on these other cases can then certainly move that debate fruitfully forward.

With that said, I think the value of poverty of the stimulus goes beyond the linguistic vs. non-linguistic nativist debate. Poverty of the stimulus is a particular (and particularly frequent) learning problem that children solve (and in fact solve a lot for years on end) when developing their language knowledge. Understanding precisely what children must solve and providing concrete ways that they might in fact solve it each time allows us to better understand how language development works. Depending on the cognitive building blocks ultimately involved, a better understanding of particular instances of language development may also help us better understand cognitive development more generally as well.

By using the approach I sketched above for making progress on poverty of the stimulus, we're forced to be explicit about our assumptions – assumptions about both the problem itself and what goes into a potential solution. We can examine these assumptions at a later time if we want, either keeping them or adjusting them, based on our future understanding of the learning problem and/or components of possible learning solutions. Poverty of the stimulus then becomes an accessible workspace for investigating our ideas about what's built into humans when it comes to learning language. Importantly, in my view, all this can be done without tears.

Acknowledgements

I'm incredibly grateful to both Greg Scontras and Richard Futrell for comments, discussions, and putting up with me while I worked out how to write this. I'm additionally grateful to both Cindy Fisher and Greg Hickok, for their comments and their unflagging support of a review on this topic. I also want to thank Stephen Goldinger and several anonymous reviewers for their very sensible and illuminating comments on earlier drafts of this manuscript.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Adriaans, F., & Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science*, *32*(1), 222–255.

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, *45*(2), 117–135.

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*(4), 319–342.

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

Baker, C. L. (1978). *Introduction to generative transformational syntax*. Englewood Cliffs, NJ: Prentice Hall.

Baker, C. L., & McCarthy, J. (1981). *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Belletti, A. (2017). Internal grammar and children's grammatical creativity against poor inputs. *Frontiers in Psychology*, *8*.

Bergelson, E., & Idsardi, W. (2009). Structural biases in phonology: Infant and adult evidence from artificial language learning. In *BUCLD 33: Proceedings of the 33rd annual Boston university Conference on Language Development* (pp. 85–96).

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

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

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition on child speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). New York: Wiley.

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

Carey, S. (2011). Précis of *The Origin of Concepts*. *Behavioral and Brain Sciences*, *34*(3), 113–167.

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

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

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

Chomsky, N. (1971). *Problems of knowledge and freedom*. London: Fontana.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon Books.

Chomsky, N. (1980a). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1–61.

Chomsky, N. (1980b). *Rules and Representations*. Oxford: Basil Blackwell.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). *The Architecture of Language*. New Delhi: Oxford University Press.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669.

Clark, A., & Eyraud, R. (2007). Polynomial identification in the limit of substitutable context-free languages. *Journal of Machine Learning Research*, *8*(Aug), 1725–1745.

Clark, A., & Lappin, S. (2011). *Linguistic Nativism and the Poverty of the Stimulus*. Oxford: Wiley-Blackwell.

Cole, P., Hermon, G., & Yanti. (2017). The grammar of binding in the languages of the world: A response to Reuland. *Cognition*, *168*, 380–384.

Collins, J. (2004). Faculty disputes. *Mind & Language*, *19*(5), 503–533.

Cowie, F. (1998). *What's within?: Nativism reconsidered*. Oxford University Press.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 597–612.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 522–543.

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

Davis, K. F., Parker, K. P., & Montgomery, G. L. (2004). Sleep in infants and young children: Part one: normal sleep. *Journal of Pediatric Health Care*, *18*(2), 65–71.

Dresher, E. (2003). Meno's Paradox and the Acquisition of Grammar. In S. Ploch (Ed.), *Living on the Edge: 28 Papers in Honour of Jonathan Kaye (Studies in Generative Grammar 62)* (pp. 7–27). Berlin: Mouton de Gruyter.

Dunbar, E. (2019). Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language*.

Fitz, H., & Chang, F. (2017). Meaningful questions: The acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225–250.

Fodor, J. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The Linguistic Review*, *19*, 105–145.

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.

Fuson, K. (1988). *Children's counting and concepts of number*. New York: Springer-Verlag.

Garfield, J. L. (1994). Innateness. In (pp. 366–374).

Gold, M. E. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Goodman, N. (1955). *Fact, fiction and forecast* (Vol. 74). JSTOR.

Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gordon, P. (1986). Level-ordering in lexical development. *Cognition*, *21*(2), 73–93.

Han, C.-h., Musolino, J., & Lidz, J. (2016). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences*, *113*(4), 942–947.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young american children*. Baltimore, MD: P.H. Brookes.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*(1), 4–9.

Heinz, J. (2016). Computational Theories of Learning and Developmental Psycholinguistics. In

J. Lidz, W. Synder, & J. Pater (Eds.), *The oxford handbook of developmental linguistics* (p. 633-663). Oxford University Press.

Heinz, J., & Rawski, J. (to appear). History of phonology: Learnability. In B. E. Dresher & H. van der Hulst (Eds.), *The Oxford Handbook of the History of Phonology*.

Horning, J. J. (1969). *A study of grammatical inference* (Tech. Rep.). Stanford University, Department of Computer Science.

Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein (Ed.), *Explanation in Linguistics: The Logical Problem of Language Acquisition* (pp. 9–31). London: Longman.

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.

Idsardi, W. J. (2005). Poverty of the stimulus arguments in phonology. *Unpublished manuscript, University of Delaware, Newark, DE*.

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

Jain, S., Osherson, D., Royer, J. S., Sharma, A., et al. (1999). *Systems that learn: An introduction to learning theory*. MIT press.

Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, *71*(4), 571–592.

Kam, X. N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the Richness of the Stimulus. *Cognitive Science*, *32*(4), 771–787.

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

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kimball, J. P. (1973). *The formal theory of grammar*. Prentice Hall.

King, D. (2015). Poverty of the stimulus arguments and behaviorism. *Behavior & Philosophy*, *43*.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, *24*(1), 118–137.

Kripke, S. A. (1982). *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

Lasnik, H. (1989). On certain substitutes for negative data. In R. Matthews & W. Demopoulos (Eds.), *Learnability and linguistic theory* (pp. 89–105). Dordrecht: Kluwer/Academic Press.

Lasnik, H., & Lidz, J. L. (2017). The Argument from the Poverty of the Stimulus. In *The Oxford handbook of universal grammar* (pp. 221–248).

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. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*, *19*, 151–162.

Li, M., & Vitányi, P. M. (1994). Inductive reasoning. In E. Ristad (Ed.), *Language computations.*

Philadelphia: American Mathematical Society.
Lidz, J. (2018). The explanatory power of linguistic theory. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics* (Vol. 129, pp. 225–242). Walter de Gruyter GmbH & Co KG.
Lidz, J., & Gleitman, L. R. (2004). Yes, we still need Universal Grammar. *Cognition*, *94*(1), 85–93.
Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65–B73.
Lightfoot, D. (1989). The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences*, *12*, 321–334.
Lightfoot, D. (1998). Promises, promises: General learning algorithms. *Mind & Language*, *13*(4), 582–587.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
Marcus, G. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.
Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.
Matthews, R. J. (2001). Cowie's Anti-Nativism. *Mind & Language*, *16*(2), 215–230.
Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.
McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.
McNeill, D. (1996). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The genesis of language* (pp. 15–84). Cambridge, MA: MIT Press.
Morgan, J. L. (1986). *From simple input to complex grammar*. The MIT Press.
Morgan, J. L., Bonamo, K. M., & Travis, L. L. (1995). Negative evidence on negative evidence. *Developmental Psychology*, *31*(2), 180.
Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.
Niyogi, P. (2006). *The computational nature of language learning and evolution*. MIT press Cambridge, MA.
Pearl, L. (2007). *Necessary Bias in Natural Language Learning* (Unpublished doctoral dissertation). University of Maryland, College Park, College Park, MD.
Pearl, L. (2011). When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition*, *18*(2), 87–120.
Pearl, L. (2017). Evaluation, use, and refinement of knowledge representations through acquisition modeling. *Language Acquisition*, *24*(2), 126–147.
Pearl, L. (2019). Fusion is great, and interpretable fusion could be exciting for theory generation: Response to Pater. *Language*, *95*(1), e109–e114. doi: 10.1353/lan.2019.0017
Pearl, L. (in press-a). How statistical learning can play well with Universal Grammar. In N. Allott, T. Lohndal, & G. Rey (Eds.), *Wiley-blackwell companion to chomsky.* Wiley.
Pearl, L. (in press-b). Modeling syntactic acquisition. In J. Sprouse (Ed.), *Oxford handbook of*

experimental syntax. Oxford University Press.
Pearl, L., & Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. *Language Learning and Development*, *5*(4), 235–265.
Pearl, 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., & Mis, B. (2016). The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, *92*(1), 1-30.
Pearl, L., & Phillips, L. (2018). Evaluating language acquisition models: A utility-based look at Bayesian segmentation. In A. Villavicencio & T. Poibeau (Eds.), *Language, Cognition and Computational Models* (pp. 185–224). Cambridge: Cambridge University Press.
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. (2018). *Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition.* University of California, Irvine and University of Connecticut, Storrs. Retrieved from https://ling.auf.net/lingbuzz/003913
Pearl, L., & Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development*, *3*(1), 43–72.
Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction ot bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.
Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.
Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607 - 642.
Perkins, L., Feldman, N., & Lidz, J. (2017). Learning an input filter for argument structure acquisition. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)* (pp. 11–19).
Piantadosi, S., & Kidd, C. (2016). Endogenous or exogenous? The data don't say. *Proceedings of the National Academy of Sciences*, *113*(20), E2764–E2764.
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.
Piattelli-Palmarini, M. (2017). From Zero to Fifty: Considerations on Eric Lenneberg's *Biological Foundations of Language* and Updates. *Biolinguistics*, *11*.
Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*, 217–283.
Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure.* Cambridge, MA: MIT Press.

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*, *31*, 949–953.

Pullum, G. (1996). Learnability, hyperlearning, and the poverty of the stimulus. In *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society* (pp. 498–513). Berkeley, CA: Berkeley Linguistics Society.

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50.

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

Quine, W. V. O. (1969). *Ontological relativity and other essays* (No. 1). Columbia University Press.

Ramsey, W., & Stich, S. (1991). Connectionism and three levels of nativism. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Rawski, J., & Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to Pater. *Language*, *95*(1), e125-e135.

Reali, F., & Christiansen, M. (2005). Uncovering the Richness of the Stimulus: Structure Dependence and Indirect Statistical Evidence. *Cognitive Science*, *29*, 1007–1028.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, *93*, 147–155.

Rey, G. (2014). Innate and learned: Carey, mad dog nativism, and the poverty of stimuli and analogies (yet again). *Mind & Language*, *29*(2), 109–132.

Rissanen, J., & Ristad, E. S. (1994). Language acquisition in the MDL framework. In E. S. Ristad (Ed.), *Language computations* (pp. 149–166).

Rohde, D., & Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*, 67–109.

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

Sampson, G. (1989). Language acquisition: Growth or learning? *Philosophical Papers*, *18*(3), 203–240.

Sarnecka, B. (2016). How Numbers Are Like the Earth (and Unlike Faces, Loitering, or Knitting). In Barner, David and Baron, A.S. (Ed.), *Core Knowledge and Conceptual Change.* New York: Oxford University Press.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language*, *24*(1), 139–161.

Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, *20*(60), 221–252.

Saxton, M. (2005). 'Recast' in a new light: Insights for practice from typical language studies. *Child Language Teaching and Therapy*, *21*(1), 23–38.

Saxton, M., Kulcsar, B., Marshall, G., & Rupra, M. (1998). Longer-term effects of corrective

input: An experimental approach. *Journal of Child Language*, *25*(3), 701–721.
Scholz, B., & Pullum, G. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review*, *19*, 185–223.
Scholz, B., & Pullum, G. (2006). Irrational nativist exuberance. In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science.*
Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.
Skidelsky, L. (2016). The poverty of the stimulus argument once again. *Análisis Filosófico*, *36*(2).
Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.
Stabler, E. P. (1998). Acquiring languages with movement. *Syntax*, *1*(1), 72–97.
Stich, S. P. (1978). Empiricism, innateness, and linguistic universals. *Philosophical Studies*, *33*(3), 273–286.
Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.
Valian, V. (2009). Innateness and learnability. In *Handbook of Child Language* (pp. 15–34). Cambridge University Press Cambridge, England.
Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 1134-1142.
Wexler, K. (1991). On the argument from the poverty of the stimulus. In A. Kasher (Ed.), *The Chomskyan Turn.* Oxford: Basil Blackwell.
Wharton, R. (1974). Approximate language identification. *Information and Control*, *26*(3), 236–255.
Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.
Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220–251.
Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*(2), 245-272.
Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Science*, *8*(10), 451–456.
Yang, C. (2012). Computational models of syntactic acquisition. *WIREs Cognitive Science*, *3*, 205–213.
Zenker, F., & Schwartz, B. D. (2017). Topicalization from Adjuncts in English vs. Chinese vs. Chinese-English Interlanguage. In LaMendola, Maria and Scott, Jennifer (Ed.), *Proceedings of the 41st annual Boston University Conference on Language Development.* Cascadilla Press.