Oxford Handbooks Online

The Argument from the Poverty of the Stimulus

Howard Lasnik and Jeffrey L. Lidz

The Oxford Handbook of Universal Grammar
Edited by Ian Roberts

Print Publication Date: Dec 2016
Subject: Linguistics, Morphology and Syntax, Language and Cognition
Online Publication Date: Feb 2017 DOI: 10.1093/oxfordhb/9780199573776.013.10

Abstract and Keywords

This article explores what Noam Chomsky called ‘the argument from poverty of the stimulus’: the argument that our experience far underdetermines our knowledge and hence that our biological endowment is responsible for much of the derived state. It first frames the poverty of the stimulus argument either in terms of the set of sentences allowed by the grammar (its weak generative capacity) or the set of structures generated by the grammar (its strong generative capacity). It then considers the five steps to a poverty argument and goes on to discuss the possibility that children can learn via indirect negative evidence on the basis of Bayesian learning algorithms. It also examines structure dependence, polar interrogatives, and artificial phrase structure and concludes by explaining how Universal Grammar shapes the representation of all languages and enables learners to acquire the complex system of knowledge that undergirds the ability to produce and understand novel sentences.

Keywords: Noam Chomsky, experience, knowledge, poverty of the stimulus argument, indirect negative evidence, Bayesian learning algorithms, structure dependence, polar interrogatives, artificial phrase structure, Universal Grammar

10.1 Introduction

THE problem of language acquisition has always been a central concern, perhaps the central concern, in generative grammar. The problem is that the learner, based on limited experience, projects a system that goes far beyond that experience. Already in Chomsky (1955), the founding document of the field, we find the following observation, which sets up a minimum explanatory criterion for a theory of linguistic knowledge and its acquisition:

> A speaker of a language has observed a certain limited set of utterances in his language. On the basis of this finite linguistic experience he can produce an indefinite number of new utterances which are immediately acceptable to other members of his speech community. He can also distinguish a certain set of 'grammatical' utterances, among utterances that he has never heard and might never produce. He thus projects his past linguistic experience to include certain new strings while excluding others.
>
> (Chomsky 1955/1975: p. 61 of 1975 version)

This general characterization of the situation leads to what Chomsky (1978) called 'the argument from poverty of the stimulus': the argument that our experience far underdetermines our knowledge and hence that our biological endowment is responsible for much of the derived state.

The argument from the poverty of the stimulus is essentially equivalent to the problem of induction. As Hume (1739) stated, 'even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience' (Hume 1739:139). Experience simply does not provide the basis for generalizing to the future. Chomsky's idea, (p. 222) following Descartes, is that the basis for generalization must come from the learner, not from the world.

Returning to the case of language, Chomsky went on to argue that the basis for generalization in language is in part distinct from the basis for generalization in other domains. The basic insight is that the character of linguistic representations is particular to just those representations and hence learning those representations must involve a mechanism designed to construct just those.

The argument can be framed either in terms of the set of sentences allowed by the grammar (its weak generative capacity) or the set of structures generated by the grammar (its strong generative capacity). From the perspective of the weak generative capacity of the system, the critical observation is that there is an indefinite number of unexperienced strings which the speaker of a language can produce, understand, and identify as grammatical/ungrammatical. Because any finite set of utterances is compatible with an infinite set of languages in extension (Gold 1967), and because speakers of a language agree about the grammaticality/interpretation of nearly all novel sentences, there must be some contribution from the learner to determine which language is acquired.

From the perspective of the strong generative capacity of the system, the observation is that the kinds of grammatical representations that speakers of a language construct on the basis of their experience are widely shared, but far removed from the data of experience. Any finite set of data is compatible with a wide/infinite range of characterizing functions (i.e., grammars, I-languages, etc.). That we all build the same kinds leads to the conclusion that learners are biased to construct certain kinds of grammatical representations and not others.

Said differently, the input to the child is degenerate in two senses (Chomsky 1967a). First, it is degenerate in *scope*: the input cannot provide evidence about all possible sentences (sentence–meaning pairs, sentence structures, etc.) that the child will encounter. Second, it is degenerate in *quality*: the input itself does not contain information about the kinds of representations that should be used in building a generative grammar of the language.

These notions of degeneracy should be not be confused with others. For example, Hornstein and Lightfoot (1981) note that speech to children contains speech errors, slips of the tongue, utterances produced by foreigners, etc., all of which might interfere with acquisition. While learners certainly do need to overcome this kind of degeneracy, what might be called the *noise* of the signal, the critical point about the degeneracy of the input from the perspective of the poverty of the stimulus is that the primary linguistic data (PLD) is (a) limited in scope and (b) uninformative with respect to choosing the appropriate representational vocabulary.

Chomsky (1971) used the phrase 'poverty of experience' for this same state of affairs, highlighting the degeneracy of the input relative to the character of the acquired knowledge. Interestingly, in that book the first case he discusses is one where the empiricist position that Chomsky rejects might seem to be the strongest—word learning:

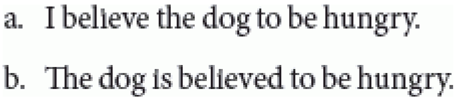
> Under normal conditions we learn words by a limited exposure to their use. Somehow, our brief and personal and limited contacts with the world suffice for us **(p. 223)** to determine what words mean. When we try to analyze any specific instance—say, such readily learned words as 'mistake,' or 'try,' or 'expect,' or 'compare,' or 'die,' or even common nouns—we find that rather rich assumptions about the world of fact and the interconnections of concepts come into play in placing the item properly in the system of language. This is by now a familiar observation, and I need not elaborate on it. But it seems to me to further dissipate the lingering appeal of an approach to acquisition of knowledge that takes empiricist assumptions as a point of departure for what are presumed to be the simplest cases.

> (Chomsky 1971:16–17)

Chomsky concludes this part of his discussion by stating that '… what little is known about the specificity and complexity of belief as compared with the poverty of experience leads one to suspect that it is at best misleading to claim that words that I understand derive their meaning from my experience' (Chomsky 1971:17). Such a claim would be misleading because the experience can only provide partial and indirect evidence about how to build a word meaning that will generalize to all relevant cases. Of course, the evidence is relevant, and in all likelihood necessary, for learning to occur, but there is no sense in which the evidence directly determines the content of the representations. As Chomsky (1965:34) notes, it is important 'to distinguish between these two functions of external data—the function of initiating or facilitating the operation of innate mechanisms and the function of determining in part the direction that learning will take.'

A particular illustration offered by Chomsky (1971) centers on what is often called the A-over-A constraint. Chomsky, informally, proposes an account of the active–passive relation (developed much further in Chomsky 1973) under which an NP following the main verb is fronted (along with other changes that won't directly concern us here). This process gives relations like those in (1).

(1)

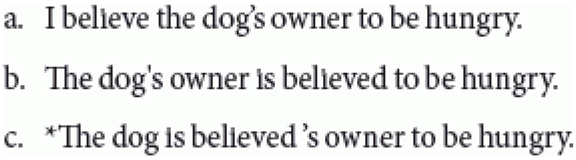


a. I believe the dog to be hungry.

b. The dog is believed to be hungry.

Chomsky observes that if the NP following *believe* is complex, an NP containing another NP, it must be the containing NP that fronts:

(2)



a. I believe the dog's owner to be hungry.

b. The dog's owner is believed to be hungry.

c. *The dog is believed 's owner to be hungry.

Chomsky's description bears directly on the 'poverty' issue:

> The instruction for forming passives was ambiguous: the ambiguity is resolved by the overriding principle that we must apply the operation to the largest noun phrase that immediately follows the verb. This, again, is a rather general property of the formal operations of syntax. There has been some fairly intensive investigation of such (p. 224) conditions on formal operations in the past decade, and although we are far from a definitive formulation, some interesting things have been learned. It seems reasonably clear that these conditions must also be part of the schematism applied by the mind in language-learning. Again, the

> conditions seem to be invariant, insofar as they are understood at all, and there is little data available to the language learner to show that they apply.
>
> (Chomsky 1971:30)

It is essential to realize how important the ambiguity-resolving property is here. *A priori*, one might expect there to be two ways of forming a passive, (2b) or (2c). Thus, evidence for the learner that (2a) is possible does not address the acquisition problem. Evidence about the existence of passives is silent with respect to the proper way of representing that construction. It should also be noted that it is not really crucial that Chomsky appeals specifically to the A-over-A condition. Any constraint that has the effect of allowing (2b) and excluding (2c) is subject to the same line of reasoning. In fact, as we will discuss in section 10.2, virtually any constraint excluding certain derivations or barring particular structure–meaning pairings is the basis for a poverty argument. Consequently, there are hundreds, if not thousands, of such arguments, either implicit or explicit, in the literature.

10.2 The Form of the Argument

Pullum and Scholz (2002) identify five steps to a poverty argument:

i) that speakers acquire some aspect of grammatical representation;
ii) that the data the child is exposed to is consistent with multiple representations;
iii) that there is data that could be defined that would distinguish the true representation from the alternatives;
iv) that that data does not exist in the primary linguistic data;
v) conclusion: the aspect of the grammatical representation acquired in (i) is not determined by experience but by properties internal to the learner.

The critical first step of the argument is in identifying the target of acquisition, whether that is a word meaning, a transformational rule, or a constraint on transformations in general. In the case of the data in (2), the target of acquisition is whatever piece of knowledge is responsible for the grammaticality of (2b) and the ungrammaticality of (2c), either a constraint on the application of a passivization transformation or, more likely, a constraint (like A-over-A) on the application of any transformational rule.

The second step is that the data is consistent with multiple representations. Assuming that the majority of passivized sentences in the PLD involve simplex subjects, then the data is equally compatible with (a) the correct grammar, (b) one which allows for

(p. 225) movement of only the most embedded NP (producing only 2c), or (c) one which allows for movement of either.

The third step involves defining what would be the relevant disambiguating evidence. In this case, the existence of (2b) is not sufficient, as this would rule out option (b), but would not rule out option (c), which allows for passivization of either NP.[1] If the existence of the actual alternative is not sufficient evidence to rule out the competitors, then what is? One possibility often raised is that explicit negative evidence would suffice. If children were simply told that sentences like (2c) were ungrammatical (characterized in a way that was sufficiently explicit, and transparent to the learner, to identify just those cases that involved moving the contained NP in an NP-over-NP structure), or if they were corrected when they produced sentences like (2c), then that evidence would distinguish the correct from the incorrect grammar.

And finally, we have steps (iv) and (v) of the argument. Since it is obvious that such explicit instruction or correction does not occur and that that is the only definable evidence that could distinguish the two grammars, it follows that the relevant constraint on the structure derives from properties internal to the learner and not from any aspect of their experience.

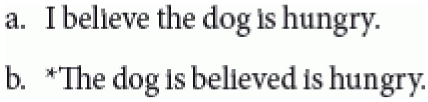
Now it is quite important to emphasize at this stage that saying that there is a constraint on possible grammars internal to children that bars them from considering the possibility that (2c) is a possible passive of (2a) is not equivalent to saying that there is no learning involved in the acquisition of English or any other language. Rather, the point is that when the learner has developed to the point of considering ways of constructing transformational rules that conform to the exposure language, there are certain hypotheses that simply will not enter into their calculations. Identifying that some construction is derived transformationally and what the surface features of that construction are (e.g., the participial morphology on the verb) must happen through some interaction between the learner and the environment (see Lidz 2010, Viau and Lidz 2011, and Lidz and Gagliardi 2015 for extensive discussion).

10.3 Further Examples

Chomsky (1971) examines one additional property of complex passive sentences of the sort exemplified in (1b), a property further developed as the Tensed-S Condition in Chomsky (1973). Corresponding to the active (1a), we found the passive (1b). But when

(p. 226) the clausal complement to the main verb is finite ('tensed') instead of infinitival, the passive becomes impossible:

(3)



a. I believe the dog is hungry.

b. *The dog is believed is hungry.

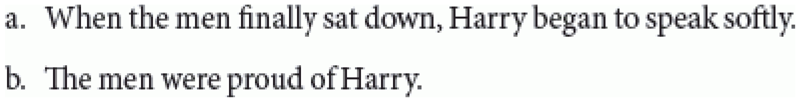
Chomsky speculates that 'nothing can be extracted from a tensed sentence.' This is a narrowing of an earlier 'clause-mate' constraint on (some) processes.[2] And again the logic of the situation is independent of the specifics of the constraint. Speakers know that (3b) is not possible, and there is no clear evidence in the input to a learner that the appropriate generalization of their experience should include the constraint that is responsible.

Chomsky explores several processes that are impeded by the boundary of a finite clause, and formulates a more general version of the constraint:

> … let us propose that no rule can involve the phrase X and the phrase Y, where Y is contained in a tensed sentence to the right of X: i.e., no rule can involve X and Y in the structure [… X … [… Y …] …], where [… Y …] is a tensed sentence.
>
> (Chomsky 1971:35)

This version of the constraint blocks not only extraction out of a finite clause, but also insertion into it, and even relating X and Y by some nonmovement rule or process. One of the most interesting instances of the latter sort was a semantic effect first investigated in Postal (1966) and Postal (1969) and called in the latter work the Inclusion Constraint. As we will immediately see, both the Inclusion Constraint (dubbed RI, for Rule of Interpretation, by Chomsky 1973) and the condition on its application provide potential poverty arguments. Postal points out a contrast between examples (4a) and (4b).

(4)



a. When the men finally sat down, Harry began to speak softly.

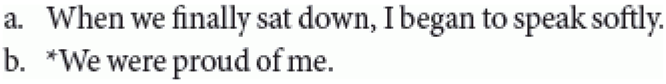
b. The men were proud of Harry.

Postal (1969) observes that

> In [(4a)] the possibility is not excluded that Harry is one of the men who sat down. In [(4b)] Harry cannot be one of the men who were proud. This is not a logical or a priori necessary fact since it is logically possible that [(4b)] could be interpreted to mean that a certain set of men were proud of one of their number, who was named Harry, and this individual Harry was proud of himself.
>
> (Postal 1969:416)

(p. 227)

Postal goes on to show that this interpretive contrast correlates with a grammaticality contrast:

(5)

a. When we finally sat down, I began to speak softly.
b. *We were proud of me.

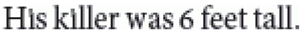
As Postal hints, and Chomsky (1971), and Chomsky (1973) claims, these are the same phenomenon. Chomsky (1971) states the generalization this way:

> … some rule of interpretation assigns the property 'strangeness' to a sentence of the form: noun phrase—verb—noun phrase—X, where the two noun phrases intersect in reference.
>
> (Chomsky 1971:38)

Since *we* and *I* must overlap in reference, (5b) cannot avoid 'strangeness.' As for (4a) and (5a), the relevant interpretive rule is limited to a local domain (roughly, the clausemate domain). After surveying a range of processes, and the locality constraints on their operation, Chomsky suggests a poverty argument:

> … there apparently are deep-seated and rather abstract principles of a very general nature that determine the form and interpretation of sentences. It is reasonable to formulate the empirical hypothesis that such principles are language universals. Quite probably the hypothesis will have to be qualified as research into the variety of languages continues. To the extent that such hypotheses are tenable, it is plausible to attribute the proposed language invariants to the innate language faculty which is, in turn, one component of the structure of mind. These are, I stress, empirical hypotheses. Alternatives are conceivable. For example, one might argue that children are specifically trained to follow the principles in question, or, more plausibly, that these principles are special cases of more general principles of mind.
>
> (Chomsky 1971:43)

Postal (1969), in the course of his discussion, makes an explicit poverty argument concerning another reference phenomenon. He observes that in (6), *his* cannot be understood as coreferential with *killer*.

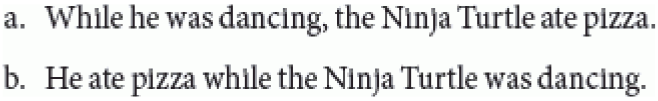
(6)

His killer was 6 feet tall.

'… [(6)] is not a clever way of saying someone killed himself and was six feet tall' (Postal 1969:421). Postal suggests an account in terms of principles of grammar, and observes that 'facts like [(6)] are of interest in relation to language learning. They are exactly the sort of thing no adult does or could teach the child directly since adults are not aware of them' (Postal 1969:422). In fact, it hardly seems likely that the child would have any evidence whatsoever for this. The same argument could have been made about the Inclusion Constraint. (p. 228)

10.4 Principle C and Indirect Negative Evidence

Indeed, Crain and McKee (1985) make such an argument about a descendent of the Inclusion Constraint, Principle C of the binding theory. Crain and McKee (1985) examined English learning preschoolers' knowledge of Principle C (Chomsky 1981a), asking whether children know that a pronoun can precede its antecedent but cannot c-command it. We refer to cases in which a pronoun precedes its antecedent as 'backwards anaphora.' In a truth value judgment experiment, children were presented with sentences like (7a) and (7b):

(7)



a. While he was dancing, the Ninja Turtle ate pizza.

b. He ate pizza while the Ninja Turtle was dancing.

In this task, participants observe a story acted out by the experimenter with toys and props. At the end of the story a puppet makes a statement about the story. The participants' task is to tell the puppet whether he was right or wrong. Crain and McKee (1985) presented children with these sentences following stories with two crucial features. First, the Ninja Turtle ate pizza while dancing. This makes the interpretation in which the pronoun (*he*) and the referring expression (*the Ninja Turtle*) are coreferentially true. Second, there was an additional salient character who did not eat pizza while the Ninja Turtle danced. This aspect of the story makes the interpretation in which the pronoun refers to a character not named in the test sentence false. Thus, if children allow coreference in these sentences, they should accept them as true, but if children disallow coreference, then they should reject them as false. The reasoning behind this manipulation is as follows. If children reject the coreference interpretation, then they must search for an additional extrasentential antecedent for the pronoun. Doing so, however, makes the sentence false. The theoretical question is whether children know that backwards anaphora is possible in sentences like (7a) but not (7b).

Crain and McKee found that, in these contexts, children as young as 3 years old accepted sentences like (7a), but overwhelmingly rejected sentences like (7b). The fact that they treated the two sentence types differently, rejecting coreference only in those sentences that violate Principle C, indicates that by 3 years of age English-learning children respect Principle C.

The observation that Principle C constrains children's interpretations raises the question of the origin of this constraint. The fact that children as young as 3 years of age behave at adult-like levels in rejecting sentences that violate Principle C is often taken as strong evidence not just for the role of c-command in children's representations, but also for the innateness of Principle C itself (Crain 1991). The reasoning behind the argument (p. 229) is that Principle C is a constraint on what is possible in language. It says that a given pairing between certain sentences and certain meanings is *im*possible. But, given that children do not have access to explicit evidence regarding what is *not* a possible form–meaning pairing in their language (see Marcus 1993 for a review), their acquisition of Principle C must be driven by internally generated constraints and not by experience alone (see Gelman and Williams 1998).

10.5 Indirect Negative Evidence and Bayesian Learning

In recent years, however, the possibility that children can learn in an indirect fashion on the basis of Bayesian learning algorithms has gained some prominence (Tenenbaum and Griffiths 2001; Regier and Gahl 2003; among others). On this view, the absence of a given form–meaning pairing might be informative about the structure of the grammar as a kind of indirect negative evidence (Chomsky 1981a).

In the context of Bayesian models, learning via indirect negative evidence is coded as the size principle (Tenenbaum and Griffiths 2001), which states roughly that smaller hypotheses are more likely than larger ones. Bayesian learners choose hypotheses by comparing the likelihood of the observed data under each hypothesis.

These models assume that learners bring to the task of learning a set of hypotheses H, each of which represents a possible explanation of the process that generated the data. In the case of language learning, this means that the class of possible grammars is defined by H, with each member h of that set representing a particular grammar. Given the observed data d, the learner's goal is to identify how probable each possible hypothesis h is, i.e., to estimate $P(h|d)$, the *posterior distribution* over hypotheses. The hypothesis with the highest posterior probability is the one that is most likely responsible for generating the observed data and hence is acquired by the learner as the correct hypothesis. Bayes' Theorem states that the posterior can be reformulated as in (8):

(8) Bayes' Theorem

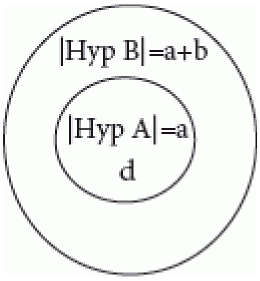
$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

The *likelihood, P(d | h)*, expresses how well the hypothesis explains the data; the *prior, P(h)*, expresses how likely the hypothesis is antecedent to any observations of data. The *evidence, P(d)*, represents the probability of the data across all hypotheses. *P(d)* functions as a normalizing factor that ensures that *P(h|d)* is a proper probability distribution, summing to 1 over all values of *h*, and is a constant that can often be safely ignored (p. 230) when comparing the relative probability of one hypothesis to another. Thus, defining a Bayesian model usually involves three steps:

i) Defining the hypothesis space: Which hypotheses does the learner consider?
ii) Defining the prior distribution over hypotheses: Which hypotheses is the learner biased towards or against?
iii) Defining the likelihood function: How does the learner's input affect the learner's beliefs about which hypothesis is correct?

Reasoning by indirect negative evidence in such models involves comparing two (or more) hypotheses. If one hypothesis produces a subset of the data that the other hypothesis produces, the likelihood of the smaller hypothesis is greater than the likelihood of the larger one. Consequently, the posterior probability of the subset grammar (i.e., the grammar generating the subset language) is greater. To see why, consider the following figure representing two grammars standing in a subset–superset relation:

(9)



In this figure, *d* represents a data point that is able to be produced by both Hypothesis A (the smaller grammar) and Hypothesis B (the larger one). For the purposes of this discussion, the size of each grammar is the set of sentences (or sentence–meaning pairs) that the grammar produces. In this case, we give the size of Hypothesis A as *a* and the size of Hypothesis B as *a+b*, where *b* is the set of sentences produced by Hypothesis B but not Hypothesis A. Hence, the likelihood that Hypothesis A produced *d* is 1/*a*; and, the likelihood that Hypothesis B produced *d* is 1/*a+b*. Since *a* is smaller than *a+b, 1/a* is larger than $1/a+b$. Consequently, the data point d is more likely to have been produced by Hypothesis A (the subset grammar) than hypothesis B (the superset grammar). Thus, as more data consistent with both grammars occurs, the posterior probability of Hypothesis A increases (even though both grammars could have produced that data).

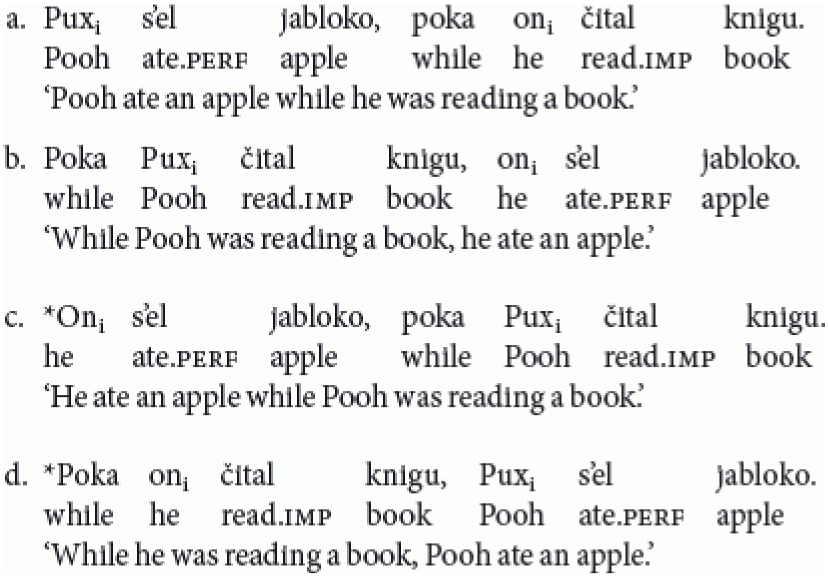
This kind of reasoning resembles the Subset Principle of Berwick (1985), which claims that if one grammar produces a language that is a subset of the language produced by another grammar, the learner should choose the subset grammar, since only this hypothesis could be disconfirmed by positive data (see also Dell 1981; Manzini and Wexler 1987; Pinker 1989). It differs from the Subset Principle in two respects. First, in Berwick's formulation the Subset Principle must be a hard-coded principle of grammar learning, whereas in the Bayesian formulation it is a general application of probability theory. Second, the Subset Principle reflects a discrete choice which can be overridden, whereas the Bayesian formulation, the preference for the subset is a probabilistic decision whose strength increases as the amount of data consistent with both grammars increases. **(p. 231)**

Returning now to the particular case under discussion, imagine that Hypothesis A is a grammar with Principle C in it and that Hypothesis B is a grammar without Principle C. Hypothesis A produces a smaller set of interpretations than Hypothesis B because any sentence to which Principle C applies has one more interpretation in Hypothesis B than it does under Hypothesis A. That is, sentences like (7b) allow either coreference between the pronoun and the name or disjoint reference under Hypothesis B but only allow disjoint reference under Hypothesis A. As learners will hear such sentences only with the disjoint reference interpretation intended, Hypothesis A becomes more likely. Thus, the lack of data that is consistent with Hypothesis B but not Hypothesis A can be treated as indirect evidence in favor of Hypothesis A over Hypothesis B.

Now, it is important to recognize that this kind of approach would have to assume that the learner has the representational capacity to formulate Principle C innately. That is, the hypothesis space H must allow for the formulation of hypotheses in hierarchical terms; and, the learner, in order to compare the likelihoods of the two hypotheses must be able to recognize c-command relations among possibly coreferential expressions and must track the relative frequency of coreference vs. disjoint reference interpretations in these environments. Without these assumptions, the indirect learner could not even begin to learn in this fashion.[3] Nonetheless, it does bring up the possibility that the existence of a constraint against certain form–meaning pairings is not by itself evidence that alternative grammars lacking that constraint are not formulable by Universal Grammar (*contra* the claims of Crain 1991).

Kazanina and Phillips (2001) addressed this issue by looking at the acquisition of backwards anaphora in Russian. Like every language, as far as we know, Russian obeys Principle C. Importantly, however, Russian exhibits a further constraint against backwards anaphora when the pronoun is contained in certain adverbial clauses but not others. These facts are illustrated in (10):

(10)



a. Pux$_i$ s'el jabloko, poka on$_i$ čital knigu.
Pooh ate.PERF apple while he read.IMP book
'Pooh ate an apple while he was reading a book.'

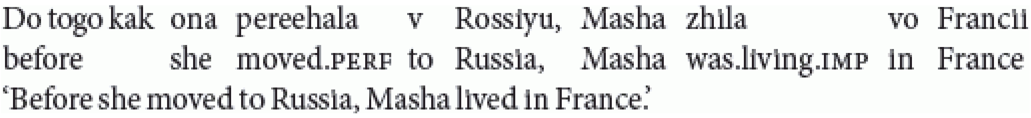
b. Poka Pux$_i$ čital knigu, on$_i$ s'el jabloko.
while Pooh read.IMP book he ate.PERF apple
'While Pooh was reading a book, he ate an apple.'

c. *On$_i$ s'el jabloko, poka Pux$_i$ čital knigu.
he ate.PERF apple while Pooh read.IMP book
'He ate an apple while Pooh was reading a book.'

d. *Poka on$_i$ čital knigu, Pux$_i$ s'el jabloko.
while he read.IMP book Pooh ate.PERF apple
'While he was reading a book, Pooh ate an apple.'

(p. 232) In (10) we see that forwards anaphora is completely free, as in English, but that backwards anaphora is more restricted than in English. In (10c), the pronoun both precedes and c-commands its antecedent and so the sentence is ruled out by Principle C. But, in (10d), the pronoun does not c-command its antecedent, but still the sentence is ungrammatical, unlike its English counterpart. The restriction on backwards anaphora appears to be tied to certain adverbial clauses, as illustrated above with the temporal adverbial *poka*, 'while.' With different temporal adverbials, backwards anaphora is possible, as shown in (11):

(11)



Do togo kak ona pereehala v Rossiyu, Masha zhila vo Francii
before she moved.PERF to Russia, Masha was.living.IMP in France
'Before she moved to Russia, Masha lived in France.'

However the restriction on backwards anaphora in *poka*-clauses is to be formulated, it is clear that it is not a universal, since it does not hold in English (see Kazanina 2005 for details).

The existence of two kinds of constraint against backwards anaphora allows us to ask about the origins of Principle C. In particular, the existence of language-particular constraints undermines the argument that Principle C is innate (in the sense that it is necessarily a part of every grammar constructable by Universal Grammar simply because it is a constraint). The existence of constraints like the Russian *poka*-constraint, therefore, makes a Bayesian approach to constraint learning more plausible. Nonetheless, Kazanina and Phillips asked whether children learning Russian demonstrate the same knowledge of Principle C as their English learning counterparts and whether they also demonstrate knowledge of the *poka*-constraint.

These researchers found a developmental dissociation between Principle C and the *poka*-constraint in Russian. While three-year-olds demonstrated adult-like knowledge for Principle C violating sentences, children at this age appeared not to know the *poka*-constraint. By five years of age, however, the Russian children had acquired the *poka*-constraint.

Because Principle C is a universal constraint but the constraint against backwards anaphora in Russian *poka*-clauses is specific to that language, Kazanina and Phillips suggest that their dissociation in acquisition derives from how they are learned. Principle C is a universal, innate, constraint on possible grammars and so does not need to be learned. Consequently, the effects of this constraint are visible in children at the earliest possible experimental observations (see also Lukyanenko, Conroy and Lidz 2014 and Sutton, Lukyanenko, and Lidz 2011). The *poka*-constraint, on the other hand, is specific to Russian and so must be learned from experience, perhaps on the basis of indirect negative evidence, as discussed earlier in this section and in section 10.4.

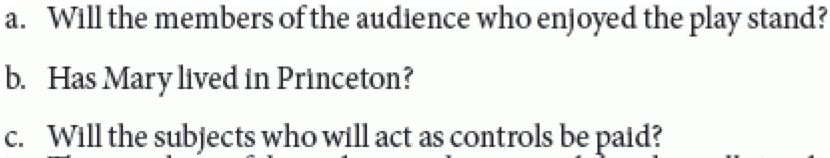
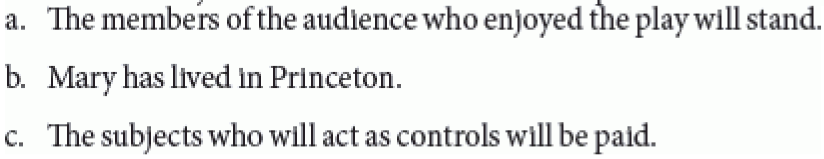
The fact that these constraints show a different learning trajectory is consistent with Principle C being innate but does not force this conclusion. In order to show that it is innate, we would need to show that the asymmetry in acquisition does not follow from asymmetries in the amount of data that learners are exposed to that provide **(p. 233)** opportunities for learning by indirect negative evidence. If the contexts for comparing a Principle C grammar against one without Principle C are more common than those for comparing a *poka*-constraint grammar against one without the *poka*-constraint, then the asymmetry in acquisition might also follow.

10.6 Structure Dependence and Polar Interrogatives

The most widely discussed poverty of the stimulus argument is based on what Chomsky (1965) called structure dependence (see also chapter 5, section 5.4). Chomsky characterized a property of transformational operations which demands that they rely on analysis into constituent structural units, and not on, say, linear sequences of words or morphemes. He argued that this property must be part of the initial state of the acquisition mechanism since, according to Chomsky, human languages invariably conform to it. The following passage provided the foundation for much further discussion and argument in subsequent years:

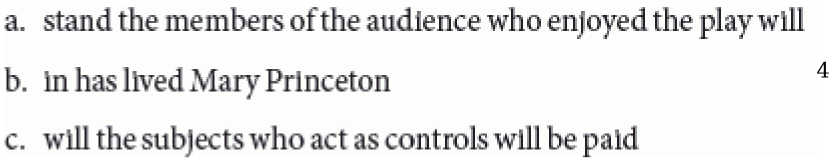
> A theory that attributes possession of certain linguistic universals to a language-acquisition system, as a property to be realized under appropriate external conditions, implies that only certain kinds of symbolic systems can be acquired and used as languages by this device. Others should be beyond its language-acquisition capacity… In principle, one might try to determine whether invented systems that fail these conditions do pose inordinately difficult problems for language learning, and do fall beyond the domain for which the language acquisition system is designed. As a concrete example, consider the fact that, according to the theory of transformational grammar, only certain kinds of formal operations on strings can appear in grammars—operations that, furthermore, have no a priori justification. For example, the permitted operations cannot be shown in any sense to be the most 'simple' or 'elementary' ones that might be invented. In fact, what might in general be considered 'elementary operations' on strings do not qualify as grammatical transformations at all, while many of the operations that do qualify are far from elementary, in any general sense. Specifically, grammatical transformations are necessarily 'structure-dependent' in that they manipulate substrings only in terms of their assignment to categories. Thus it is possible to formulate a transformation that can insert all or part of the Auxiliary Verb to the left of a Noun Phrase that precedes it, independently of what the length or internal complexity of the strings belonging to these categories may be. It is impossible, however, to formulate as a transformation such a simple operation as reflection of an arbitrary string (that is, replacement of any string a_1 … a_n, where each a_i is a single symbol, by a_n … a_1), or interchange of the $(2^n$-1)th word with the 2^nth word throughout a string of arbitrary length, or insertion of a symbol in the middle of a string of even length. … Hence, one who proposes this theory would have to predict that although a language might form interrogatives, for (p. 234) example, by interchanging the order of certain categories (as in English), it could not form interrogatives by reflection, or interchange of odd and even words, or insertion of a marker in the middle of the sentence. Many other such predictions, none of them at all obvious in any a priori sense, can be deduced from any sufficiently explicit theory of linguistic universals that is attributed to a language-acquisition device as an intrinsic property.
>
> (Chomsky 1965:55–56)

In later discussions, Chomsky often focused on the English auxiliary-fronting phenomenon as a clear and accessible example of the general idea. Unfortunately, other scholars were frequently misled by this into taking one particular aspect of the aux-fronting paradigm as the principal structure dependence claim, or, worse still, as the principal poverty of the stimulus claim. We will return to further discussion of this, after first looking at the early explicit presentation in Chomsky (1968/1972/2006):

> … grammatical transformations are invariably *structure-dependent* in the sense that they apply to a string of words [fn.: More properly, to a string of minimal linguistic units that may or may not be words.] by virtue of the organization of these words into phrases. It is easy to imagine structure-independent operations that apply to a string of elements quite independently of its abstract structure as a system of phrases. For example, the rule that forms the interrogatives of **71** from the corresponding declaratives of **72** (see note 10 [I should emphasize that when I speak of a sentence as derived by transformation from another sentence, I am speaking loosely and inaccurately. What I should say is that the structure associated with the first sentence is derived from the structure underlying the second.]) is a structure-dependent rule interchanging a noun phrase with the first element of the auxiliary.
>
> **(71)**
>
> 
>
> a. Will the members of the audience who enjoyed the play stand?
>
> b. Has Mary lived in Princeton?
>
> c. Will the subjects who will act as controls be paid?
>
> **(72)**
>
> 
>
> a. The members of the audience who enjoyed the play will stand.
>
> b. Mary has lived in Princeton.
>
> c. The subjects who will act as controls will be paid.

In contrast, consider the operation that inverts the first and last words of a sentence, or that arranges the words of a sentence in increasing length in terms of phonetic segments ('alphabetizing' in some specified way for items of the same length), or that moves the left-most occurrence of the word 'will' to the extreme left—call these O_1, O_2, and O_3, respectively. Applying O_1 to **72a**, we derive **73a**; applying O_2 to **72b**, we derive **73b**; applying O_3 to **72c**, we derive **73c**: (p. 235)

73



a. stand the members of the audience who enjoyed the play will

b. in has lived Mary Princeton

c. will the subjects who act as controls will be paid

[4]

> The operations O_1, O_2, and O_3 are structure-independent. Innumerable other operations of this sort can be specified.
>
> There is no a priori reason why human language should make use exclusively of structure-dependent operations, such as English interrogation, instead of structure-independent operations, such as O_1, O_2, and O_3. One can hardly argue that the latter are more 'complex' in some absolute sense; nor can they be shown to be more productive of ambiguity or more harmful to communicative efficiency. Yet no human language contains structure-independent operations among (or replacing) the structure-dependent grammatical transformations. The language-learner knows that the operation that gives **71** is a possible candidate for a grammar, whereas O_1, O_2, and O_3, and any operations like them, need not be considered as tentative hypotheses.
>
> If we establish the proper 'psychic distance' from such elementary and commonplace phenomena as these, we will see that they really pose some nontrivial problems for human psychology. We can speculate about the reason for the reliance on structure-dependent operations … but we must recognize that any such speculation must involve assumptions regarding human cognitive capacities that are by no means obvious or necessary. And it is difficult to avoid the conclusion that whatever its function may be, the reliance on structure-dependent operations must be predetermined for the language-learner by a restrictive initial schematism of some sort that directs his attempts to acquire linguistic competence.
>
> (Chomsky 1968/1972/2006: p. 52 of the 1968 edition)

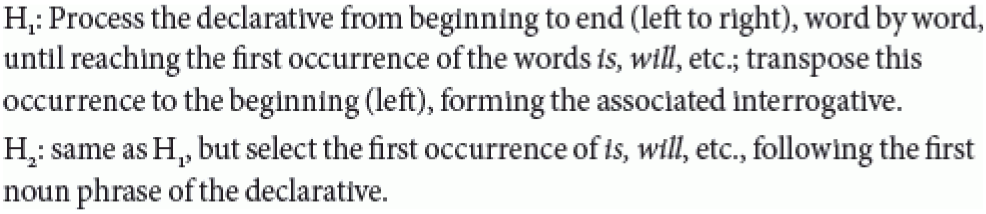
And here is Chomsky's explicit poverty argument based on auxiliary-fronting and structure dependence:

> Notice further that we have very little evidence, in our normal experience, that the structure dependent operation is the correct one. It is quite possible for a person to go through life without having heard any relevant examples that would choose between the two principles. It is, however, safe to predict that a child who has had no such evidence would unerringly apply the structure-dependent operation the first time he attempts to form the question corresponding to the assertion 'The dog that is (p. 236) in the corner is hungry.' Though children make certain kinds of errors in the course of language learning, I am sure that none make the error of forming the question 'Is the dog that in the corner is hungry?' despite the slim evidence of experience and the simplicity of the structure-independent rule. Furthermore, all known formal operations in the grammar of

> English, or of any other language, are structure-dependent. This is a very simple example of an invariant principle of language, what might be called a formal linguistic universal or a principle of universal grammar.
>
> (Chomsky 1971:27–28)

In Piattelli-Palmarini (1980), there is a very interesting interchange between Chomsky and Hilary Putnam concerning this poverty argument. Chomsky formulates two imaginable versions of auxiliary fronting (related to some we have seen earlier in this section), calling H_1 structure-independent and H_2 structure-dependent:

(12)

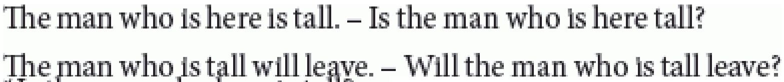


H_1: Process the declarative from beginning to end (left to right), word by word, until reaching the first occurrence of the words *is*, *will*, etc.; transpose this occurrence to the beginning (left), forming the associated interrogative.

H_2: same as H_1, but select the first occurrence of *is*, *will*, etc., following the first noun phrase of the declarative.

Chomsky observes that the following data refute H_1 but are predicted by H_2:

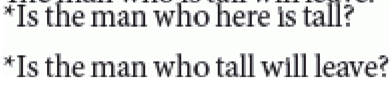
(13)



The man who is here is tall. – Is the man who is here tall?

The man who is tall will leave. – Will the man who is tall leave?

(14)



*Is the man who here is tall?

*Is the man who tall will leave?

Chomsky then asks how the child knows that H_1 is false.

> It is surely not the case that he first hits on H_1 (as a neutral scientist would) and then is forced to reject it on the basis of data such as [(13)]. No child is taught the relevant facts. Children make many errors in language learning, but none such as [(14)], prior to appropriate training or evidence. A person might go through much or all his life without ever having been exposed to relevant evidence, but he will nevertheless unerringly employ H_2, never H_1 on the first relevant occasion (assuming that he can handle the structures at all). … If humans were differently designed, they would acquire a grammar that incorporates H_1 and would be none the worse for that. In fact, it would be difficult to know, by mere passive observation of a person’s total linguistic performance, whether he was using H_1 or H_2. … Such observations suggest that it is a property of S_0 [the initial state of the language faculty] that rules (or rules of some specific category, identifiable on quite general grounds by some genetically determined mechanism) are structure-dependent. The child need not consider H_1; it is ruled out by properties of his initial mental state, S_0.
>
> (Piattelli-Palmarini 1980:40)

Putnam, in rejecting Chomsky's conclusion, responds, in part 'H_1 has never been "put forth" by anyone, nor would any sane person put it forth …' (Piattelli-Palmarini 1980: 287).

But Chomsky has a powerful counter-response:

> Putnam considers my two hypotheses H_1 and H_2, advanced to explain the formation of yes-or-no questions in English. He observes that the structure-independent rule H_1 would not be put forth by any 'sane person,' which is quite true, but merely constitutes part of the problem to be solved. The question is: Why? The answer that I suggest is that the general principles of transformational grammar belong to $S_0{}^L$ as part of a schematism that characterizes 'possible human languages'.
>
> (Piattelli-Palmarini 1980:311)

Perfors et al. (2006) offer a related argument against this particular poverty argument. Perfors et al. summarize the situation and their argument as follows:

> The Poverty of the Stimulus (PoS) argument holds that children do not receive enough evidence to infer the existence of core aspects of language, such as the dependence of linguistic rules on hierarchical phrase structure. We reevaluate one version of this argument with a Bayesian model of grammar induction, and show that a rational learner without any initial language-specific biases could learn this dependency given typical child-directed input.

Curiously, they wind up doing no such thing, nor do they actually even attempt any such thing. Instead, as pointed out by Lasnik and Uriagereka (2007), they wind up repeating one error that Putnam made. Their system, when presented with a context-free language, learns a context-free language. All of the grammars presented as targets of the learning are particular phrase structure grammars. Thus, as Chomsky observed in discussing a point Putnam made, to talk of structure dependence is to make a category mistake. Structure dependence (and structure independence) are properties of transformations:

> Note that both of my hypotheses, H_1 and H_2, present rules that apply to a sentence, deforming its internal structure in some way (to be precise, the rules apply to the abstract structures underlying sentences, but we may put this refinement aside). Both the structure-independent rule H_1 and the structure-dependent rule H_2 make use of the concepts 'sentence,' 'word,' 'first,' and others; they differ in that H_2 requires in addition an analysis of the sentence into abstract

> phrases. A rule that does not modify the internal structure of a sentence is neither structure-dependent nor structure-independent. For example, a phrase structure rule, part of a phrase structure grammar in the technical sense of the term, is neither structure-dependent nor structure-independent.
>
> (Piattelli-Palmarini 1980:315)

Pullum and Scholz (2002), in their extensive discussion of poverty of stimulus arguments, call the auxiliary-fronting argument discussed earlier in this section 'the apparently strongest case of alleged learning from crucially inadequate evidence discussed in (p. 238) the literature, and certainly the most celebrated.' But they reject Chomsky's claim of unavailability of relevant evidence for the learner, citing especially Sampson (1989). They say that it has not at all been established that children are not exposed to interrogative sentences where the subject contains an auxiliary verb, and where that auxiliary is not what has been fronted, and they present several occurring examples found in various searches. They state that 'Chomsky's assertion that you can go over a vast amount of data of experience without ever finding such a case is unfounded hyperbole. We have found relevant cases in every corpus we have looked in' (Pullum and Scholz 2002:44). Pullum and Scholz (2002) conclude:

> Our preliminary investigations suggest the percentage of relevant cases is not lower than 1 percent of the interrogatives in a typical corpus. By these numbers, even a welfare child would be likely to hear about 7,500 questions that crucially falsify the structure-independent auxiliary-fronting generalization, before reaching the age of 3. But assume we are wrong by a whole order of magnitude on this, so there are really only 750. Would exposure to 750 crucial examples falsifying the structure-independent generalization not be enough to support data-driven learning?
>
> (Pullum and Scholz 2002:45)

This argument misses two essential points. First, empirically, it is simply not correct. The corpora examined by Pullum and Scholz were all newspaper corpora which are not representative of speech to children. When corpora of child-directed speech were examined (Legate and Yang 2002), the relevant disambiguating data occurred at rates of less than 0.07% of all utterances, a number substantially (nearly two orders of magnitude) less than that of other constructions which are acquired at age 3.

Second, as Lasnik and Uriagereka (2002) observe, essentially following Freidin (1991), this number of relevant sentences is largely beside the point. That is, while one might read Chomsky as implying that the availability of the relevant grammatical questions in the child's data would undermine the poverty argument, that is not actually the case. Suppose the child is presented with data like (15).

(15) Is the dog that is in the corner hungry?

While this does seem to indicate the incorrectness of a rule demanding that the first auxiliary must be fronted to form an interrogative (16a) and the relative superiority of (16b), it provides no basis for excluding, say, a rule like (17a) or one like (17b).

(16)
a. Front the first auxiliary.
b. Front the auxiliary in the matrix Infl.

(17)
a. Front any auxiliary.
b. Front any finite auxiliary.
(p. 239) c. Front the last auxiliary.
d. Front either the first or last auxiliary.

As Lasnik and Uriagereka (2002) observe,[5]

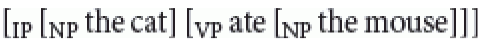
> … P&S are missing Freidin's point: [(15)] is not direct evidence for anything with the effects of hypothesis [(16b)]. At best, [(15)] is evidence for something with the logical structure of [(16a)] or X, but certainly not for anything having to do with [(16b)] ….
>
> (Lasnik and Uriagereka 2002:149)

In sum, while there have been many attempts to undermine the argument from the poverty of the stimulus through analysis of auxiliary fronting rules, these generally miss the point in several ways. First, while this argument is a prominent one, it does not represent the strongest case (indeed, when Chomsky 1975:30 discusses this case, he calls it 'the simplest one that is not entirely trivial'); nor does the entire argument depend on the validity of this one instantiation. Second, the argument is about the relation between sentences and their structures. It is not about the grammaticality of particular yes–no questions; it is about the relation between these questions and the corresponding declaratives. Third, any poverty argument is based on a comparison of hypotheses. But, as noted above, any finite dataset is compatible with a potentially infinite number of grammars and so in essence a complete response to a poverty argument would have to compare perhaps an infinite number of hypotheses, and surely not just two. So, even if there is evidence in speech to children that favors the correct analysis over one explicit alternative, this does not decide the question against the poverty argument. In order to do so, it would be necessary to show that the available evidence would also rule out all other possible reformulations of the rule.

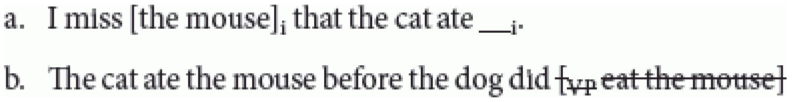
10.7 Structure Dependence and Statistical Analysis in Artificial Phrase Structure

In research on animal learning, one of the most informative approaches to separating the contribution of the learner from the contribution of the environment is to withhold certain kinds of experiences from the learner and to see what gets acquired. If what is (p. 240) ultimately acquired is not affected by removing certain kinds of experiences from the organism's normal life, then it follows that those experiences played no role in shaping the knowledge attained. While experiments like these are unethical to conduct with human children (but see Feldman, Goldin-Meadow, and Gleitman 1978, Landau and Gleitman 1985, and Senghas and Coppola 2001 for some natural variants), we can recreate this kind of selective rearing experiment in the laboratory through the use of artificial languages.

Takahashi (2009) conducted just such an experiment examining the structure dependence of transformational rules. As is well known, constituent structure representations provide explanations for (at least) three kinds of facts. First, constituents provide the units of interpretation. Second, the fact that each constituent comes from a category of similar constituents (e.g., NP, VP, etc.) makes it such that a single constituent type may be used multiple times within a sentence, as in (18):

(18) 
$[_{IP}$ $[_{NP}$ the cat] $[_{VP}$ ate $[_{NP}$ the mouse]]]

Third, constituents provide the targets for grammatical operations such as movement and deletion:

(19) 
a. I miss [the mouse]$_i$ that the cat ate __$_i$.
b. The cat ate the mouse before the dog did ~~[$_{VP}$ eat the mouse]~~

It is the third property which corresponds to the structure dependence of transformational rules.

Thompson and Newport (2007) make a very interesting observation about phrase structure and its acquisition: because the rules of grammar that delete and rearrange constituents make reference to structure, these rules leave a kind of statistical signature of the structure in the surface form of the language. The continued co-occurrence of certain categories and their consistent appearance and disappearance together ensures that the co-occurrence likelihood of elements from within a constituent is higher than the co-occurrence likelihood of elements from across constituent boundaries.

Thompson and Newport (2007) go on to argue that this statistical footprint could be used by learners in the acquisition of phrase structure. And, they show that adult learners are able to use this statistical footprint in assigning constituent structure to an artificial language. But showing that learners are sensitive to the statistical features of the environment does not yet provide information about the acquired representations. It is impressive that learners learned about the constituent structure of an artificial language given only statistical information about that structure. But this demonstration remains silent about the character of the acquired representations and the inferences that these representations license. (p. 241)

In order to determine whether the acquired representations have properties that derive from the structure of the learner, it is important to identify their deductive consequences. Do learners know things about constituent structure (even if this structure is acquired using statistical features of the environment) that are not evident in the statistics themselves? More narrowly, does the structure-dependence of grammatical rules follow from the fact that a grammar leaves a statistical signature on the input or does it follow from structure imposed by the learner on that statistical signature?

In order to answer this question, Takahashi and Lidz (2008) constructed a miniature artificial grammar containing internally nested constituents. In addition, the grammar contained rules which allowed for the repetition of constituents of a certain type, the movement of certain constituents and substitution of certain constituents by pro-forms. They then created a corpus of sentences from this language in which these rules applied often enough to provide statistical evidence for the constituent boundaries. In other words, the language provided statistical cues to the internal structure of the sentences.

Their first question, using this artificial language, was whether adults and infants could acquire constituent structure using only statistical information. The language was presented in contexts that did not provide any referential information, so that no meaning could be assigned to any of the words. And, there was no prosodic or phonological information of any kind that could serve as a cue to the phrase structure. So, to the extent that learners could acquire the phrase structure, they would have to do so through the statistical features of the exposure. In order to test whether the learners acquired the phrase structure, they asked whether the learners could distinguish novel sentences containing either moved constituents or moved nonconstituents. Since only constituents can move in natural languages, they reasoned that if learners could distinguish moved constituents from moved nonconstituents, it must be because they had learned the constituent structure of the artificial language. They found that both adults, after 36 minutes of exposure, and 18-month-old infants, after only two minutes of exposure, were able to do so (Takahashi and Lidz 2008; Takahashi 2009). Thus, the statistical footprint of constituent structure is detectable by learners and is usable in the acquisition of phrase structure.

Now, the exposure provided to the learners in this experiment included sentences containing movement. Although the particular sentences tested were novel, they exhibited structures that had been evident during the initial exposure to the language. Takahashi and Lidz thus went on to ask whether the inference that only constituents can move derives from the learner's exposure to movement rules which apply only to constituents or whether this inference derives from the child's antecedent knowledge about the nature of movement rules in natural language.

To ask this question, they created a new corpus of sentences from their artificial language. In this novel corpus were included sentences in which (a) certain constituents were repeated in a sentence, (b) certain constituents were optionally absent from a (p. 242) sentence, and (c) certain constituents were replaced by pro-forms. This combination of operations created a statistical signature of the phrase structure of the language such that it was possible to identify the constituent boundaries in the language. However, in this input corpus they included no examples of movement. This made it possible for them to identify the locus of the learner's knowledge that only constituents can move. If this knowledge derives from the learner's experience in seeing movement rules, then we would expect learners to be unable to distinguish moved constituents from moved nonconstituents. On the other hand, if the learner brings knowledge about what kinds of movement operations are possible in natural language to the learning task, then we would expect learners to correctly distinguish moved constituents from moved nonconstituents.

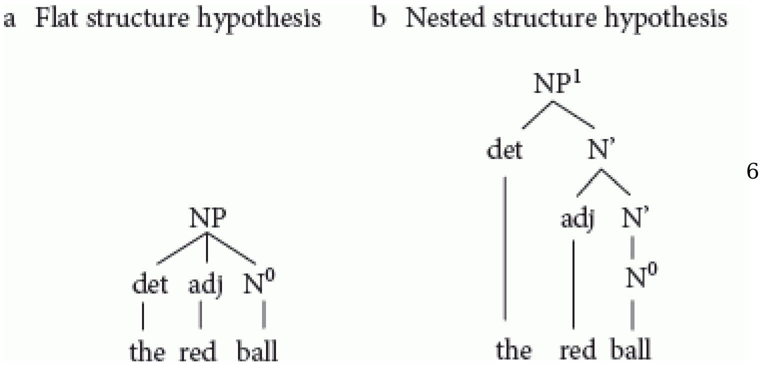
They found that both adults and 18-month-old infants displayed knowledge of the constraint that only constituents can move, even when their exposure to the artificial language contained no instances of movement whatsoever. Thus, we can conclude that some of what is acquired on the basis of statistical information is not itself reflected in the statistics. Since the learners in this experiment had seen no examples of movement, their knowledge of the constraint that only constituents can move could not have come from the exposure to language but rather must have come from the learners themselves. More generally, while learners may get evidence about constituency from movement operations, the possibility that only constituents can move comes not from those experiences, but from constraints on possible movement rules that are part of the architecture of the language faculty. In this case, the learners knew about the structure dependence of transformational rules even without having any experience of transformations whatsoever.

In sum, identifying the constituency of a language has consequences for novel sentences with structures never before encountered. These deductive consequences reveal the structure of the learner over and above any role of distributional learning. Distributional learning therefore functions as part of a process of mapping strings onto the grammar that generated them (perhaps in the Bayesian fashion described in section 10.5). But some properties of the identified grammar are contributed by the learner's antecedent knowledge of the class of possible grammars.

10.8 *One*-Substitution and the Trouble with Indirect Negative Evidence

One further well-known illustration of the poverty of the stimulus argument concerns the hierarchical structure of NP and the anaphoric uses of *one* (Baker 1978; Hornstein and Lightfoot 1981; Lightfoot 1982; Hamburger and Crain 1984). Consider two hypotheses for the structure of NP, given in (20). Both would, in principle, be possible analyses of strings containing a determiner, adjective, and noun. (p. 243)

(20)



[6]

We know, on the basis of anaphoric substitution, that for adults (20b) is the correct representation (Baker 1978). In (21), the element *one* refers anaphorically to the constituent [*red ball*].

(21) I'll play with this red ball and you can play with that one.

Since anaphoric elements substitute only for constituents and since it is only under the nested structure hypothesis that the string *red ball* is represented as a constituent (i.e., with a single node containing only that string), it follows that (20b) is the correct structure.[7]

Now, although we know that the nested structure hypothesis reflects the correct adult grammar, and that *one* is anaphoric to N′ and not N^0, how children acquire this knowledge is more mysterious. Consider the following learning problem (Hornstein and Lightfoot 1981). Suppose that a learner is exposed to small discourses like (21) in which *one* is anaphoric to some previously mentioned discourse entity and that the learner has recognized that *one* is anaphoric. In order to understand this use of *one*, the learner must know that it is anaphoric to the phrasal category N′, which is possible only under the nested structure hypothesis. However, the data to support this hypothesis is not available to the learner for the following reason. For positive assertions like (21), every situation that makes *one* = $[_{N'}$ *red ball*] true also makes *one* = $[_{N^0}$ *ball*] true. Thus, if the learner (p. 244) had come to the hypothesis that *one* is anaphoric to N^0 and not N′, evidence that this is wrong would be extremely difficult to come by.

Evidence that could support the N′ hypothesis over the N^0 hypothesis comes from negative sentences like (22) in contexts in which Max has a blue ball.

(22) Chris has a red ball but Max doesn't have one.

In such a situation, the learner who posited that *one* was anaphoric to the N^0 *ball*, would have to conclude that he had built the wrong grammar (or that the speaker was lying) and thus be led to change the hypothesis. Now, in order for learners to build the correct grammar, such situations would have to be common enough for them to show up at levels distinguishable from noise in every child's linguistic environment. Since such situations are not likely to be so common, we conclude that neither the flat structure hypothesis nor the hypothesis that *one* is anaphoric to N^0 could be part of the hypothesis space of the learner. If they were, then some learners might never come upon the evidence disconfirming those hypotheses and would therefore acquire the wrong grammar. Since there is no evidence that English speakers actually do have that grammar, it simply must never be considered.

The logic of the argument is unquestionable; however, it is based on the crucial assumption that the evidence that unambiguously supports the nested structure hypothesis does not occur often enough to impact learning. In addition, because it is an argument based on what adults know about their language, it is missing the important step of showing that at the earliest stages of syntactic acquisition, children do know that *one* is anaphoric to the phrasal category N′.

Hamburger and Crain (1984), in response to Matthei (1982), addressed the latter issue by testing 4- to 6-year-old children and found that they do represent the NP with a nested structure and that they also know that *one* is anaphoric to the phrasal category N′.

However compelling, evidence based on preschool-aged children cannot reveal the initial state of the learner or the mechanisms responsible for the acquisition of this syntactic structure. This type of evidence leaves open the possibility that learners begin the process of acquisition with a flat structure grammar, discover somehow that this structure is wrong, and subsequently arrive at the nested structure grammar to better capture the input. Lidz, Waxman, and Freedman (2003) addressed this concern by testing infants at the earliest stages of syntactic acquisition, since these infants are more likely to reveal the initial state of the learning mechanism.

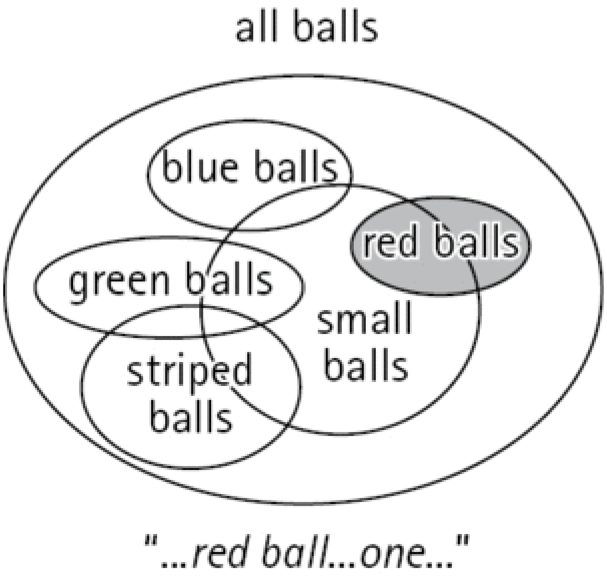
Lidz, Waxman, and Freedman (2003) tested 18-month-old infants in a preferential looking study (Hirsh-Pasek and Golinkoff 1996) in order to determine whether children represent strings like *the red ball* as containing hierarchical structure. Each infant participated in four trials, each consisting of two phases. During the familiarization phase, an image of a single object (e.g., a red ball) was presented three times, appearing in alternating fashion on either the left or right side of the television monitor. Each presentation was accompanied by a recorded voice that named the object with a phrase consisting of a determiner, adjective, and noun (e.g., 'Look! A red ball'). During the test phase, two (p. 245) new objects appeared simultaneously on opposite sides of the television monitor (e.g., a red ball and a blue ball). Both objects were from the same category as the familiarization object, but only one was the same color. Infants were randomly assigned to one of two conditions which differed only in the linguistic stimulus. In the control condition, subjects heard a neutral phrase ('Now look. What do you see now?'). In the anaphoric condition, subjects heard a phrase containing the anaphoric expression *one* ('Now look. Do you see another one?').

The assumption guiding the preferential looking method is that infants prefer to look at an image that matches the linguistic stimulus, if one is available (Spelke 1979). Given this methodological assumption, the predictions were as follows. In the control condition, where the linguistic stimulus does not favor one image over the other, infants were expected to prefer the novel image (the blue ball), as compared to the now-familiar image (the red ball). In the anaphoric condition, infants' performance should reveal their representation of the NP. Here, there were two possible outcomes. If infants represent *one* as anaphoric to the category N^0, then both images would be potential referents of the noun (*ball*). In this case, the linguistic stimulus is uninformative with regard to the test images, and so infants should reveal the same pattern of performance as in the control condition. However, if infants interpret *one* as anaphoric to N′, then they should reveal a preference for the (only) image that is picked out by N′ (the red ball).

Subjects in the control condition revealed the predicted preference for the novel image, devoting more attention to it than to the familiar image. This preference was reversed in the anaphoric condition, where infants devoted more attention to the familiar than to the novel image. This constitutes significant evidence for the hypothesis that by 18 months, infants interpret *one* as anaphoric to the category N′, despite the fact that nearly any instance of anaphoric *one* in the input is consistent with both the *one* = N^0 hypothesis and the *one* = N′ hypothesis.

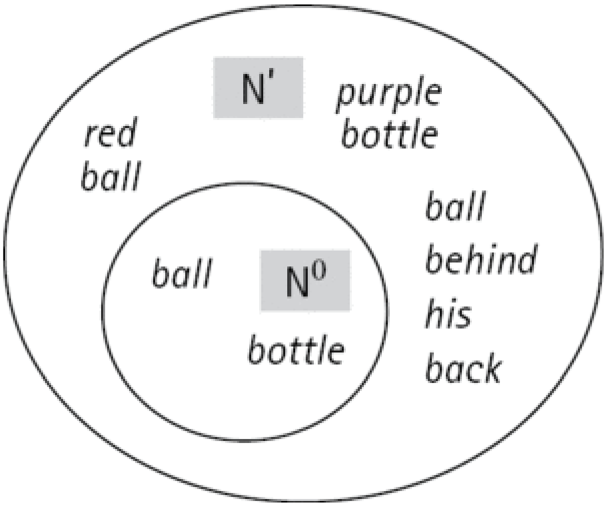
In addition to this behavioral data, Lidz, Waxman, and Freedman (2003) also conducted a corpus study aimed at determining whether it was really true that the evidence that would distinguish the *one* = N^0 hypothesis from the *one* = N′ hypothesis did not occur in speech to children. They found that in a corpus of child-directed speech, unambiguous data that distinguishes the two hypotheses occurred at a rate of 0.2%, roughly ½ the rate of occurrence of ungrammatical sentences. Because learners should treat ungrammatical sentences as noise, they should not treat any data that occurs less often than that as useful. If they did, they would also learn that the ungrammatical sentences were grammatical, contrary to fact. Thus, the corpus analysis supports the step of the poverty argument concerning the unavailability of informative data.

Now, Regier and Gahl (2003) noted that the sparseness of unambiguous data could be overcome if learners could use ambiguous data in conjunction with the predictions of alternative hypotheses to evaluate them. They reasoned that if *one* were anaphoric to N^0, then sentences containing *one* should allow a wider range of interpretations than they would if *one* were anaphoric to N′. As noted, since every red ball is also a ball, but there are balls that are not red, then a grammar in which *one* was anaphoric to N^0 would allow interpretations that excluded the property denoted by the adjective in the antecedent (p. 246) NP, as depicted in Figure 10.1. Hence, if the property mentioned by the adjective in the antecedent always holds of the referent of the anaphoric NP containing *one*, this would be a conspicuous coincidence under the *one* = N^0 hypothesis but is fully expected on the *one* = N′ hypothesis. Thus, a general learning mechanism that disfavors coincidences, like the Bayesian learner discussed in section 10.5, would come to prefer the N′ hypothesis over the N^0 hypothesis.



Click to view larger

Figure 10.1 The red balls are a subset of the balls. If *one* is anaphoric to *ball*, it would be mysterious if all of the referents of the NPs containing *one* were red balls. Learners should thus conclude that *one* is anaphoric to *red ball*.



Click to view larger

Figure 10.2 Syntactic predictions of the alternative hypotheses. A learner biased to choose the smallest subset consistent with the data should favor the *one* = N^0 hypothesis.

This idea shows the potential informativeness of indirect negative evidence. If learners are comparing the predictions of alternative hypotheses, then even if the data is ambiguous it might be more or less likely under each of the alternatives, providing some reason to favor one hypothesis over the other.

However, in the particular case at hand, this kind of learning mechanism has been shown not to work, primarily because it concerns itself only with the semantic predictions of each hypothesis. Because the semantic predictions of the *one* = N′ hypothesis are a subset of the *one* = N^0 hypothesis, learners should be biased towards the former. However, as shown in Figure 10.2, the two hypotheses also make predictions about the syntactic properties of their antecedents. In particular, the set of strings covered by the *one* = N′ hypothesis is larger than the set of strings covered by the one = N^0 hypothesis.

Pearl and Lidz (2009) show, with a range of computational learning simulations, that considering both the semantic and syntactic consequences of the two hypotheses leads to a learner that favors the *one* = N^0 hypothesis, largely due to the 95% of utterances (p. 247) containing *one* that do not have an adjective in the antecedent. In those cases, the two hypotheses make the same semantic predictions, but the *one* = N^0 hypothesis is favored because it is the smallest hypothesis consistent with the data.

The moral of this story is that when we consider the possibility of learning by indirect negative evidence, we must be sure to consider all of the syntactic and semantic predictions of each hypothesis under consideration. When we do so, we find that learning by indirect negative evidence may not be as effective as it seems when only considering a portion of the relevant data. Pearl and Lidz (2009) go on to show that the problem can be overcome by ignoring the vast majority of the data that the learner is exposed to (namely all of the data in which the antecedent NP contains no modifiers). However, because this discounting of certain data is not motivated by general learning principles, it follows that a general purpose learning mechanism (e.g., one that uses indirect negative evidence) can function only with domain-specific constraints on either the hypotheses under consideration or the set of data that the learner takes to be informative.

10.9 Conclusions

The argument from the poverty of the stimulus remains one of the foundational cornerstones of generative linguistics. Because a grammatical theory must contribute to our understanding of how children come to have grammars (the ‘explanatory adequacy’ of Chomsky 1965), questions of learnability are intimately tied up with the proper formulation of the theory of syntax (see also chapter 11). Because of this central place in the theory, it is important to understand the argument for what it is. Learning a language requires internalizing a grammar (i.e., a system for representing sentences). The internalized grammars have properties which do not follow from facts about the distributions of words and their contexts of use. Nor do these properties follow from independently understood features of cognition. Consequently, the way to force grammars to have these properties as opposed to others is to impose some constraints on the hypotheses that learners consider as to how to organize their experience into a system of grammatical knowledge.

The point of these arguments is not that there is no way of organizing or representing experience to get the facts to come out right. Rather, there must be something inside the learner which leads to that particular way of organizing experience. The puzzle is in defining what forces learners to organize their experience in a way that makes the right divisions. This organizing structure is what we typically refer to as Universal Grammar: the innate knowledge of language that (a) shapes the representation of all languages and (b) makes it possible for learners to acquire the complex system of knowledge that undergirds the ability to produce and understand novel sentences.

That said, it just as important to note that claims about the poverty of the stimulus and the existence of constraints on possible grammars do not eliminate the environment as a critical causal factor in the acquisition of a particular grammar. A complete theory of (p. 248) language development must show how the particular constraints of Universal Grammar (e.g., the necessary structure dependence of grammatical rules) makes it possible for learners to leverage their experience in the identification of a grammar for the language they are exposed to (see chapter 12). Positing a universal grammar constrains the learning mechanism to be a selective one, rather than instructive, in the sense that learning involves using the data in the exposure language to find the best-fitting grammar of that language, subject to the constraints imposed by Universal Grammar (Fodor 1966; Pinker 1979; Lightfoot 1982; and chapter 11). Even if learners come fully loaded with innate knowledge about the range of abstract structures that are possibly utilized in language, they must still use evidence from the surface form of language to identify which particular abstract structures underlie any given sentence in the language to which they are exposed (Fodor 1966; Pinker 1979; Tomasello 2000; Viau and Lidz 2011; Lidz and Gagliardi 2015). But the fact that the input to children plays a causal role in the construction of a grammar does not undermine arguments from the poverty of the stimulus. Rather, the rich inferences that children make on the basis of partial and fragmentary data still provide strong arguments for the poverty of the stimulus and the contribution of innate principles of grammar in the acquisition of a language.

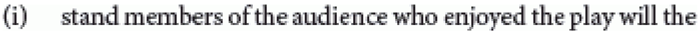
Notes:

([1]) One might argue that the more restrictive option (a) is to be preferred over option (c) through the use of indirect negative evidence (Chomsky 1981a). Given that option (a) predicts only one type of passive and option (c) predicts two types, and only one type occurs in the PLD, option (a) is simply more likely. We return to the difficulty of identifying the scope of such arguments in section 10.5.

([2]) This locality condition is incorporated into the notion governing category of Chomsky (1981).

([3]) More generally, as Lasnik (1989) notes, learning by indirect negative evidence should be possible only to the extent that the learner has a constrained representational vocabulary over which to build and compare hypotheses.

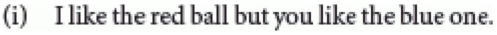
([4]) Chomsky's point is clear, but this example seems not to be the correct one, as it evidently involves both the correct and the incorrect transformation, the latter following the former. More immediately relevant would be:


(i) stand members of the audience who enjoyed the play will the

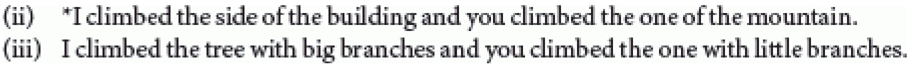
([5]) Pullum and Scholz (2002) are aware of Freidin's argument, but choose to overlook its deep significance, saying: 'We ignore this interesting point (which is due to Freidin 1991), because our concern here is to assess whether the inaccessibility claim is true.'

([6]) The argument goes through in exactly the same way if we change the constituent labels of NP to DP and N′ to NP, as in Abney (1987).

([7]) Note also that *one* can be anaphoric to certain strings containing only a single noun, as in (i):


(i) I like the red ball but you like the blue one.

Here, however, *one* is still anaphoric to N′. This can be seen by examining the difference between cases where the NP contains an argument and those in which it contains an adjunct:


(ii) *I climbed the side of the building and you climbed the one of the mountain.
(iii) I climbed the tree with big branches and you climbed the one with little branches.

Because *one* cannot be anaphoric to a complement-taking noun (ii) without including the complement, it follows that *one* cannot be anaphoric to N^0 (Lee (20)) and that cases in which it apparently is, such as (iii), represent cases of the head noun being contained in a larger N′ constituent.

Howard Lasnik

Howard Lasnik is Distinguished University Professor in the Department of Linguistics at the University of Maryland. He is one of the world's leading theoretical linguists and has produced influential and important work in areas such as syntactic theory, logical form, and learnability. His publications include *Essays on Anaphora* (1989), *Minimalist Syntax* (Blackwell 1999), and *Minimalist Investigations in Linguistic Theory* (2003).

Jeffrey L. Lidz

Jeffrey Lidz is Professor of Linguistics at the University of Maryland. His work examines the relation between grammatical theory, on-line understanding mechanisms and learning. Bringing data to bear from languages as diverse as English, French, Korean, Kannada, and Tsez, he has published papers on quantification, argument structure, morphosyntax, A-bar movement, and reference relations, Lidz is currently editor in chief of *Language Acquisition: A Journal of Developmental Linguistics*.