When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity.

Lisa Pearl

Department of Cognitive Sciences

3151 Social Science Plaza A

University of California

Irvine, CA 92697

lpearl@uci.edu

Jeffrey Lidz

Linguistics Department

1401 Marie Mount Hall

University of Maryland

College Park, MD 20742

jlidz@umd.edu

Abstract

We identify three components of any learning theory: the representations, the filters on data intake, and the knowledge updating procedure. With these in mind, we model the acquisition of the English anaphoric pronoun *one*. We show first that an unconstrained domain-general updating procedure fails to learn anaphoric *one*. However, when this procedure is paired with a domain-specific filter on data intake, then it succeeds. Thus, we find that the strengths of domain-general learning regimes are amplified when paired with domain-specific representations and domain-specific filters on data intake.

Keywords: learnability, representations, data intake filters, probabilistic updating, Bayesian learning, anaphoric *one*, domain-specific learning, domain-general learning

1. Introduction

Vast quantities of ink and hard feelings have been spilt and spent on the nature of learning in humans and other animals. Are there domain-specific learning mechanisms or is learning the same across all domains? One of the most frequent battlegrounds in this debate is the case of language learning. Is there a domain-specific language acquisition device or does language acquisition rely solely on domain-general learning mechanisms? We believe that the phrase "domain-specific learning" can be and has been interpreted in several distinct ways, leading to the illusion of incompatibility with domain-general learning. However, by examining these interpretations, we believe we can create a synthesis of these two viewpoints in a way that amplifies the strengths of both approaches.

There are three pieces to any learning theory. First, learners must have a way of representing the data to be learned from. In the domain of language learning, these would be the linguistic representations, such as phonemes, morphemes and phrase structure trees. If learners come equipped with a space of possible linguistic representations, then we have a domain-specific representational format in our learning theory. On the other hand, if learners represent the information in the linguistic signal in terms of cooccurrence probabilities between properties of the acoustic signal, then we do not have a domain-specific representational format in our learning theory. Of course, it is possible that domain-specific representations can be constructed out of the domain-general representations of the input. In this case, then, we do not have domain-specific representations as part of the learning theory. Instead, we would have domain-specific representations as the output of learning.

Second, learners must decide which data to learn from. For language learning, one might propose that only some data are useable by the learner. For example, Lightfoot (1991) proposes that main clause data are privileged for the learner. Data in embedded clauses are initially ignored for the purposes of grammar learning. Because such filters are defined over the linguistic representations, they instantiate domain-specific filters on data intake. On the other hand, if what looks like a constraint against embedded clause data were in fact due to learners only having a finite amount of working memory, causing them to use, say, only the first 4 words of an utterance, this would be a domain-general filter. As with the representations, it is possible that a domain-specific filter is the output of a domain-general procedure. For instance, if the learner has domain-specific representations such as main clauses and embedded clauses, a finite working memory constraint might lead to using the first structural "chunk" available – i.e. the main clause data. Of course, it is also possible that learners treat all of their linguistic information sources equally and that there is no constraint from the learning mechanism on what data are relevant for learning.

Third, learners must have a way of updating their knowledge on the basis of the selected data. Any learning algorithm that is used only for language learning (e.g. Fodor, 1998) would count as an example of a domain-specific learning procedure. On the other hand, if the same learning algorithm is used across different domains (e.g., Bayesian learning, Tenenbaum and Griffiths, 2001), then the learning algorithm is domain-general.

In principle, it is an independent question for each aspect of the learning theory whether it is domain-specific or domain-general. Although these aspects have typically been equated, they are, in fact, separate and should be addressed independently. Any one of these components might be domain-general while the others are domain-specific. This is how we reconcile the opposing viewpoints on linguistic nativism. In the current paper, we provide a case study in which we show that language acquisition can take advantage of a domain-general learning procedure.

However, we also show that this learning procedure can only work when paired with domain-specific filters on data intake.[1]

1.1 Why a domain-general learning procedure?

Probabilistic reasoning (particularly Bayesian) has been shown to be the optimal strategy for solving problems and making decisions given noisy or incomplete information (J. Pearl, 1996). But it is important to keep in mind that reasoning, probabilistically or otherwise, requires an adequate understanding of the representations that are used in the relevant mental computations. A probabilistic learner takes probabilistically available information to derive a conclusion about a discrete representation from a range of antecedently available options (cf. Shannon, 1948).

Crucially, for probabilistic learning such as Bayesian updating to be able to function, the hypothesis space must already be specified (cf. Tenenbaum, Griffiths, & Kemp (2006) for theory-based Bayesian models that emphasize this point). Otherwise, the Bayesian updating procedure has nothing over which to operate. In short, if the learner has no options to select from, probabilistic learning cannot help. Probabilistic learning dovetails with a defined hypothesis space; it does not replace it.

The appeal of probabilistic learning is that one can apply this domain-general learning procedure to representations in any domain, allowing for the same domain-general learning procedure to apply independent of the character of the representations that are to be learned. For language learning, adopting a probabilistic learning approach over linguistic representations allows us to synthesize a learning theory that combines the strengths of both domain-general and domain-specific learning (cf. Yang, 2002). We argue below that, in addition, the domain-specific filters on the learner's data intake play a vital role in constraining the usefulness of a probabilistic (domain-general) learning procedure.

1.2. Filtering & Computational Models

When exploring data intake filtering as a viable strategy for a learner, we must keep in mind the feasibility, sufficiency, and necessity of whatever filter is proposed. For feasibility, we must ensure that the useable data for a learner appears in sufficient quantity. If the data a learner would need are too sparse, then learning will be impossible. In addition, we must make sure the learner can identify the relevant data. Specifically, the data intake filter must be feasible to implement. For sufficiency, we can ask if learners who use the filter exhibit correct behavior. In short, do these learners converge on the correct hypothesis? If filters are sufficient to produce correct behavior, we can examine if filters are necessary. If the correct behavior results regardless of filtering, then filters are sufficient but not necessary. Alternatively, if incorrect behavior results without filtering, then we support the necessity of filtering for the learning problem under consideration.

Exploring the efficacy of data intake filtering is not always so straightforward, however. The situation we want is to precisely control a learner's data intake and then see the effect this has on learning. This is logistically and ethically difficult to achieve with standard experimental techniques in naturalistic settings – we cannot simply restrict children's intake for years on end

[1] The overall message of the paper is unaffected by the question of whether the representations that comprise the hypothesis space of the learner are built in or derived by a domain-general learning procedure. In order for the learning algorithm to work, the learner must be equipped with hierarchical representations of phrase structure by the time the learner attempts to solve the particular problem we examine here. However, we remain neutral as to whether these representations are innate or derived previously.

and then see how it affects their learning. However, this is precisely what we can do with a computational model. It is perfectly feasible to restrict the data intake of a simulated learner in any way we choose and then observe the effect on the model's learning.

One question that might reasonably arise is how much use a simulated learner actually is. Why do we believe that a model of a learner is at all realistic? As Goldsmith & O'Brien (2006) note:

"When the model displays unplanned (i.e. surprising) behavior that matches that of a human in the course of learning from the data, we take some satisfaction in interpreting this as a bit of evidence that the learning models sheds light on human learning."

In short, if the simulated learner accords with human behavior in some non-trivial way that is not purposefully built into the model, we conclude that the assumptions the learning model has made accord with the human learning algorithm. And indeed, there has been a recent surge of computational modeling work examining the effect of data filtering on language acquisition (Sakas & Fodor, 2001; Sakas & Nishimoto, 2002; Yang, 2002; L. Pearl, 2005; Pearl & Weinberg, 2007; among others).

In the case we examine here, we explicitly ground the model with available empirical data from experimental work (Lidz, Waxman, & Freedman, 2003 – henceforth LWF) as well as from corpora of child-directed speech (CHILDES database: MacWhinney, 2000). Again, this is because we want the computational model to reflect the learning process of a human learner as much as possible. Thus, we provide it with estimates of realistic data to learn from and estimates of an appropriate time course of learning.

1.3 English anaphoric one interpretation

The phenomenon under investigation is the interpretation of the anaphoric element *one* in English. There are two different levels of representation that must be considered for interpreting anaphoric *one*: the syntactic level and the semantic level. At the syntactic level, the infant must learn what the linguistic antecedent of *one* is; at the semantic level, the infant must determine what object in the world an NP containing *one* refers to. Both of these levels contribute to the information a Bayesian learner would use when converging on the correct representation of *one.* A linguistic antecedent (syntax) can be translated into a reference to an object in the world (semantics) and so both syntactic and semantic representations are implicated in knowledge of *one*. In terms of learning, the syntactic antecedent of *one* has semantic consequences on what referents are picked out; the semantic referent has syntactic implications of what the linguistic antecedent is. A probabilistic learner can thus use multiple sources of information to inform its hypotheses about the interpretation of anaphoric *one*.

As we will see below, the correct syntactic representation for English adults can be described as the linguistic antecedent of *one* being a string classified as N'. This syntactic knowledge has semantic consequences, which are what LWF used to determine if 18-month olds had that specific syntactic representation. In this way, the knowledge that *one* refers to N' strings traverses both the syntax domain and the semantics domain. So, the learner's hypotheses likewise involve hypotheses in both the syntax and semantics domains.

We will show that a probabilistic learner who does not filter the available data will in fact have a very low probability of converging on the correct interpretation of anaphoric *one*, which involves both the correct syntactic category of the antecedent of *one* and the correct set of referents in the world. However, a learner that ignores *some* of the available data will succeed.

The data that are crucial to ignore are a type of ambiguous data that is quite pervasive in the learner's input.

The probabilistic learning procedure the simulated learner uses is an adapted form of Bayesian updating that is layered with information exploited from the hypothesis space layout. Specifically, in order to learn from the ambiguous data available, the Bayesian learner incorporates the information implicit in hypotheses with a subset-superset relationship. The entire updating procedure is domain-general. Notably, however, this Bayesian learner does not succeed without filters on data intake that are domain-specific.

While we investigate only the interpretation of anaphoric *one* here, the problem we find for an unconstrained Bayesian learner is likely to reappear in language learning more generally. This is because most linguistic knowledge requires learners to align representations across domains (e.g. syntax and semantic reference). Linguistic representations are inherently multi-dimensional, and there are correspondences across different levels of representation. The main difficulty in this case study is that the Bayesian learner uses any data that are informative about any of the representations under consideration. This seems a perfectly reasonable strategy, but leads to failure for anaphoric *one* acquisition. Success requires domain-specific constraints that cause the learner to ignore some potentially useful information. This conclusion casts doubt on Bayesian learning as the primary source of constraints on learners. In other words, the domain-specific theory of representations and a set of domain-specific filters on data intake are required to overcome the overly general nature of the domain-general learning procedure. The point of our paper is not that domain-general learning procedures are useless in language acquisition, but rather that in some situations they work best when supported by domain-specific components of the learning mechanism.

1.4 This article's plot

The computational model we develop in this article uses all the available informative data (unambiguous and ambiguous) as well as both syntactic and semantic data to converge on the probabilities of competing representations. We will show that, even under the most generous estimates of the various parameters involved in such a model, a Bayesian learner lacking domain-specific filters on data intake will fail to converge on the correct interpretation of anaphoric *one*. The crucial conclusion is not that there is an in principle problem with Bayesian learning, but rather that there is an in principle problem with using all of the available data to decide between hypotheses in this case study. Bayesian learning is an effective learning regime, but only when the data are appropriately filtered.

The remainder of the article proceeds as follows. First, we briefly describe the Bayesian learning framework that we will be adopting and detail how the learning procedure operates in the face of variously structured hypothesis spaces. We then describe the grammar of anaphoric *one*, the linguistic phenomenon we use as our case study. We also describe the behavioral evidence indicating that 18-month-olds have acquired the adult representation of anaphoric *one*, and repeat the argument from LWF that the unambiguous data available to children is too sparse to support acquisition of this knowledge. This leads to a feasibility problem for learning only from unambiguous data, namely data sparseness.

We address various proposals to circumvent the sparse data problem, the most prominent of which is to learn from ambiguous data as well as unambiguous data. We argue that a prior proposal advocating a domain-general solution to this problem (Regier & Gahl, 2004) in fact implements implicit domain-specific filters on data intake. Following that, we describe a Bayesian learning model that is truly domain-general, in that it removes all implicit filtering on

the data. We show that such a model fails to converge on the adult representations of anaphoric *one* with high probability, unlike 18-month olds. This highlights the necessity of some additional learning strategy. Moreover, to underscore that the model provides an upper bound on the probability of convergence on the correct interpretation, we describe how the Bayesian learning model would perform even more poorly under a set of less charitable parameter values.

Next, we identify the source of the model's failure. One contributing factor to the rather spectacular failure of the model derives from the link between syntax and semantics. A second contributor to this failure is the abundance of ambiguous data, which given Bayesian learning techniques, causes to the learner to misconverge. We argue that successful acquisition depends on a domain-specific filter on the data. In particular a subset of the ambiguous data must be ignored by the learner. Filtering the data in this way yields sufficient learning behavior in the simulated learner. Because the unfiltered learner fails to acquire the correct interpretation, we infer that filtering is both necessary and sufficient in this case. Finally, we speculate on the origin of the necessary domain-specific filter, which addresses the feasibility of a learner implementing it. We suggest that its roots may lie in a syntactocentric approach to learning anaphoric *one*.

2. Bayesian Updating: How It Works

Bayesian updating is a probabilistic updating procedure that is widely used in natural language processing tasks to update the probabilities of alternate available hypotheses (Manning & Schütze, 1999). Specifically, it calculates the conditional probability of the hypothesis, given the data. There is also evidence for the psychological validity of a procedure like Bayesian updating as a method used by adult humans (Tenenbaum & Griffiths, 2001; Cosmides & Tooby, 1996; Staddon, 1988), children (Xu & Tenenbaum, 2007), and infants (Gerken, 2006). Specifically, these studies demonstrate probabilistic convergence on the more restrictive hypothesis compatible with the observable data. This is in line with the Bayesian updating procedure adopted here when there are two hypotheses under consideration that differ in their level of restrictiveness (section 2.2).

The main purpose of Bayesian updating is to infer the likelihood of a given hypothesis, given a series of examples as input. The implementation of Bayesian updating depends greatly on the structure of the hypothesis space, since the relation of the hypotheses to each other affects how probability is shifted between the different hypotheses. We examine two instances of hypothesis spaces below and their effect on Bayesian updating.

2.1 A Simple Case: Two Hypotheses, Equally Likely

Suppose there are two non-overlapping hypotheses in the set: A and B. By non-overlapping, we mean that the examples in the input will either favor A or favor B unambiguously. There are no examples that signal (or can be accounted for by) both A and B – each hypothesis covers a distinct set of data points. Suppose also that the learner who will be using Bayesian updating has no reason to be biased towards one hypothesis, so the initial probabilities assigned to both A and B are 0.5. These are the prior probabilities associated with each hypothesis.

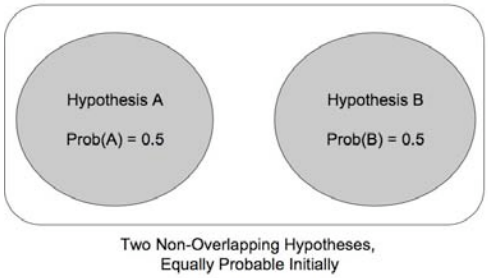
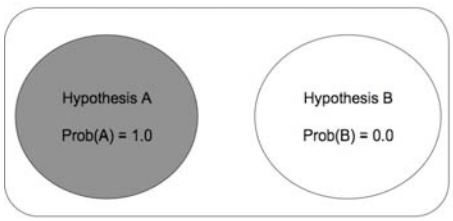
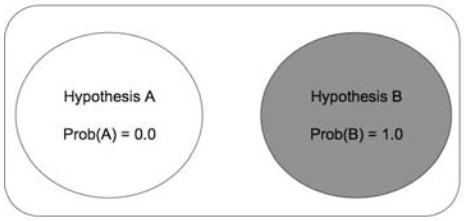


Figure 1. Two non-overlapping hypotheses, equally probable initially. The shading reflects how much probability is associated with each hypothesis.

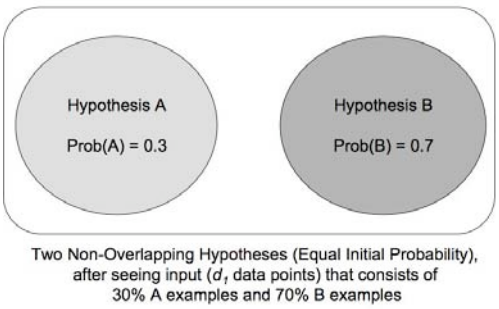
The learner then encounters some amount of data (say d_1 data points) and uses Bayesian updating to shift the probability mass between A and B to reflect the distribution in the data intake. Each data point will cause the learner to shift the probabilities a small amount until the probability distribution among the hypotheses reflects the probability distribution encountered in the intake.



(a)



(b)



(c)

Figure 2. Two non-overlapping hypotheses with equal initial probability after seeing various types of input. The shading reflects how much probability is associated with each hypothesis.

If the data intake consists only of examples of A, the learner will eventually shift the probability so A is close to 1.0 and B is close to 0.0 (2a). Conversely, if the data intake consists only of examples of B, the learner will eventually shift the probability so A is close to 0.0 and B is close to 1.0 (2b). In each of these cases, the learner shifts the majority of the probability to a single hypothesis, thereby converging on one hypothesis as correct. However, it is possible that

the learner will encounter a mixed distribution between A and B in the data intake. If so, the learner will shift the probability to reflect the bias in the perceived distribution since the target state is a probabilistic distribution between A and B. As a concrete example, if the input is consistently 30% A examples and 70% B examples, the learner will eventually shift the probability of A to be significantly less than that of B, reflecting the 30-70 distribution (2c).

2.2. A Not-So-Simple Case: Two Hypotheses in a Subset Relation, Equally Likely

Suppose the hypothesis space again consist of two hypotheses, but one hypothesis is a subset of the other hypothesis. Let A be a subset of B, so all examples of A are also examples of B (Tenenbaum & Griffiths, 2001; Manzini & Wexler, 1987; Berwick, 1985; Berwick & Weinberg, 1984; Pinker, 1979). That is, while B has unambiguous examples, there are no unambiguous examples for A – all examples covered by hypothesis A can also be covered by hypothesis B. Suppose the initial probabilities assigned to both A and B are 0.5, and the learner is trying to decide whether the subset A or the superset B is the correct hypothesis.

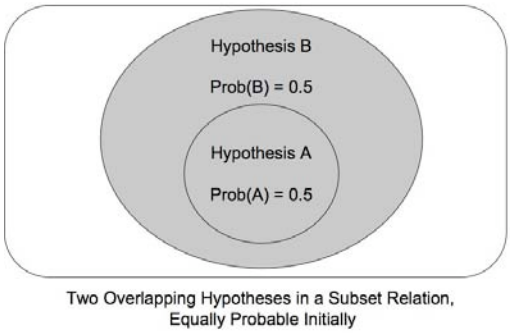


Figure 3. Two overlapping hypotheses in a subset relation, with equal probability initially. The shading reflects how much probability is associated with each hypothesis.

Suppose the learner encounters only unambiguous examples for B in the data intake (say, d_2 data points). Eventually, the learner will shift most of the probability to B (in the limit, B = 1.0, A = 0.0).

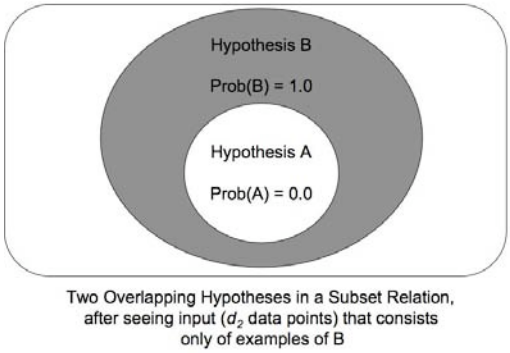


Figure 4. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing d_2 data points that are unambiguous for hypothesis B. The shading reflects how much probability is associated with each hypothesis.

But what if hypothesis A (the subset hypothesis) is the correct one for the target language? All examples covered by hypothesis A are also covered by hypothesis B – they are thus ambiguous data points. It is *impossible* for the learner to encounter any unambiguous data points for hypothesis A. If the data intake consists only of these ambiguous data points, one might expect the learner to remain at a neutral probability of 0.5 for each hypothesis since these data points are compatible with each hypothesis. The learner would be doomed never to converge on the correct hypothesis, the subset hypothesis A.

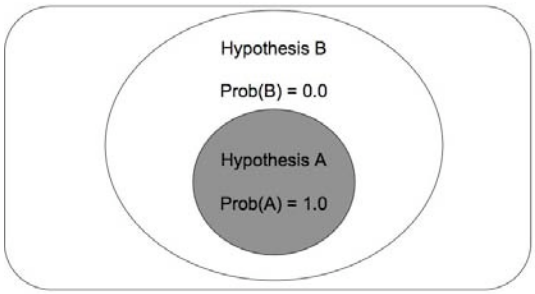
One way to save the learner from this fate is to exploit the layout of the hypothesis space. The Bayesian updating procedure can take advantage of the subset-superset relation of the hypotheses to favor hypothesis A when encountering an ambiguous data point. The logic is as follows:

(1) Logic of Favoring the Subset Hypothesis for an Ambiguous Data point
   (a) If hypothesis B (the superset hypothesis) was correct, the data intake should contain at least *some* examples covered only in the superset B (i.e. unambiguous B examples).
   (b) If only examples covered by the subset A are encountered in the data intake, it becomes more and more unlikely that hypothesis B is correct.
   (c) Therefore, the more the learner encounters only data points in the subset A (even though these are ambiguous data points), the more the learner will favor the subset hypothesis A.

A learner taking advantage of this logic will therefore consider a restriction to the subset A more and more probable as time goes on if only subset data points are encountered. This logic can be implemented in the Bayesian updating procedure itself, and has been referred to as the *size principle* (Tenenbaum & Griffiths, 2001) and more generally in the learning literature as *indirect negative evidence* (Chomsky 1981, Lasnik, 1987; among many others).

Essentially, the smaller size of the set of examples covered by hypothesis A benefits hypothesis A when ambiguous examples are encountered. Specifically, the likelihood of encountering these examples given the smaller set covered by A is greater than the likelihood of encountering these examples given the larger set covered by B. So, A is slightly favored when encountering an ambiguous example covered in its subset.[2] After a sufficient number of ambiguous examples in the data intake (and, importantly for the basic version of the size principle, *no* unambiguous examples of the superset B), A will be highly favored.

[2] The amount A is favored depends on the relative sizes of A and B, which the learner must already know (perhaps as a separate prior) or empirically derive from the data. The smaller A is compared to B, the more A is favored given an ambiguous data point.



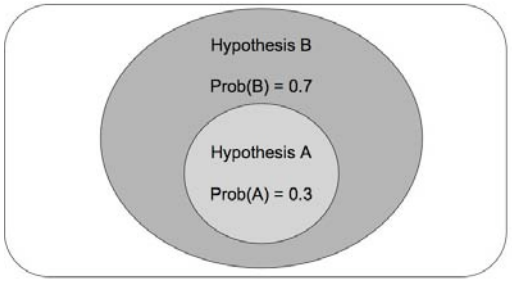
Two Overlapping Hypotheses in a Subset Relation, after seeing input (> d_2 data points) that consists only of examples of A

Figure 5. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing more than d_2 data points that are examples of A. The learner uses the size principle to converge on hypothesis A. The shading reflects how much probability is associated with each hypothesis.

We note that there is a disparity between the quantity of data points required to converge on B when using unambiguous data points as compared to the quantity required to converge on A using ambiguous data points. In particular, if the learner requires d_2 data points to reach probability p for B when encountering unambiguous B data points, the learner will require *more* than d_2 data points to reach p for A when encountering ambiguous data points. This is because the size principle allows A to only be *somewhat* favored for an ambiguous data point while B is *exclusively* favored for an unambiguous B data point, though the actual amount of favoring depends on the relative sizes of A and B. In sum, unambiguous data points carry more information than ambiguous data points. Hence, inferences from unambiguous data points have more impact than inferences from ambiguous data points.

If the data intake has a mixed distribution (both unambiguous B examples and ambiguous examples), the unambiguous B examples will have more effect on the learner's probability distribution than the ambiguous examples that slightly favor A. Both types of data points, however, will contribute to the final probability the learner converges on. Again, the number of data points required to converge on the final probability will be greater in this case (more than d_2 data points) than if only unambiguous B examples were encountered and the correct hypothesis was B exclusively.

It is important to note that exploiting the hypothesis space layout using the heuristic of the size principle is a non-trivial contribution to the learning problem for hypotheses arrayed in a subset-superset relationship. Though it is a heuristic and so not guaranteed to succeed for all cases, it nonetheless has an advantage over approaches that do not exploit the hypothesis space layout. Specifically, if only subset data are encountered, it will converge on the subset.



Two Overlapping Hypotheses in a Subset Relation, after seeing input (> d_2 data points) that consists of 30% A examples and 70% B examples

Figure 6. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing more than d_2 data points that are a mix of examples of A and B. The learner uses the size principle to converge on the probability that reflects the distribution observed in the input. The shading reflects how much probability is associated with each hypothesis.

2.3 Summary

Having given a brief overview of how Bayesian learning can be used to choose from a set of predefined hypotheses in the general case, we turn now to the specific case of anaphoric *one*. As we will see, the hypothesis space for anaphoric *one* is an instantiation of the subset-superset scenario just described. Before detailing the specifics of Bayesian learning for anaphoric *one*, we first describe the correct representation for anaphoric *one*.

3. Anaphoric *One*

3.1. Adult Knowledge: *Grammar*

For English adults, the element *one* is anaphoric to strings that are classified as N' (i.e., the antecedent for *one* is an N' string), as in example (2) below. The structures for the N' strings are represented in figure 7.[3]

(2a) *One* is anaphoric to N' (*ball* is antecedent)
"Jack likes this *ball* and Lily likes that *one.*"
(2b) *One* is anaphoric to N' (*red ball* is antecedent)
"Jack likes this *red ball* and Lily likes that *one*."

[3] Note that the precise labels of the constituents here are immaterial. If the structure is $[_{DP}$ this $[_{NP}$ red $[_{NP}$ $[_{N^o}$ ball]]]], the conclusions reached in this paper would not be changed. Under this labeling, the generalization would be that *one* is anaphoric to constituents labeled NP. Similarly, if syntactic labels were appended with semantic types reflecting the combinatorics of the semantic system, as in (i), then the generalization would be that *one* is anaphoric to nominal constituents of type <e,t>.

(i) $[_{DP, <e>}$ this $[_{NP, <e,t>}$ red $[_{N^o, <e,t>}$ ball]]]

What is important is only that there is some labeling convention by which nominal constituents lacking complements are distinguished from those taking complements.

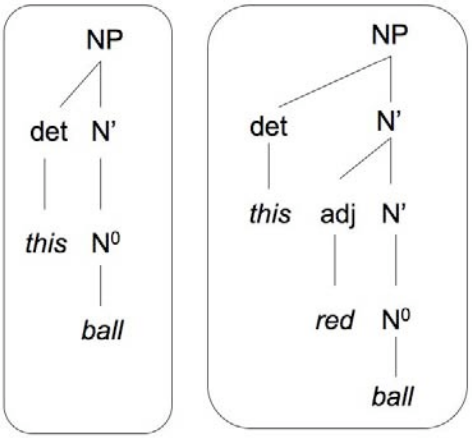


Figure 7. Structures for the N' strings *this ball* and *this red ball*.

These representations encode two kinds of information: constituency structure and category structure. The constituency structure tells us that in a Noun Phrase (NP) containing a determiner (det), adjective (adj) and noun (N^0), the adjective and noun form a unit within the larger Noun Phrase. The fact that *one* can be interpreted as a replacement for those two words (as in (2b)), tells us that those two words form a syntactic unit. The category-structure tells us which pieces of phrase structure are of the same type. That is, both *ball* and *red ball* are of the type N'[4]. The following argument explains how we know this.

Consider the following examples in which *one* cannot be anaphoric to a noun (cf. Baker (1979)):

(3i) a. I met the member of Congress…
b. * …and you met the one of the Society for Creative Anachronism.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ member] [$_{PP}$ of Congress]]]

(3ii) a. I reached the conclusion that syntax is innate…
b. * …and you reached the one that learning is powerful.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ conclusion] [$_{CP}$ that syntax is innate]]]

These contrast with cases in which what follows the head noun is an adjunct/modifier. Here, *one* can substitute for what appears to be only the head noun.

(3iii) a. I met the student from Peoria…
b. … and you met the one from Podunk.
c. [$_{NP}$ the [$_{N'}$ [$_{N'}$ [$_{N0}$ student]] [$_{PP}$ from Peoria]]]

(3iv) a. I met the student that Lily invited to the party
b. … and you met the one that Jack invited.
c. [$_{NP}$ the [$_{N'}$ [$_{N'}$ [$_{N0}$ student]] [$_{CP}$ that Lily invited to the party]]]

[4] Again, the precise labels for the categories do not play a critical role here, provided that there is a distinction between complement-taking nouns and other nominal constituents. See previous footnote.

These cases differ with respect to the status of what follows the noun. In (3i) and (3ii) what follows the noun is a complement, but in (3iii) and (3iv) what follows the noun is a modifier. We can see that *one* can take a noun as its antecedent only when that noun does not take a complement. We represent this by saying that *one* must take N' as its antecedent and that in cases in which there is no complement, the noun by itself is categorized as both N^0 and N'. In other words, in cases like (2a), it must be the case that *ball* = N', as in the structure in Figure 7. If it weren't, we would have no way to distinguish this case from one in which *one* cannot substitute for a single word, as in (3i) and (3ii).

3.2. Pragmatics of anaphoric one

In addition, when there is more than one N' to choose from (as in (2b) above), adults prefer the N' corresponding to the longer string (*red ball*). For example, in (2b) an adult (in the null context) would often assume that the ball Lily likes is red – that is, the referent of *one* is a ball that has the property red (cf. Akhtar et al. (2004)). This semantic consequence is the result of the syntactic preference for the larger N' *red ball*. If the adult preferred the smaller N' *ball*, the semantic consequence would be no preference for the referent of *one* to be red, but rather for it to have any property at all. Importantly, though, this preference is not categorical. It is straightforward to find cases where it is overridden, as in (4):

(4) I like the yellow bottle but you like that one.

Here, it is quite easy to take *one* to refer to *bottle* and not *yellow bottle.*

3.3 Children's knowledge of anaphoric one

But do children prefer *one* to be anaphoric to an N' string (and more specifically the larger N' string if there are two), rather than to an N^0 string? If so, the semantic consequence would be readily apparent: the antecedent for *one* would be phrasal, and hence the referent of *one* would be sensitive to properties mentioned by modifiers in the antecedent. LWF conducted an intermodal preferential looking paradigm experiment (Golinkoff et al., 1987; Spelke, 1979) to see if infants did, in fact, have a preference for the referent of *one* to have properties mentioned by the modifier in the antecedent (i.e., for a red bottle if a potential antecedent of *one* is *red bottle*).

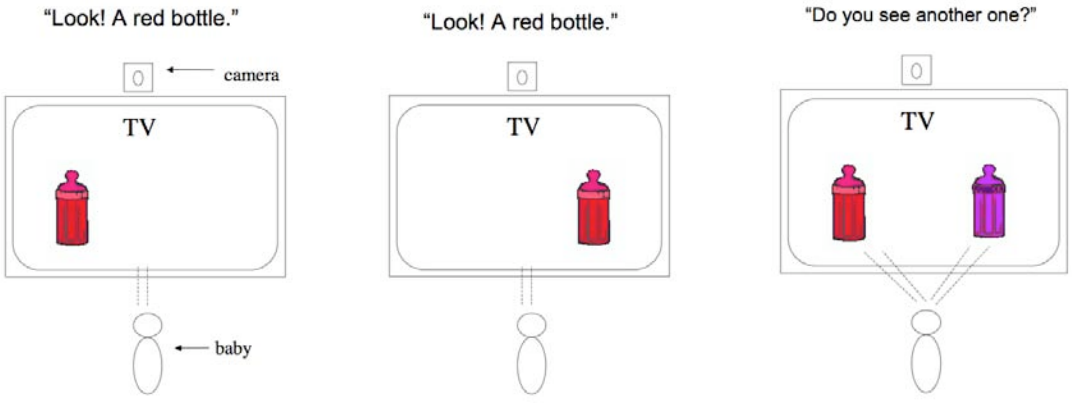


Figure 8. LWF experimental setup.

The infant in the LWF experiment is first shown a bottle of one color, e.g., red, while several utterances of the form "Look! A red bottle." are played. Then, in the test stage, two bottles are shown – one red one and one of another color. The utterance "Do you see another one?" is played simultaneously and the infant's looking preferences are recorded.

The 18-month olds demonstrated a significant preference for looking at the bottle that had the same property mentioned in the N' string – e.g. the bottle that was red when the N' string *red bottle* was a potential antecedent. These same results were obtained when the infants listened to, "Look! A red bottle" followed by "Do you see another red bottle?" (See Lidz & Waxman (in prep.) for more empirical data supporting this.) This suggests that the infants were interpreting these utterances similarly, namely that *one* referred to *red bottle* in the original test condition.

Notably, the infants' response differed from the baseline condition where they heard, "Look! A red bottle" followed by "What do you see now?" In the baseline condition, the infants had a novelty preference and looked longer at the non-red bottle if they had previously seen a red bottle and heard, "Look! A red bottle".

LWF explained this behavior as a semantic consequence of the syntactic preference that *one* be anaphoric to the larger N' string (*red bottle*). If the children had allowed *one* to be anaphoric to N^0 (bottle), they would have behaved similarly to the baseline condition and had a preference for the new bottle they hadn't seen before. Since infants preferred the larger N' string (as adults do) and this larger N' string could not be classified as N^0, LWF concluded that the 18-month olds have the syntactic knowledge that *one* is anaphoric to N' strings in general.

3.4. Hypothesis spaces for anaphoric one

Anaphoric *one* has two hypothesis spaces associated with it: one for the syntactic domain and one for the semantic domain. The fact that there are two separate hypothesis spaces in our learning model, one for syntax and one for semantics, is an assumption that requires justification. An alternative would be to have one hypothesis space that contains both syntactic and semantic components for each hypothesis (as in Regier & Gahl, 2003). There are two reasons that we adopt this assumption. First, there is information available in the input that bears on only one or the other of these components. Thus, separating the components into distinct hypothesis spaces allows the learner to make more targeted inferences. Second, as we will see, successful learning requires a filter on data intake and this filter can feasibly be implemented by separating the syntax from the semantics.[5]

Returning to our hypothesis spaces, the syntactic domain contains hypotheses about what strings can be antecedents for *one*. Each hypothesis covers a set of strings, and is classified by the syntactic category that can generate all the strings in the hypothesis. The semantic domain is a projection of the syntactic domain and contains hypotheses about the interpretation of *one* (specifically what referents in the world *one* can refer to). Each hypothesis covers a set of referents, and is classified by what properties the referents in that set must have. In both domains, there are two hypotheses to choose from. Each hypothesis makes predictions about the data that will be encountered and, consequently, the elements that will be analyzable under that hypothesis.

For each domain, the elements analyzable by one hypothesis are a subset of the elements analyzable by the other. For syntax, the hypotheses under consideration are (a) that *one* is

[5] It is not impossible to implement this filter in the one hypothesis space model. However, this implementation requires a partitioning of the hypotheses that is equivalent to the one we implement. Thus, the decision to have two linked hypothesis spaces as opposed to one composite hypothesis space does not affect the results we report here.

anaphoric to strings that are classified as N^0 and (b) that *one* is anaphoric to strings that are classified as N'. Every string in N^0 can also be classified as N' but there are strings in N' that cannot be classified as N^0. Therefore, the strings that comprise the N^0 set are a subset of the strings that comprise the N' set.

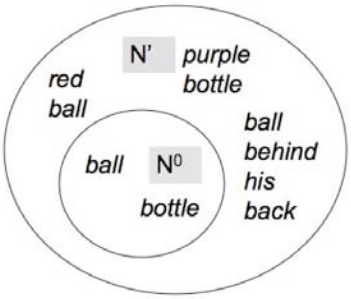


Figure 9. The syntax hypothesis space, N^0 vs. N'. All the elements in the sets are strings that are possible antecedents of *one*. Every string classified as N^0 can also be classified as N'. In addition, there are strings in N' that are not in N^0, and so the N^0 set is a subset of the N' set.

For the semantic interpretation, the referents of *one* could have the restriction that they must have the property named by the modifier; alternatively, the referents of *one* could have no restriction on what property they have. Since the modifier is linguistically not part of the N^0 (recall figure 7) and instead is part of the N' phrase, we will refer to the property named by the modifier as the *N'-property*. We will refer to referents with no restrictions as being *any-property* referents, since these referents can have any property (though of course they must still be instances of the noun in the antecedent, e.g. balls, if the antecedent is *red ball*). So, in the semantic domain, the two hypotheses under consideration are (a) that the referent of *one* is restricted to have the N'-property and (b) that the referent of *one* can have any property (is not restricted to have the N'-property).

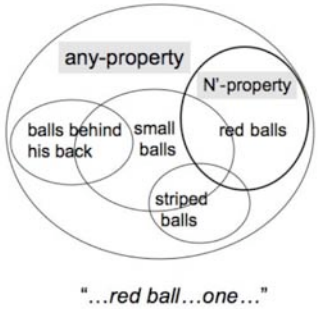


Figure 10. The semantic hypothesis space, N'-property vs. any-property. Any-property is a superset of N'-property. Note that in order to define the sets (N'-property vs. any-property), the utterance must be used to determine the salient property that the referent of the antecedent has. The salient property can be determined from the linguistic antecedent of *one*.

Just as in the syntactic domain, the elements predicted by one hypothesis are a subset of the elements predicted by the other. Every referent that has the N'-property (say red for *red ball*) is a member of the N'-property set. By definition, every member of the N'-property set is also a member of the any-property set, since the N'-property is one of the properties available for

objects to have. However, there are members of the any-property set (say green balls for the linguistic antecedent *red ball*) that do not have the N'-property (red). So, since all the members of the N'-property set are members of the any-property set, the N'-property set is a subset of the any-property set. Moreover, some members of the any-property set are *not* members of the N'-property set. So, the any-property set is a superset of the N'-property set in the semantic domain.

The difficulty for a Bayesian learner using the size principle becomes apparent when we examine how the two prediction spaces defined by the hypotheses are connected. Specifically, in the syntactic domain, the relative complement of the subset in the superset (the set of strings that are in the superset but not the subset, such as *red ball*) is linked to the subset in the semantic domain; the subset in the syntactic domain is linked to the superset in the semantic domain. For ease of exposition, we will refer to the relative complement of the subset in the superset as the "exclusive superset".

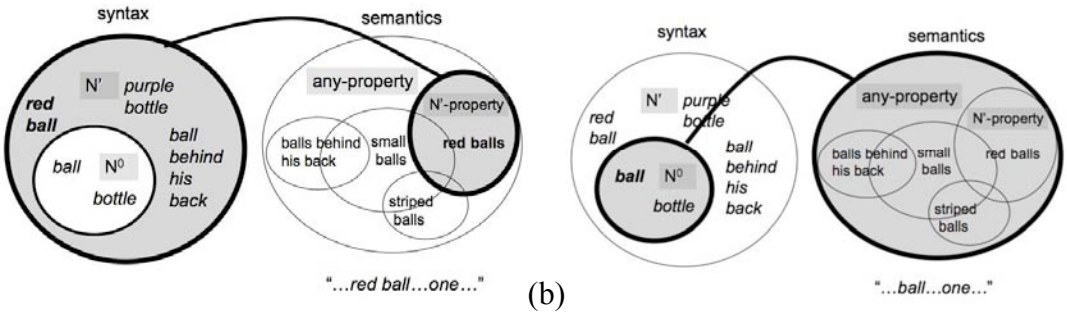


Figure 11. In the syntactic domain, the exclusive superset is linked to the subset in the semantic domain (a). The subset of the syntactic domain is linked to the superset in the semantic domain (b).

This is due to the compositional property of syntactic representations: larger syntactic constituents (such as the N' *red ball*) have meanings that are restrictions on the meanings (and so the referents) of their constituent subparts. In syntax, the strings in the exclusive superset (e.g. *red ball*) designate a subset of referents in the semantics (e.g. the red balls); the strings in the subset of the syntax (e.g. *ball*) designate the superset of referents in the semantics (e.g. all balls).

Because the syntactic and semantic representations are linked in this fashion, a Bayesian learner that relies on indirect evidence to shift probability towards the subset will receive conflicting information from across the two domains. For instance, the learner will encounter ambiguous data that favors the syntactic subset (figure 11b: the wrong answer for English anaphoric *one*). The learner will also encounter ambiguous data that favors the semantic subset which is linked to the exclusive superset in the syntax that implicates N' (figure 11a: the correct answer for English anaphoric *one*). However, this will not negate the aforementioned syntactic evidence that favors the syntactic subset N^0. Yet, the learner shouldn't ignore available syntactic information since anaphoric *one* has a representation at the syntactic level. Thus, we can see that an unrestricted Bayesian learner that uses all available data (syntactic and semantic) will need to overcome conflicting information across domains in order to converge with a high probability on the correct representation of anaphoric *one*.

It is important to recognize that the problem of linked hypothesis spaces extends far beyond the particular case of anaphoric *one*. Because syntactic structures are semantically compositional, this problem may persist across the acquisition of any aspect of the grammar that

depends on the link between syntax and semantic reference, depending on the distribution of the data.

3.5. Sparse data problems

In order to determine whether children's knowledge could have been acquired on the basis of experience with the relevant forms and structures, LWF conducted a corpus analysis on child-directed speech. The important empirical question was how frequently data appeared in child-directed speech that unambiguously signaled that *one* was anaphoric to N' instead of N^0. If the data were not frequent, learning this syntactic knowledge would be difficult. The distribution LWF found is displayed in table 1 below.

Total Data in Corpus	Total # with anaphoric *one*
54,800	792
Data Type	# of data points
Unambiguous	2
"*Jack wants a red ball, but Lily doesn't have another one for him*." (Lily doesn't have another ball with the property red.)	
Type I Ambiguous	36
"*Jack wants a red ball, and Lily has another one for him.*" (Lily has another red ball for Jack.)	
Type II Ambiguous	750
"*Jack wants a ball, and Lily has another one for him."* (Lily has a ball with any number of properties.)	
Ungrammatical	4
"*…you must be need one.*" (Adam19.cha, line 940)	

Table 1. The distribution of utterances in the corpus examined by LWF.

All data are defined by a pairing of utterance and environment. We now elaborate on the pairings for each data type. Unambiguous antecedent data have the form in (5) and bias the learner strongly towards the syntactic exclusive superset and the semantic subset:

(5) Unambiguous antecedent example
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."
Environment: Jack wants a red ball, but Lily doesn't have another red ball – she has another ball with different properties.

Because Lily does indeed have a ball, the antecedent of *one* cannot be *ball*. However, Lily's ball is not red, so the antecedent of *one* can be *red ball*. Since *red ball* can only be classified as N', these data are unambiguous evidence that *one* can be anaphoric to N'.

An example of this type taken from the Adam corpus in CHILDES (MacWhinney, 2000) is given here. (Adam40.cha, line 890)

(6) CHI: Do you have another flat tire?
MOT: No. I don't think I have one.

In this context, the mother had a tire, but not a flat tire, so the antecedent of *one* is unambiguously *flat tire*.

Type I ambiguous antecedent data have the form in (7) and bias the learner towards the semantic subset, which is linked to the syntactic exclusive superset:

(7a) Type I ambiguous antecedent example
Utterance: "Jack wants a red ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has the property red.

(7b) Type I ambiguous antecedent example
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him ."
Environment: Lily doesn't have another ball at all.

For data of the form in (7a), Lily has a ball, so the antecedent of *one* could be *ball.* However, Lily also has a ball that is red, so the antecedent of *one* could be *red ball.* Because *ball* could be classified as either N' or N^0, these data are ambiguous between *one* anaphoric to N' and *one* anaphoric to N^0.

An example of this type taken from the Adam corpus in CHILDES (MacWhinney (2000)) is given here (Adam01.cha, line 291).

(8) MOT: That's a big truck.
MOT: There goes another one.

In this context, *one* could be taken to refer to either *truck* or *big truck*.

For data of the form in (7b), Lily does not have a ball – but it is unclear whether the ball she does *not* have has the property red. For this reason, the antecedent of *one* is again ambiguous between *red ball* and *ball,* and *one* could be classified as either N' or N^0. There were no examples in either Adam or Nina's corpus of this form.

Type II ambiguous antecedent data have the form in (9), and bias the learner (incorrectly) towards the syntactic subset:

(9a) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has various properties.

(9b) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, but Lily doesn't have one for him."
Environment: Lily does not have another ball.

For both forms of type II ambiguous data, the antecedent of *one* must be *ball*. However, since *ball* can be classified as either N' or N^0, such data are ambiguous with respect to what *one* is anaphoric to.

An example of this type taken from the Adam corpus of CHILDES (MacWhinney (2000)) is given here (Adam01.cha, line 566).

(10) CHI: my pillow my

MOT: That's a good one to jump on.

Because there are no modifiers in the antecedent, *my pillow*, this data is uninformative about the structure of *one*.

There were no examples in either Adam or Nina's corpus of the form (8b).

Ungrammatical data involve a use of anaphoric *one* that is not in the adult grammar, such as in (11):

(11) Ungrammatical antecedent example
Utterance: "…you must be need one."

Since the utterance is already ungrammatical, it does not matter what environment it is paired with. The child will presumably be unable to resolve the reference of *one*. Such data is therefore noise in the input and is not used by the learner to update the hypotheses.

The vast majority of the anaphoric *one* input consists of type II ambiguous data (750 of 792, 94.7%). Type I ambiguous data makes up a much smaller portion (36 of 792, 4.5%). Ungrammatical data are quite rare (4 of 792, 0.5%), and unambiguous data rarer still (2 of 792, 0.25%). Since LWF considered unambiguous data as the only informative data, they concluded that such data seemed far too sparse to definitively signal to a learner that *one* is anaphoric to N'.

This seems in line with theory-neutral estimations of the quantity of data required for acquisition by a certain age (Legate & Yang, 2002). Specifically, other linguistic knowledge acquired by 20 months required at least 7% unambiguous signatures in the available data (Yang (2004) referencing Pierce (1992)). At least 1.2% unambiguous data was required for acquisition by 36 months (Yang (2004) referencing Valian (1991)). So, independent of what acquisition mechanism is assumed, having 0.25% unambiguous data makes it unlikely that the learner would be able to acquire the correct interpretation of anaphoric *one* by 18 months.

LWF's experimental results, however, suggested that 18-month olds know that *one* is anaphoric to N'. They therefore claimed that such knowledge does not need to be learned from the available data. Instead, the learner would have other biases that would allow this knowledge to be derived from the data available. One possibility (cf. Hornstein & Lightfoot (1981), Baker (1979)), which LWF advocated, would be that the child is constrained only to hypothesize phrasal antecedents for pronouns. Thus, once the child identified *one* as a pronominal form, the possibility that it was anaphoric to N^0 would simply never be considered as a potential hypothesis. An alternative possibility is that the child had somehow previously realized that *one* should be treated differently from other nouns, and had conducted a statistical analysis over its syntactic distribution with respect to complements and modifiers to determine that *one* was of the category N' [6] (Foraker et al., 2007). The child would then realize *one* is an N' category, and so, if it is anaphoric, it will be anaphoric to something of the same syntactic category. The child would then not need to learn this from the syntactic and semantic data available when *one* is anaphoric to a linguistic antecedent explicitly mentioned in the discourse.

[6] Recall that we are using the term N' as a cover term for any syntactic category that excludes complement-taking nouns. (see fn 4).

4. Learning Anaphoric *One*

4.1 Suggestions for learning that one is anaphoric to N'

Two replies to LWF made suggestions for how this syntactic knowledge and semantic interpretation would be learnable from the available data, and would not need to already be previously known. Akhtar et al. (2004) noted that even if the percentage of unambiguous data is quite small, 18-month olds have still been exposed to an estimated 1,000,000 utterances; this should yield a larger number of unambiguous data than the LWF corpus analysis obtained. However, we note that this does not address the problem of the *frequency* of unambiguous data being so low. As highlighted in the section 3.5, a child would need the frequency of unambiguous data to be much higher (over 25 times more frequent) in order to converge on the knowledge by 18 months.

Yet, it is reasonable to question the actual amount of data available to learners by 18 months, particularly if we wish to instantiate a learning model for anaphoric *one*. To estimate what this quantity would be, we must consider that learning the syntactic and semantic properties of *one* can only commence once the child has some repertoire of syntactic categories. Thus, we posited that the learning period begins at 14 months because there is experimental data supporting infant recognition of the category Noun and the ability to distinguish it from other categories such as Adjective at this age (Booth & Waxman, 2003). If learners hear approximately 1,000,000 sentences from birth until 18 months, they should hear approximately 278,000 sentences between 14 months and 18 months. The adjusted expected distribution of anaphoric *one* data is displayed in table 2.

Total Data before 18 months	Total # with anaphoric *one*
~278,000	4017
Data Type	# of data points
Unambiguous	10
"*Jack wants a red ball, but Lily doesn't have another one for him*." (Lily doesn't have another ball with the property red.)	
Type I Ambiguous	183
"*Jack wants a red ball, and Lily has another one for him.*" (Lily has another red ball for Jack.)	
Type II Ambiguous	3805
"*Jack wants a ball, and Lily has another one for him."* (Lily has a ball with any number of properties.)	
Ungrammatical	19
"*…you must be need one.*"	

Table 2. The expected distribution of utterances in the input to learners between 14 and 18 months.

Perhaps the most striking feature of this distribution is that there are still pitifully few unambiguous data points available. With only 10 chances to hear unambiguous data (on this estimate), a learner could well miss out due to fussiness, distraction, or other vagaries of toddler life. And again, this is still 0.25% of the anaphoric *one* data, which is well below the estimate of the amount of unambiguous data needed to acquire knowledge by 36 months (estimated at 1.2%, Yang (2004)), let alone by 18 months.

Regier & Gahl (2004) (henceforth, R&G) offer a solution: use a domain-general Bayesian learner that can extract information from ambiguous data as well. This will give the learner significantly more data to learn from. Using a Bayesian learning model that implements the size principle of Tenenbaum & Griffiths (2001), R&G demonstrate how a learner could use both unambiguous and type I ambiguous data to converge on the correct representation. We review their learning model in the next section.

4.2. A Regier & Gahl Bayesian learner

The power of R&G's model comes from using indirect evidence available in the type I ambiguous data. This is an attractive strategy, since there are nearly 20 times as many type I ambiguous data points as there are unambiguous data points (183 to 10). This raises the useable data for the learner up to 4.8% (193 of 4017), which seems more in line with the amount required for acquisition as early as 18 months (Yang (2004)). The indirect evidence itself is derived solely from the environment in which type I ambiguous data are uttered – specifically, by the learner examining the distribution of the referents of *one*. For example, suppose the learner hears type I ambiguous data such as the example in (7a) (repeated below as (12)):

(12) Type I Ambiguous
Utterance: "Jack wants a red ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has the property red.

Since the adult preference is to choose the larger N' as the antecedent, the antecedent of *one* will nearly always be *red ball* and the referent of the NP containing *one* will have the property red. The learner is able to observe the simultaneous presence of the larger N' as potential antecedent (*red ball*) and a referent in the world of *one* with the property mentioned in the N' (red). We note that this observation requires the learner to have a very abstract notion of what to generalize over. It is insufficient to generalize over a single property such as "red" or "behind his back"; instead, the learner must generalize over "property mentioned in the N' antecedent".

Now, the connection between the N' antecedent and a referent with the property mentioned in the N' will be true for some portion of the type I ambiguous data.[7] Crucially, for R&G's model, it is *never* true that the referent of *one* definitively lacks the property mentioned in the N' antecedent (i.e. the referent of *one* is definitively not red when the antecedent is *red ball*). A Bayesian learner using the size principle is very sensitive to this fact in the following way:

(13) Bayesian Learner Logic

(a) For type I ambiguous data, suppose that the referent of *one* could have any property, and not necessarily have the property mentioned in the larger N' antecedent. Suppose also that the set of potential referents for an utterance like (12) is represented in figure 12.

[7] This reasoning will not work for type I ambiguous data of the form in (7b): "Jack wants a red ball, but Lily doesn't have another one for him", where Lily does not have a ball. This is because the learner cannot tell whether or not the ball Lily doesn't have has the property red. These data are therefore not useful as indirect evidence. Such data did not occur in the Adam and Nina corpora from which our estimates are derived.

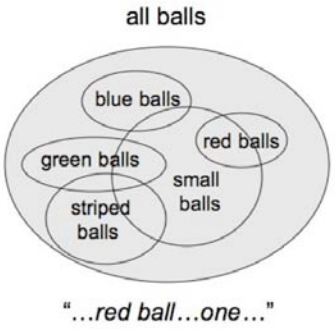


Figure 12. The set of potential referents for *one* in the world when an utterance such as "Jack wants a red ball, and Lily has another one for him" is heard.

(b) The actual distribution of referents observed by the learner, however, is only a particular subset of all the possible referents.

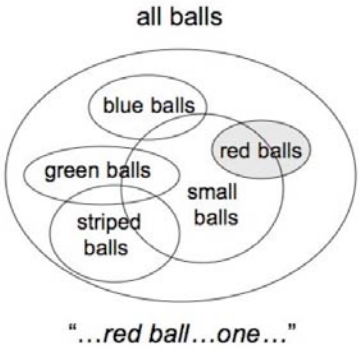


Figure 13. The observed set of referents for *one* when an utterance such as "Jack wants a red ball, and Lily has another one for him" is heard.

(c) It is highly unlikely that the referent of *one* is only ever a member of the subset if the referent could be any member of the superset. The Bayesian learner will therefore consider a restriction to the subset to be more and more probable as time goes on. This is the size principle of Tenenbaum & Griffiths (2001): if there is a choice between a subset and the superset, and only data from the subset is seen, the learner will be most confident that the subset is the correct hypothesis. Thus, the learner uses the lack of data for the superset as indirect evidence that the subset is correct.

Correlatively, the learner uses the relative sizes of the hypotheses to favor the subset when encountering an ambiguous data point. The specific instantiation of the bias for the subset (red balls) given a single subset data point is based on the likelihood of encountering that subset data point. The likelihood of choosing a specific member of the subset (a red ball) is higher if members can be drawn only from the subset (red balls), as opposed to if members can be drawn from the superset (all balls). This occurs because the superset necessarily has more members to choose from, and therefore there is a lower probability of choosing a specific subset member.

The amount of bias a subset data point gives the subset depends on the relative sizes of the subset and superset. If the superset (all balls) has many more members than

the subset (red balls), the likelihood of drawing a specific member from the subset (a red ball) when any member from the superset could have been chosen is low. The bias towards the subset (red balls) given a subset data point (a red ball) is then higher. In contrast, if the superset (all balls) has only a few more members than the subset (red balls), the likelihood of drawing a specific member from the subset (a red ball) when any member from the superset could have been chosen is higher. The bias towards the subset (red balls) given a subset data point (red ball) is then lower.

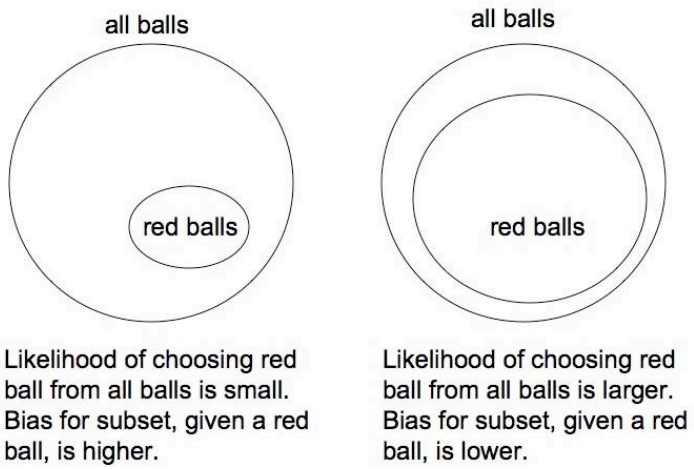


Figure 14. Comparison of different ratios of superset to subset, the likelihood of choosing a member of the subset, and the effect on subset bias.

(d) Once the learner is biased to believe that there is a restriction to the subset of referents described by the property mentioned in the N' (*red* in *red ball*), the learner then assumes that the correct antecedent is, in fact, the larger N'.[8] Since the larger N' cannot be classified as N^0, the learner then knows that *one* always has an N' antecedent.

(e) For the LWF experiment, a Bayesian learner would have converged on the subset of red bottles as the potential referents of *one* in the test utterance. Given a choice between a red and a non-red bottle, the Bayesian learner therefore looks at the bottle that belongs to the correct subset: the red bottle.

A great strength of the R&G model is that the bias to choose the subset, given indirect evidence, does not need to be explicitly instantiated as a constraint on learning. Instead, it falls out neatly from the mathematical implementation of the Bayesian learning procedure itself – the size principle of Tenenbaum & Griffiths (2001). This model therefore draws on a domain-general learning strategy.

However, the model implemented in the R&G study still harbors two implicit biases about domain-specific data filters on the learner's intake. The first bias is that only unambiguous and type I ambiguous data are used; type II ambiguous data are ignored even though they may also provide indirect evidence to a Bayesian learner. The second bias is that only semantic data

[8] R&G's model demonstrates how this could happen after very few type I ambiguous data.

(the referents of *one*) are used to converge on the syntactic knowledge of what *one* is anaphoric to; syntactic data are ignored.

In the remaining sections of the article, we will see that stripping away these two biases (and thus creating an unbiased learner truer to the spirit of R&G's proposal) leads to markedly different results from those of R&G. Specifically, once we remove these two biases, we will discover that a Bayesian learner will *not* learn that *one* is anaphoric to N' with high probability and will *not* choose the adult interpretation of the larger N' constituent with high probability when there is a choice between N' constituents. So, this unrestricted Bayesian learner will (a) have a preference for the wrong syntactic analysis (N^0) and (b) a preference for the wrong combined syntactic and semantic interpretation.

The benefit that comes from using indirect negative evidence to shift the majority of the probability to the subset in the hypothesis space is tempered by the link between the two levels of representation. In particular, the semantic interpretation is a projection from the syntax. If indirect learning leads to the subset N^0 in the syntax, then the semantic preference to choose the interpretation consistent with the larger N' constituent when there is a choice between two N' constituents will not be helpful to the learner in most cases. This is simply because the learner will not choose the N' analysis very often, and so will have no need to access the semantic interpretation preference. Thus, the existence of multiple levels of representation reduces the efficacy of this kind of learner, at least for this data set.

5. An Equal-Opportunity Bayesian Learner

We have named the unrestricted domain-general learning model the Equal-Opportunity (EO) Bayesian Learner since it removes the two implicit biases of R&G's Bayesian learner and so gives equal treatment to all data. First, it denies privileged status to a subset of the data and instead uses all the informative data available: unambiguous, type I ambiguous, and type II ambiguous. Second, it denies privileged status to semantic data – syntactic and semantic data are both used to shift probability between opposing hypotheses. There is an intuitive logic to using both types of data, since one should presumably use syntactic data (among other kinds of data) to converge on syntactic knowledge.[9] This syntactic knowledge has semantic consequences, which are displayed in the LWF experiment. If a Bayesian learning procedure, unconstrained by domain-specific filters, is to be an effective domain-general learning solution, it should correctly acquire knowledge that spans domains such as syntax and semantics as well as knowledge contained completely within these domains.

5.1. EO Bayesian Learning

The EO Bayesian learning model uses Bayesian reasoning to update the learner's confidence in each of two alternative hypotheses.[10] We detail the learning process

---

[9] Note that even if we believed the knowledge about *one* was stated purely in semantic terms, the data that any grammar predicts will include both syntactic data (i.e. what the linguistic antecedent for *one* is) and semantic data (what the referent of *one* is). So, excluding either kind of data is an arbitrary restriction on the learner that would need to be justified. For this reason, the modeling choice to include both syntactic and semantic data does not rely on a particular specification of knowledge about anaphoric *one*.

[10] The implementation we use differs from the R&G Bayesian learner by being more conservative about updating the probabilities of the competing hypotheses. The R&G learner is quite liberal about shifting probability to the superset hypothesis – in fact, a *single* piece of data for the superset is enough to shift *all* the probability to that hypothesis. See Appendix B for details.

independently for each of the two domains (syntax and semantics) that are relevant for determining the appropriate structure of anaphoric *one*. We then describe how to implement the updating algorithm, given that these two domains are linked and that there are multiple sources of information that are informative about anaphoric *one*.

5.2. Updating the Syntax Hypotheses

Recall that there are two hypotheses under consideration in the syntactic domain: the N' hypothesis and the N^0 hypothesis. The N' hypothesis takes the antecedent of *one* to be a constituent of the category N'; the Nº hypothesis takes the antecedent of *one* to be a constituent of the category N^0.

We represent the probability that the N' hypothesis is correct with $p_{N'}$. Because there are only two hypotheses in the hypothesis space, and because probabilities range from 0 to 1, the probability that the N^0 hypothesis is correct is $1 - p_{N'}$. We set the initial value of $p_{N'}$ before the learner has observed any data to 0.5 as an instantiation of the assumption that both hypotheses are equiprobable.

The update function requires a single parameter *t*, which represents the total amount of data expected during the learning period, and can be thought of as the total amount of change the real learner's brain is allowed to undergo before settling into the final state. In the simulated learner here, we simply quantify that amount of change as the total estimated amount of useable data available during the learning period (4017 data points, if using all available data).[11] The model uses *t* to determine how much probability shifting should be done, given a single piece of data. If *t* is small, only a small number of changes are allowed and each piece of data shifts the probability quite a lot; conversely, if *t* is large, a large number of changes are allowed and each piece of data shifts the probability a smaller amount. The value of *t* we use here will allow the modeled learner to converge as close as possible to an endpoint (e.g. $p_{N'} \approx 1.0$). In this way, we hope to estimate the best-case scenario for this kind of learner.

The exact update functions for $p_{N'}$ depend on the data type observed – unambiguous, type I ambiguous, or type II ambiguous. Unambiguous and type I ambiguous data cause the learner to use the function in (14a), which is essentially an implementation of the indirect negative evidence update function used by the R&G model. Type II ambiguous data, which were not considered by the R&G learner, cause the EO Bayesian learner to use the function in (14b).[12]

(14a) Update function for unambiguous and type I ambiguous data

Utterance: "…red ball…one…"

World: referent has the property red (unambiguous & some type I ambiguous) or it is unknown if referent has the property red (some of type I ambiguous)

$$p_{N'} = \frac{p_{N' \text{ old}} * t + 1}{t + 1}$$

[11] Of course, the value of *t* is essentially arbitrary, but in order to model this learning process, it needed to be estimated. While the estimate presented here seems fair, we present a range of possible *t*-values in Appendix C. What we see there is that the size of t does not influence the final probability of the correct grammar.

[12] For details of how these functions are derived, see appendix A.

(14b) Update function for type II ambiguous data

Utterance: "…ball…one…"

World: referent has various properties (type II ambiguous)

$$p_{N'} = \frac{p_{N' \text{ old}} * t + p_{N' \mid a}}{t + 1}$$

The update functions are derived by using the mathematical framework laid out in Appendix A. To briefly summarize, a binomial distribution centered at $p_{N'}$ is used to approximate the learner's expectation of the distribution of the data to be observed. Data points from this distribution fall into two classes: they either have the "property" of being an N' data point or they do not have this property (and are instead N^0 data points). If $p_{N'}$ is 0.5 (as it is initially), the learner expects half the informative data points to be N' data points. Using the derivations described in Appendix A, we can then derive equation (14a) for updating $p_{N'}$ after the learner has encountered an unambiguous data point.

The update function for unambiguous and type I ambiguous data (which comprise 193 of the data points) depends only on the prior probability that N' is the correct hypotheses ($p_{N' \text{ old}}$) and *t*. An intuitive interpretation of the unambiguous data update function is that the numerator represents the learner's confidence that the observed unambiguous N' data point *u* is a result of the N' hypothesis being correct; the denominator represents the total data observed so far. Thus, 1 is added to the numerator because the learner is fully confident that *u* indicates the N' hypothesis is correct; 1 is added to the denominator because a single data point has been observed.

Unambiguous data signal that the N' hypothesis is correct (in that only the N' hypothesis could have produced *u*) and so should be treated with full confidence by the learner. In contrast, the type I ambiguous data do not indicate that only the N' hypothesis could have produced *u* – these data are *ambiguous* between the N^0 and N' hypotheses. Thus, a smaller value should be added to the numerator for such data to indicate less than full confidence that only the N' hypothesis could have produced *u*.

However, we will allow the EO Bayesian learner to treat the type I ambiguous data with full confidence in the N' hypothesis. We make this allowance for two reasons. First, we know of no principled way to reasonably estimate how much confidence should be associated with a type I ambiguous data point. Second, this allowance is generous towards the Bayesian learner because it allows the model to overestimate the confidence the learner has in the N' hypothesis. If we were less generous and lessened the confidence in the type I ambiguous data, the probability of N' would only be lower than what we present here. As we will see below, even with this generous estimate, the learner will fail to assign sufficient probability to the N' hypothesis.

The update function for type II ambiguous data (14b), which comprise 3805 of the data points, depends on the prior probability that N' is the correct hypotheses ($p_{N' \text{ old}}$), *t,* and a confidence value ($p_{N' \mid a}$). The intuitive interpretation for this function remains the same as the interpretation for the function in (14a): the numerator represents the learner's confidence that the observed ambiguous utterance-world pairing *a* is a result of the N' hypothesis being correct; the denominator represents the total data observed so far. Thus, a value less than 1 ($p_{N' \mid a}$) is added to the numerator because the learner is only partially confident that ambiguous data point *a* indicates the N' hypothesis is correct; and, 1 is added to the denominator because a single data point has been observed. The partial confidence value $p_{N' \mid a}$ depends on the likelihood that the

utterance in *a*, which has only a noun string as the antecedent of *one* (ex: "…ball…one…"), would be produced if any N' string could have been chosen from the set of N' strings ($p_{N \text{ from } N'}$). See appendix A for details about how we derive this value.

The likelihood value $p_{N \text{ from } N'}$ is what allows the learner to retrieve information from the type II ambiguous data. The more unbalanced the ratio of noun-only strings to other strings in the N' set, the stronger the effect of the size principle will be that biases the learner towards the subset N^0 hypothesis. Example (15) displays how much biasing occurs after a single piece of type II ambiguous data, assuming a current $p_{N'}$ of 0.5, a ratio of noun-only strings to total N' strings of 0.25, and a *t* of 4017.

(15) Updated $p_{N'}$ after a single type II ambiguous data point *a*
Let $p_{N'} = 0.5$, $p_{N \text{ from } N'} = 0.25$, and $t = 4017$.
Updated $p_{N'} = .499925$ (a very slight bias for the N^0 hypothesis)

While the amount of bias towards the N^0 hypothesis is quite small, keep in mind that the majority of the data is type II ambiguous and so these small biases will add up over time.

5.3. Updating the Semantics Hypotheses

Recall that there are two hypotheses under consideration in the semantic interpretation domain that are projections from the syntactic domain: the N'-property hypothesis and the any-property hypothesis. The N'-property hypothesis requires the referent of *one* to have the property mentioned in the N' antecedent (e.g. red if the potential antecedent was *red ball*); the any-property hypothesis allows the referent of *one* to have any property. In this case, it's the N'-property hypothesis that represents the subset hypothesis. Thus, as above, the size principle will favor this hypothesis for any data that is compatible with both hypotheses.

We represent the probability that the N'-property hypothesis is correct with $p_{N'\text{-prop}}$. Because there are again only two hypotheses in the hypothesis space, the probability that the any-property hypothesis is correct is $1 - p_{N'\text{-prop}}$. We set the initial value of $p_{N'\text{-prop}}$ before the learner has observed any data to 0.5 as an instantiation of the assumption that both hypotheses are equiprobable.

The update function requires two parameters: *t* and *c*. As before, *t* represents the total amount of data expected during the learning period and is instantiated in this model as 4017, the estimated amount of data available during the learning period. The parameter *c* represents the number of properties (or *categories* of referents) in the world that the learner is aware of (e.g. red, striped, behind his back, etc.).

For the semantic domain, the data are divided according to how the properties of the referent of *one* compare to the salient property in the N' antecedent. The data types, representing the utterance-world pairings, are same-property, different-property, and unknown-property.

Same-property examples are those in which the potential antecedent of one mentions some property and the referent of one also has that property. Some of the data analyzed as type I ambiguous in the syntactic domain are same-property data. There are 183 or less data points of this form (because some portion of type I ambiguous are unknown-property data points).

(16a) Example of same-property data (syntax: type I ambiguous)
Utterance: "Jack wants a red ball, and Lily has another one for him."
World: Lily has another red ball for Jack.

The referent of *one* (the ball that Lily has) has the same property mentioned in the N' antecedent (red).

The data analyzed as unambiguous in the syntactic domain are also same-property data in the semantic domain. There are 10 data points of this form. Because these data necessarily include negation, seeing why they are same-property data is a bit complicated. Consider the example in (16b).

(16b) Example of same-property data (syntax: unambiguous)
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."
World: Lily has a non-red ball for Jack.

The speaker in this situation is asserting the absence of a red ball. The referent of *one* is a red ball that is not present in the situation. Thus, the meaning of *one* includes the property mentioned in the antecedent.

Because the N'-property hypothesis depends on matching the property overtly mentioned in the modifier (e.g. *red* of *red ball*), type II ambiguous data are not informative for choosing between the two hypotheses. This is simply because there is no overtly mentioned modifier, as shown in (16c). Therefore, the semantic interpretation projection from the syntactic hypothesis space is a single hypothesis (the any-property hypothesis).[13] Since the semantic domain only has one hypothesis for type II ambiguous data, no information can be inferred about what the correct hypothesis would be when there is more than one semantic interpretation to choose. The learner therefore ignores the semantic hypothesis space when encountering type II ambiguous data.

(16c) Example of same-property data (syntax: type II ambiguous)
Utterance: "Jack wants a ball, and Lily has another one for him."
World: Lily has a ball with some property for Jack.

A different-property example is given in (17), when the potential antecedent has a property mentioned in the modifier (e.g. red of *red ball*), but the referent of *one* does not have this property. This situation would occur in rare cases, perhaps as noise or perhaps because of a pragmatic bias.

(17) Example of different-property data (syntax: type II ambiguous)
Utterance: "Jack likes a red ball, and Lily likes that one."
World: Lily likes a ball that is not red. (i.e., the referent of *one* is a non-red ball, even though the potential antecedent mentions the property *red*).

In this case, the semantic interpretation hypothesis unambiguously favored is the any-property hypothesis, since the data point is specifically in the exclusive superset of balls that do

[13] Alternatively, the learner could view the N' hypothesis space as being equivalent to the any-property hypothesis space, because there is no explicitly mentioned modifier. On this approach, the sets are identical in size and so neither hypothesis is favored by this kind of data point. Moreover, a Bayesian learner confronted with this type of data will be pushed towards the uniform probability distribution between two hypotheses of 0.5. So, the benefit gained from other more informative data that would raise the probability above 0.5 would be negated by the type II ambiguous data. By choosing to implement the model the way we have in the paper, we are providing a generous estimate of the learner's probability of landing on the correct interpretation.

not have the N'-property (red). So, this kind of data strongly biases the learner towards the any-property hypothesis, the superset hypothesis in the semantic domain. That, in turn, biases the learner towards the subset in the syntactic domain (the smaller N' constituent, if the N' analysis is chosen). However, we will be generous and assume that this data does not occur in the EO Bayesian learner's dataset. This assumption will cause the EO Bayesian learner to (again) overestimate the probability assigned to the N'-property hypothesis, $p_{N'\text{-prop}}$.

Finally, we come to the unknown-property data, as in (18).

(18) Example of unknown-property data (syntax: type I ambiguous)
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."
World: Lily has no ball for Jack.

In the example in (18), the speaker is asserting the absence of a ball. The referent of *one* is a ball, with some unknown properties, that is not present in the situation. Thus, the referent of *one* may or may not include the property (red) mentioned in the potential antecedent.

A portion of type I ambiguous data consists of unknown-property data. Such data cannot be used for updating the probabilities of the opposing semantic hypotheses. However, we will be generous and allow R&G's assumption to hold true: none of the type I ambiguous data are of this form. Therefore, we will allow all type I ambiguous data to be of the form in (16a), which is an example of same-property data. This again gives an overestimation of $p_{N'\text{-prop}}$, which is the subset in the semantic hypothesis space. Consequently, this will bias the learner towards the exclusive superset in the syntactic hypothesis space, N'. Thus, the model here will again overestimate the amount of probability the learner will assign to the correct hypothesis for the structure and interpretation of anaphoric *one*, given an utterance with more than one potential antecedent.

Table 3 represents the expected distribution of data for updating the semantic hypotheses in this model.

Total Data before 18 months	Total # with anaphoric *one*
~278,000	4017
Data Type	# of data points
Same-Property	10 + 183
"*Jack wants a red ball, and Lily has another one for him.*" (Lily has a red ball for Jack.) "*Jack wants a red ball, but Lily doesn't have another one for him*." (Lily has a non-red ball for Jack.)	
Different Property	0
"*Jack likes this red ball, and Lily likes that one.*" (Lily likes a ball without the salient property that the antecedent referent has.)	
Unknown Property	0
"*Jack wants a red ball, but Lily doesn't have another one for him."* (Lily has no ball for Jack.)	

Table 3. The expected distribution of utterances in the input to the Bayesian learner for updating the semantics hypotheses. Note that the type II ambiguous data points are uninformative in the semantic interpretation domain, so those 3805 data points are ignored.

The exact update functions for $p_{N'\text{-prop}}$ depend on the data type observed. However, the only update function relevant for this model is the same-property update function (19), which is similar to its syntactic counterpart in (14b). [14] In both cases, the subset hypothesis is favored upon encountering an ambiguous data point.

(19) Update function for same-property data

$$p_{N'\text{-prop}} = \frac{p_{N'\text{-prop-old}} * t + p_{N'\text{-prop} \mid s}}{t+1}$$

The update function for same-property data depends only on the prior probability that the N'-property hypothesis is correct ($p_{N'\text{-prop-old}}$) and t. An intuitive interpretation of the update function is that the numerator represents the learner's confidence that the observed utterance-world pairing s is a result of the N'-property hypothesis being correct; the denominator represents the total data observed so far. For the different-property data, 0 is added to the numerator because the learner has no confidence that the N'-property hypothesis is correct. The same-property data behave as the type II ambiguous data in the syntactic domain: a value less than 1 ($p_{N\text{-prop} \mid s}$) is added to the numerator since the same-property data point is consistent with both the N'-property hypothesis and the any-property hypothesis (the N'-property is, by definition, a member of the any-property set). The learner is therefore *not* fully confident that s indicates the N'-property hypothesis. The partial confidence value $p_{N'\text{-prop} \mid s}$ depends on the likelihood that the referent of *one* in s, which has the same property mentioned in the N' antecedent (ex: "…red ball…one…", referent has property 'red'), would have been chosen if a referent with any property could have been chosen from the set of potential referents. This value is $1/c$, given c properties in the world. See appendix A for details about how we derive the partial confidence value $p_{N'\text{-prop} \mid s}$. As for the denominator, we add 1 for both the same-property data point and the different-property data points because a single data point has been observed.

5.4. The Updating Algorithm for Linked Domains and Multiple Information Sources

Recall that there is an inherent connection between the syntax and the semantics. In particular, the subset hypothesis in the syntax corresponds to the superset hypothesis in the semantics, and vice versa. That is, the N' hypothesis in the syntax, which represents the superset in this domain, is connected to the N'-property hypothesis in the semantics, which represents the subset in that domain. Similarly, the N^0 hypothesis in the syntax, which represents the subset in this domain, is connected to the any-property hypothesis in the semantics, which represents the superset in that domain. Given this arrangement of hypothesis spaces, any piece of data impacting a hypothesis in one domain should impact the corresponding hypothesis in the other domain by the same amount. We now provide a description of how we model this process.

First, suppose the learner receives an unambiguous or type I ambiguous data point (which have two strings as potential antecedents, e.g. *red ball* or *ball*). This data point can be analyzed in either domain, syntax or semantics. So, the learner chooses which one to analyze it in first. Then, the update functions described above are employed to determine the amount the probability that should be shifted within that domain. Next, the probability is shifted in the other domain by the same amount. See figure 15, which shows the learner analyzing the data in syntax and updating both syntax and semantics. Now, the learner analyzes the data point in the other domain, applies the update functions described previously to determine the amount the

[14] For details of how this update function is derived, see Appendix A.

probability that should be shifted within this domain. Next, the probability is shifted in the other domain by the same amount. See figure 16, which shows the learner analyzing the data in the semantics and updating both semantics and syntax.

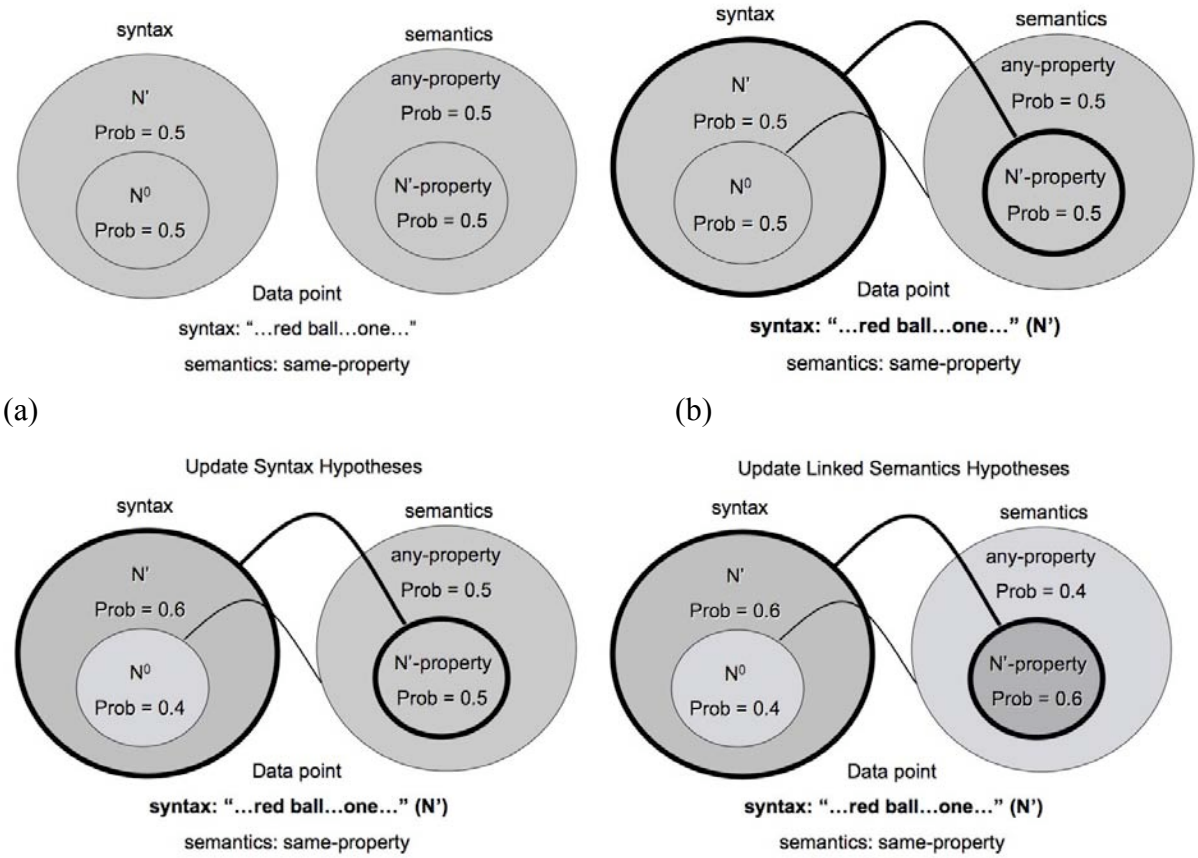


Figure 15. The learner encounters a data point (a) and analyzes it first in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c) and the probability of the linked semantics hypotheses (d).

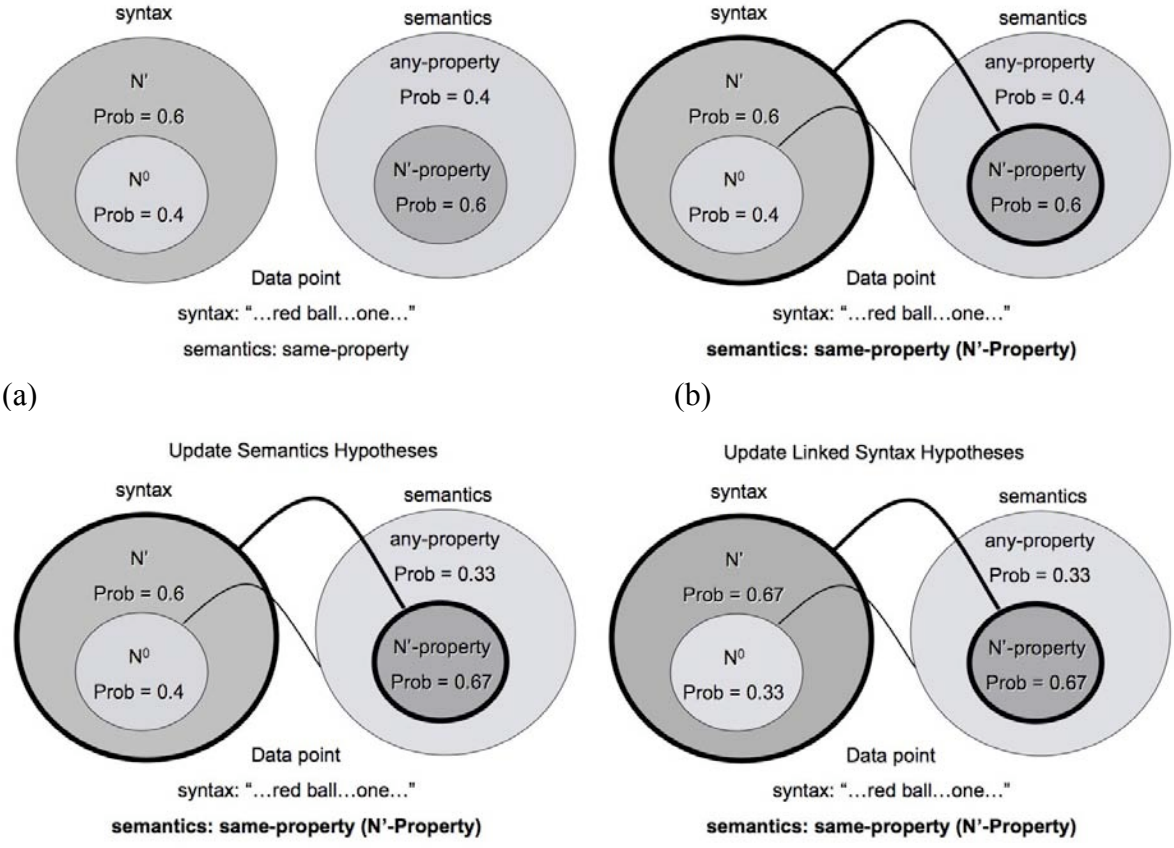
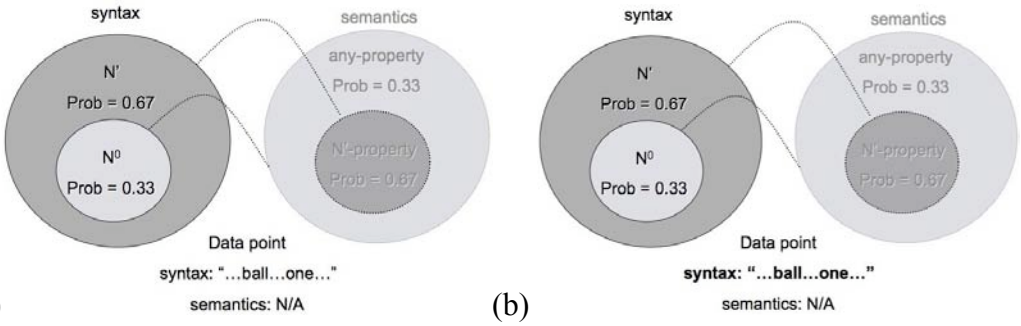
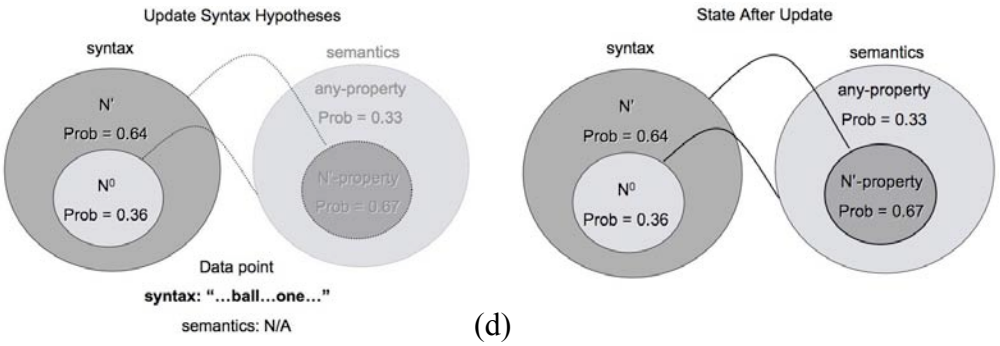


Figure 16. After analyzing the data point in the syntax domain and updating the probabilities across the domains, the learner then starts at the state in (a) and analyzes the data point in the semantics domain (b). Then, the learner updates the probability of the semantics hypotheses (c) and the probability of the linked syntax hypotheses (d).

The update process differs for a type II ambiguous data point, however. This is because there is only one string that is the potential antecedent (e.g. *ball*), and the projection from the syntax to the semantics leaves only one interpretation (any-property). Type II ambiguous data points are thus uninformative for the semantic interpretation domain. So, the learner simply updates in the syntax domain alone, as shown in figure 17. The semantic interpretation domain is ignored for this type of data.





(c) (d)

Figure 17. The learner encounters a type II ambiguous data point (a) and analyzes it in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c). The final state after update is show in (d). Importantly, the semantic domain is not influenced by the type II ambiguous data point because there is only one semantic interpretation available for an antecedent with no modifiers (e.g. *ball*), the any-property hypothesis. The semantic domain is only influenced when there is more than one potential antecedent, leading to more than one semantic interpretation.

The type II ambiguous data updating highlights how the two domains (syntax and semantics) can have different probabilities associated with linked hypotheses despite the link. For example, just because the syntactic exclusive superset (N') is linked to the semantic subset (N'-property) does not mean these two hypotheses will have the same probability. Type II ambiguous data only updates the syntactic hypothesis space, and so will alter the syntactic domain probabilities without affecting the semantic domain probabilities.

5.5. What Good Learning Would Look Like

In the model, the learner initially assigns equal probability to the two hypotheses in each of the two domains: in the syntax, N^0 and N', and in the semantics, N'-property (corresponding to the larger N' constituent interpretation, e.g. *red ball*) and any-property (corresponding to the smaller N' constituent interpretation, e.g. *ball*). The probability of choosing the preferred adult interpretation, given an utterance with two potential antecedents, depends on choosing the correct hypothesis in each domain. So, if the learner hears, "Look! A red bottle! Do you see another one?" (as in the LWF experiment), the interpretation of *one* is calculated as in (20), which is schematized in the decision tree in figure 18.

(20) Interpreting *one* in "Look! A red bottle! Do you see another one?"

(a) Determine if the antecedent of *one* should be N^0 or N', using $p_{N'}$.

(b) If the antecedent is N^0, then the referent can have any-property.

(c) If the antecedent is N', use $p_{N'\text{-prop}}$ to determine if the smaller N' constituent interpretation (any-property) or the larger N' constituent interpretation (N'-property) should be used.

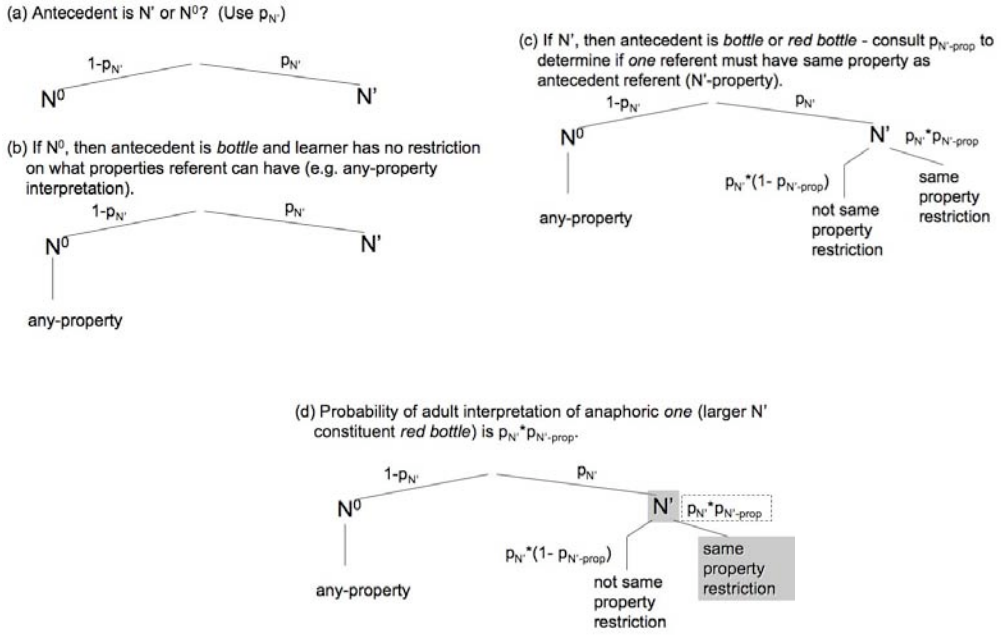


Figure 18. Decision tree to interpret anaphoric *one* in utterances with more than one potential antecedent, such as "Look! A red bottle! Do you see another one?" The probability of having the adult interpretation (*one* = *red bottle*) is $p_{N'}*p_{N'\text{-prop}}$.

The probability of choosing the preferred adult interpretation (the larger N' constituent is the antecedent of *one*) is the product of the probability of choosing the correct hypothesis in the syntax (N') and that of choosing the correct hypothesis in the semantic interpretation (N'-property = larger N' constituent): 0.500 * 0.500 = 0.250. Given that the end state should be a probability near 1, a good learning algorithm should have a trajectory like that illustrated in figure 19. In short, the learner should steadily increase the probability of choosing the preferred adult interpretation.[15]

[15] Note that the modeled learner will not converge to 1.0 in this simulation (or indeed, in any Bayesian model, except in the limit). However, we emphasize that we are looking for qualitatively correct behavior – specifically, an increasing probability of the correct interpretation of anaphoric *one*. Moreover, if we want to map the model onto real learners, along with the assumption that real learners have categorical knowledge, we could implement a thresholding function where the probability is mapped to 1.0 if it exceeds some threshold (e.g., 0.8).

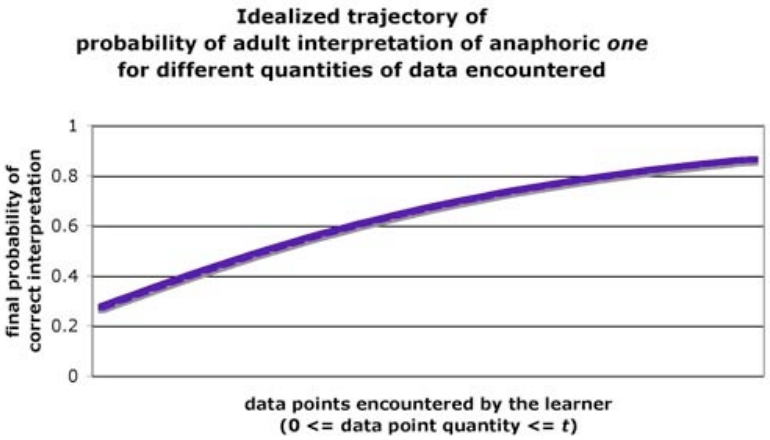


Figure 19. The idealized trajectory of the probability of the correct interpretation for anaphoric *one* as a function of the data points encountered by the learner.

5.6. Simulating an EO Bayesian Learner

Now that we have established how an EO Bayesian Learner learns and what the ideal learning outcome would be, we can simulate learning over our estimate of the set of data that 18-month olds have been exposed to. Each data point is analyzed in both the syntax and semantics domains, as relevant to the data type; and, each data point is classified for both syntax (unambiguous, type I ambiguous, or type II ambiguous) and semantics (same-property only, by generous assumption).

5.6.1. Syntax

The probability $p_{N'}$ is updated as each data point is observed. The model requires a value for $p_{N \text{ from } N'}$, the probability of choosing a noun-only string from the N' string set. This requires that we determine how many strings are in the N' set. There are two ways of doing this. First, we could allow a string to consist of individual vocabulary items ("bottle", "ball", "ball behind his back", etc.). Alternatively, we could allow a string to consist of individual categories (Noun, Noun PrepositionalPhrase, etc.). Recall that as $p_{n \text{ from } N'}$ increases, the ratio between superset size and subset size decreases and the N'-hypothesis is not penalized as much by a type II ambiguous data point. This means that a higher $p_{n \text{ from } N'}$ will generate a higher estimate for $p_{N'}$. Therefore, to be generous and maximize the model's estimate of $p_{N'}$, we choose the option that maximizes the value of $p_{n \text{ from } N'}$ and allow the strings in the N' set to consist of individual categories instead of vocabulary items. The number of categories is necessarily smaller than the number of vocabulary items in those categories, and so this yields a larger value for $p_{n \text{ from } N'}$.

Let the set of strings in N' = {Noun, Adjective Noun, Noun PrepositionalPhrase, Adjective Noun PrepositionalPhrase}.[16] The probability of producing a Noun string from this N' string set is 1/4 or 0.25. We can now look at the semantic domain.

5.6.2. Semantics

The probability $p_{N'\text{-prop}}$ is updated as each data point is observed. The model requires a value for *c*, the number of properties in the learner's world. Recall that as *c* increases, the ratio

[16] This is still a conservative estimate – there are likely to be additional category strings in N', such as Adjective Adjective Noun, because language is recursive. Additional strings would again lower $p_{n \text{ from } N'}$.

between the superset (any-property) and subset (N'-property) increases; the higher this ratio, the more the subset hypothesis (N'-property) is rewarded whenever a same-property data point is encountered. Data from the MacArthur CDI (Dale & Fenson, 1996) suggest that 14-16 months olds know at least 49 adjectives. Therefore, we estimate that an 18-month old learner should be aware of at least 49 properties in the world.[17]

Note however that it is unlikely all 49 properties to choose from would be represented in a given situation (nice balls vs. red balls vs. blue balls vs. pretty balls, etc.). Instead, a subset of the available categories the learner knows would be available in each case (perhaps as few as two: a red ball vs. a blue ball, for instance). So, assuming the learner considers the potential 49 properties the semantic referent in a given situation *could* have had will be an overestimation of the categories the learner actually considers. Because of this, the simulated learner will receive more bias towards the semantic subset (the correct interpretation of anaphoric *one*) than a real learner would. This will yet again yield an overestimation of a real learner's probability of choosing the more restricted referent set in the semantics, and thus an overestimation of the probability of the learner choosing the correct interpretation of anaphoric *one*.

5.6.3. Linked Domain Updating

Recall that the update algorithm analyzes each data point in two domains and shifts the probability between the opposing hypotheses within each domain and across domains accordingly, as relevant. As we can see in figure 20, the learning trajectory as a function of the amount of data seen does not match our ideal learning outcome. In fact, as the learner encounters more data, the probability of the adult interpretation steadily drops to a final value of 0.171. This final value represents the product of the probability of the correct syntactic hypothesis ($p_{N'}$), which is 0.310 (1000 simulations, sd = .00377) and that of the correct semantic interpretation hypothesis ($p_{N'\text{-prop}}$), which is 0.551 (1000 simulations, sd = .00382).[18] Thus, based on the data observed, the learner is extremely unlikely to access the preferred adult interpretation for *one* (i.e., that *one* is anaphoric to strings described by N', and that the referent of *one* must have the N' property) in an utterance with two potential antecedents.

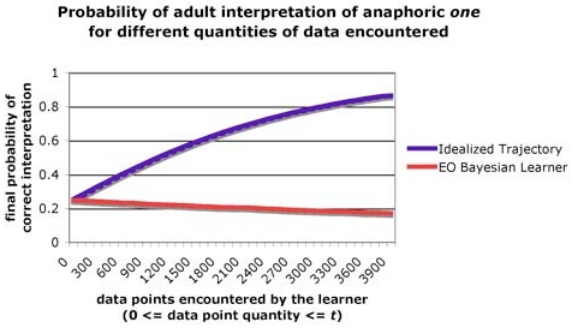


Figure 20. The EO Bayesian Learner's trajectory as a function of the amount of data encountered compared against the idealized trajectory for a learner.

[17] In reality, there are still more properties due to the combination of adjectives (nice red, big striped) and prepositional phrases (nice…behind his back, big striped…in the corner). We will not consider the consequences of recursive modification here.

[18] Note that this value is obtained using the procedure in which the learner chooses at random whether to analyze the data point in the syntax first or in the semantics first for unambiguous and type I data. The same value is obtained if the learner always analyzes the data point in the syntax first and if the learner always analyzes the data point in the semantics first.

6. The Outcome of an EO Bayesian Learner

To summarize, even with conservative estimates of various parameters, the EO Bayesian learner is heavily biased against the preferred adult interpretation of anaphoric *one* in an utterance with two potential antecedents. In fact, the probability of converging on the preferred adult interpretation of anaphoric *one* is quite small (0.171). In short, there is less than a one in five chance of an EO Bayesian learner converging on the correct interpretation for anaphoric *one*.

This result is strikingly different from that reported in R&G, who found overwhelming success for a Bayesian learner. What is the source of this difference? Recall that R&G's domain-general learner made use of only a subset of the available data and gave priority to semantic data over syntactic data. However, if a Bayesian learner is unconstrained in its data intake, then we would expect that it does not favor one type of data over any other - favoring one type of data over another represents a domain-specific filter.

This EO Bayesian model, in contrast, lacks any domain-specific filter on data intake. It uses all the available data (unambiguous, type I ambiguous, and type II ambiguous) and treats syntactic and semantic data as equally relevant to the learner. As we can see, such an unconstrained domain-general learning procedure on its own fails to converge on the correct interpretation of anaphoric *one* with high probability.

This failure is especially striking because of how generous we were regarding the data available to the EO Bayesian learner and how the learner interpreted that data. Below, we highlight where we was generous and see that revoking that generosity only pushes the final probability of choosing the preferred adult interpretation closer to zero. So, we will conclude that unconstrained (and specifically, unfiltered) Bayesian learning by itself is not sufficient to model human learning or behavior in this domain.

As noted above, there were two places in the construction of the model where we biased the learner towards the correct interpretation of anaphoric *one*. First, we gave a generous interpretation of the available data by providing a liberal estimate of the amount of informative data in the environment. Second, we made conservative assumptions about the learner's understanding of the environment. Even in the face of this generosity, the EO Bayesian learner failed.

In the first case, we were unable to determine a fair estimate of the amount of informative data in the environment – for example, the confidence a learner had in the type I ambiguous data (section 5.2), the quantity of type I ambiguous data that were informative (sections 5.2 & 5.3), and the quantity of data indicating the non-preferred adult interpretation (section 5.3). Consequently, we maximized the size of the informative data set in order to get an upper bound on the probability of converging on the correct interpretation. In what follows, we leave these assumptions as is.

In the second case however, we will show one way in which we can relax the conservative assumptions about the learner's understanding of the environment to make these assumptions more realistic. As we will see, the results reported above represent an upper bound on the probability of converging on the correct interpretation of anaphoric *one* when there are two potential antecedents. Changing the relevant assumptions only decreases this probability further.

The conservative assumption we examine concerns the value of $p_{N \text{ from } N'}$, which is the probability of observing a Noun-only string, given the set of all the N' strings. We previously described the elements of the N' string set as category strings, such as Noun and Adjective Noun.

However, if we describe the elements of the N' string set as strings consisting of vocabulary items, such as "bottle" and "red bottle", the probability of observing a Noun-only string is much smaller: it is the number of Noun-only strings divided by the total number of N' strings in the learner's language. The MacArthur CDI (Dale & Fenson, 1996) suggests that 14-16 month olds know about 247 nouns and 49 adjectives. Therefore, the total number of N' strings for an 18-month old learner consists of at least all the nouns and adjective+noun combinations, which is 247+49*247=12350.[19] Using these (still somewhat conservative) estimates, $p_{N \text{ from } N'}$ is 0.0201. This is considerably smaller than the previous value of 0.25. Recall that the smaller the value of $p_{N \text{ from } N'}$, the more the N' hypothesis is penalized whenever a type II ambiguous data point is encountered.

Using this less generous value of $p_{N \text{ from } N'}$ (0.0201, instead of 0.25), the probability of converging on the adult interpretation is the product of the probability of the correct syntactic hypothesis (0.235, 1000 simulations with sd = 0.00316) and the probability of the correct semantic interpretation hypothesis (0.554, 1000 simulations with sd = 0.00358), which is 0.130. On the current, more realistic estimate of the model's parameter, the learner now has less than a one in six chance of converging on the preferred adult interpretation of anaphoric *one* in a situation where there are two potential antecedents for *one.*

7. On the Necessity of Domain-Specific Filters on Data Intake

We began our discussion with the observation that a learning theory can be divided into three components: the representational format, the filters on data intake, and the learning procedure. The EO Bayesian learner attempted to solve the problem of anaphoric *one* using a prespecified representational format[20], but no domain-specific filters or learning procedures. In contrast, the model presented by R&G, which also used a prespecified representational format and a domain-general learning procedure, implicitly used two domain-specific filters on data intake. This model succeeded. We can now examine whether both of these filters are necessary to converge on the preferred interpretation of anaphoric *one*. In particular, we can ask what happens to the EO Bayesian learner if we use these filters, separately and together, and thus examine the necessity of the domain-specific filters.

The first filter implicit in R&G's model was to systematically exclude type II ambiguous data. These are examples in which the antecedent for anaphoric *one* is an NP containing no modifiers (e.g. …*ball…one…*). We will instantiate a variant of the model that follows R&G's model in excluding this data. This variant will, like the original EO Bayesian learner, take into account both the semantic and syntactic consequences of its hypotheses, but ignore the type II ambiguous data.

To simulate this no-type-II-data filter, we considered only the unambiguous and type I ambiguous data points (193, by our estimate). Both the syntactic data and semantic data were used for updating, thus making use of the link across the two domains and the fact that there are multiple sources of information. When we run the model on this data set, the final probability for the N' hypothesis in the syntax and the N'-property hypothesis in the semantics is 0.930. There is no deviation, since the data points consist of the 10 unambiguous data points, which are

---

[19] Again, this is a conservative estimate since there are still more N' strings from combinations of prepositional phrases as well as adjectives with prepositional phrases, for instance – e.g. "bottle in the corner", "big striped ball behind his back", etc. The effects of recursive modification only exacerbate the problem.

[20] Although our model requires antecedent knowledge of X-bar theoretic structures, it is an independent question whether these are innate or derived from experience.

maximally informative for the N' and N'-property hypotheses, and the 183 type I ambiguous data points, which we generously assumed were maximally informative for the N' and N'-property hypotheses. Moreover, there are no countervailing data points for the alternative hypotheses (N^0 in the syntax and any-property in the semantics). Thus, the probability for the correct hypotheses is continually increased. The product of the two probabilities, which represents the probability of converging on the correct interpretation for anaphoric *one*, is 0.865. This is a sharp improvement over the filter-free variant of the model (over 5 times more likely to converge on the correct interpretation).

The second filter implicit in R&G's model is to use only semantic data. That is, alternative syntactic hypotheses were evaluated only with respect to the predictions they made about the referents of phrases containing anaphoric *one*. These are the semantic consequences of the syntactic hypotheses. However, these hypotheses were not evaluated with respect to the predictions they made about the set of possible strings that would be available as antecedents for anaphoric *one*. So, the syntactic implications of the syntactic hypotheses were not considered.

Consider now a variant of the EO Bayesian learner that learns only from the semantic consequences of its syntactic hypotheses. In the semantic interpretation domain, that learner maintained two hypotheses: the N'-property hypothesis and the any-property hypothesis. The probabilities of these two hypotheses are updated on the basis of semantic data. Moreover, these hypotheses are linked to the syntactic hypotheses. The N'-property hypothesis is linked to the N' hypothesis (specifically, the exclusive superset of the N'-hypothesis); and, the any-property hypothesis is linked to the N^0-hypothesis. Consequently, by updating the probabilities of the semantic hypotheses, we also update the probabilities of the syntactic hypotheses. If we ignore the syntactic consequences of the hypotheses, then the only way to update the syntactic hypotheses is via the link to the semantic hypothesis space.

If we simulate an EO Bayesian learner that only learns via the semantic analysis of the data, the final probability for $p_{N'}$ and $p_{N'\text{-prop}}$ is 0.810. As with the previous filter, there is no deviation because there are no countervailing data points and so the probability is continually increased. This is because only data with semantic consequences are considered, and so the type II ambiguous data is ignored. Its effect on the final probability is thus nullified. The final probability of converging on the correct interpretation is the product of the two probabilities, which is 0.656. This is a marked improvement over the unfiltered Bayesian learner; the semantics-only filtered Bayesian learner is nearly four times as likely to converge on the preferred adult interpretation of anaphoric *one*. However, this probability is significantly below the one obtained by using the no-type-II ambiguous data filter (.865 probability against .656 probability). Analyzing the data only in terms of its semantic interpretation can generate significant improvement over the original EO Bayesian learner, but seems to fall short of the benefit gained by simply ignoring the type II ambiguous data outright.

We now consider the consequences of using both of these filters simultaneously. Recall that the effect of the semantics-only filter, which restricted the learner to using only the semantic analysis, was that only semantic data could impact the hypotheses. This results in the type II ambiguous data being excluded from consideration, as it is uninformative with respect to the alternate semantic interpretations since it has only one potential antecedent. The no-type-II-data filter explicitly excludes type II data. So, if the model use these two filters in concert, the result is *the same* as when it used the semantics-only filter alone; the type II ambiguous data is excluded (by the semantics-only filter, due to its lack of semantic consequences, and by the no-type-II-data filter explicitly) and only semantic data can impact the probabilities associated with the hypotheses (due to the semantics-only filter). Thus, the resulting probabilities for the N'

hypothesis and N'-property hypothesis are 0.810 and the probability of the preferred adult interpretation of anaphoric *one* is 0.656. Since using both filters yields an identical result to using the semantics-only filter alone, the benefit gained from using the no-type-II-filter is lost. It is therefore in the interest of the learner to apply only the no-type-II-filter. That is, the learner should ignore type II ambiguous data, but still use both syntactic and semantic data equally to update the hypothesis spaces.

In fact, R&G essentially implemented this filter in their model. They implemented it by providing their learner with only a restricted set of data as input. Though they called this a domain-general learner, the restricted data set provided to the learner should actually be seen as the expression of two domain-specific filters on data intake. As it turns out, at least one of these filters appears to be necessary for successful acquisition. And, as we have seen, filtering out the type II ambiguous data provides the strongest benefit to the domain-general learner.

To summarize, the EO Bayesian learner shows us that a learner not equipped with domain-specific filters on data intake cannot converge on the correct interpretation for anaphoric *one*. Figure 21 displays the learning trajectories and outcomes for the full set of simulations: no filter, no-type-II-data filter, semantics-only filter, both filters. As we can see, using the no-type-II-data filter by itself yields the highest probability for the correct interpretation. Moreover, the efficacy of this filter is negated when used with the semantics-only filter. In other words, the ideal learner must use both syntactic and semantic evidence, but be restricted in which sentences it takes as opportunities to learn from. A domain-specific filter on data intake thus appears necessary.

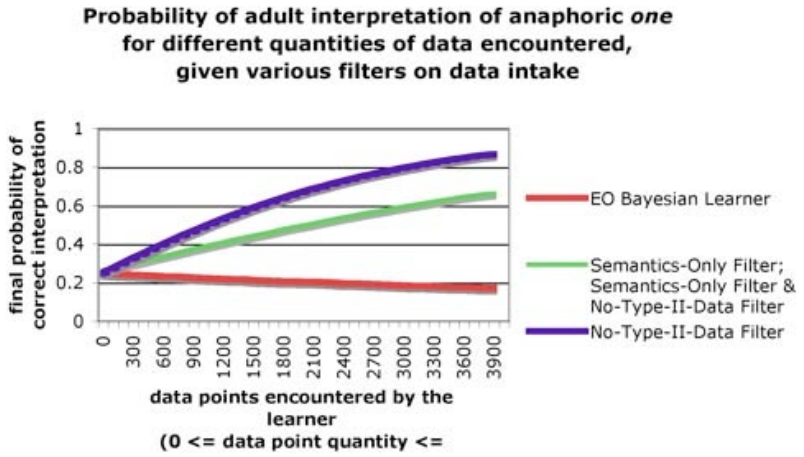


Figure 21. The Bayesian Learner's trajectory as a function of the amount of data encountered: no filters, no-type-II-data filter, semantics-only filter, and both semantics-only filter and no-type-II-data filter.

8. Feasibility of implementing the necessary filter

The necessity of a filter on data intake now raises an important question. Where does this filter come from? It seems fairly obvious that the learner cannot come equipped with a filter that says "ignore type II ambiguous data" without some procedure for identifying this data.
What we really want to know is whether there is a principled way to derive the existence of this filter. Specifically, we want the filter that ignores type II ambiguous data to be a consequence of

some other principled learning strategy. In this way, the necessary filter would become feasible for a learner to implement.

Suppose there is a general principle that learning occurs only in cases of uncertainty, because it is only in cases of uncertainty that information is conveyed (Shannon, 1948; cf. Gallistel, 2001). The learning algorithm therefore engages only when there is uncertainty about the identity of the antecedent.

One suggestion would be to call on the semantics-only filter, arguing that interpreting anaphoric *one* is simply a semantic problem. This could be termed a semantocentric approach to learning, and so the syntactic implications are irrelevant for learning. The result of this strategy would be that the learner only uses the semantic consequences of the data to update the hypotheses. As we saw in the previous section, this would rule out type II ambiguous data (with a single noun as potential antecedent, such as *ball*), because such data has only one semantic interpretation available (any-property)– thus, there is no uncertainty. However, as we also saw in the previous section, this causes the learner to lose the useful effect that the *syntactic* data can have. Specifically, if only semantic data are used, the benefit gained from having linked domains is lost. The learner uses only semantic data to update both hypothesis spaces; the learner does not also use the syntactic aspect of the data to update both hypothesis spaces. This leads to a lower probability of the adult interpretation of anaphoric *one*.

Another possibility is that the learner takes a syntactocentric approach, and the problem the learner faces is solely to identify the string that is the antecedent of anaphoric *one*. The only influence semantic interpretation data has is as a reflection of various syntactic hypotheses that are entertained. Suppose that the learner comes equipped with a constraint against anaphora to X^0 categories (Baker, 1979; Hornstein & Lightfoot, 1981) or is able to have derived it previously using a syntactocentric filter on the available data (Foraker et al., 2007). The syntactic hypothesis space is reduced to a single hypothesis: *one* = N'. In this situation, the learner needs only to solve a different problem in the syntax domain: namely, which N' is the appropriate antecedent in cases in which there are multiple N's available.

For example, if the learner hears "Here's a red ball. Give me another one, please," there are two N's available, *red ball* and *ball*. These two different antecedents have different semantic interpretations: *red ball* is restricted to red balls whereas *ball* is not. In other words, the N'-property hypothesis is linked to the larger N' *red ball*, whereas the any-property hypothesis is linked to the smaller N' *ball*. Choosing the appropriate antecedent can be achieved using the update functions described for the EO Bayesian learner.

Now, in cases in which there is only one N' available (as in type II ambiguous data), there are no choices to be made in finding an antecedent. That is, if the learner hears, "Here's a ball. Give me another one, please," the only possible antecedent is the N' *ball*. Consequently, the learner has no uncertainty about the meaning of the expression and so does not invoke the learning algorithm.

This last point is critical for motivating the learner's choice to ignore type II ambiguous data. As noted above, having a range of available antecedents causes uncertainty about the antecedent. It is this uncertainty that triggers the learning algorithm. It is important to see at this point that this syntactocentric approach requires the learner to be concerned not with the category of the antecedent (N' vs. N^0), but rather the identity of the antecedent when there are two or more N's to choose from. However, allowing the learner to view this as a problem of which syntactic antecedent to choose rather then merely as a problem of interpretation causes the learner to use the syntactic aspect of the data as well, which we found was crucial for a more successful learner.

We note that the syntactocentric learner succeeds because the acquisition of anaphoric *one* is effectively partitioned into two stages, as opposed to a single stage. In the EO Bayesian learner formulation, the learner attempts to learn both the syntactic category of anaphoric *one* (and its antecedent) and the properties of its referent in the world at the same time. This is shown below in figure 22.

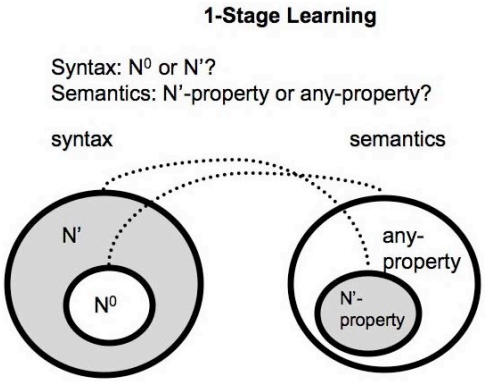


Figure 22. The EO Bayesian learner formulation of anaphoric *one* acquisition: 1-stage learning.

The syntactocentric learner, in contrast, has two stages of learning. In the first stage, the learner determines the syntactic category of anaphoric *one* and its antecedent. This may be solved via a constraint against anaphora to X^0 categories (Baker, 1979; Hornstein & Lightfoot, 1981), in which case the first stage is effectively instantaneous. This problem may also be solved via a syntactocentric filter on the available data (Foraker et al, in press), which will presumably take some time. In the second stage, the learner determines which N' should be chosen when encountering a data point with more than one option by looking at properties that the referent of *one* has. This can be solved via the Bayesian updating methods described in the previous section.

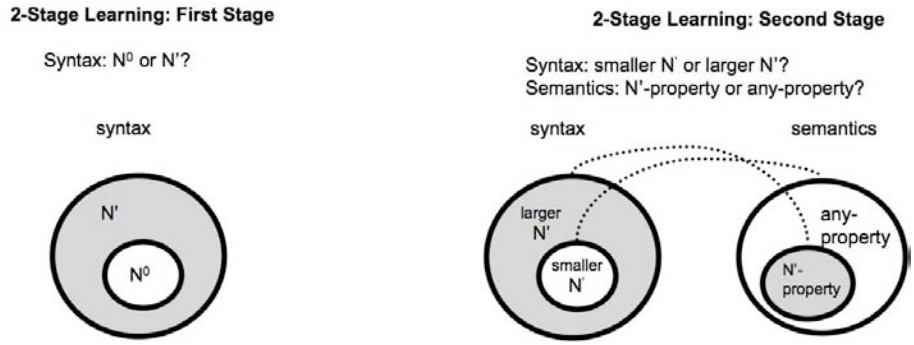


Figure 23. The syntactocentric learner partitions the acquisition of anaphoric *one* into two stages. The first stage solves the problem of the category of *one* (and its antecedent). The second stage solves the problem of which N' *one* refers to by using information available in the referent of anaphoric *one*.

Recall that we have posited a syntactocentric Bayesian learner who uses the domain-general learning strategy of "learn only in cases of uncertainty". By partitioning the learning problem into two stages, this learner will then implement the necessary filter on the input (ignore

type II ambiguous data). The intake will consist only of the data points where there is uncertainty in the syntactic antecedent of *one*, namely the unambiguous and type I ambiguous data points. The syntactocentric learner will not be led astray by the type II ambiguous data points because they will be uninformative in the second learning stage. Crucially, this filter does not need to explicitly be built into the learner's knowledge. So, in summary, the learner can implement a domain-specific filter (ignore type II ambiguous data) by using a domain-general learning strategy (learn in cases of uncertainty), provided there is a domain-specific bias on how to view the learning problem (syntactocentric).

9. Conclusion

The case of anaphoric *one* demonstrates the interplay between domain-specificity and domain-generality in learning. What we have seen here is that a domain-general learning procedure can be successful, but crucially only when paired with domain-specific filters on data intake. Moreover, we have suggested that the particular domain-specific filter that yields the best result can plausibly be derived from a domain-specific bias on learning.

In examining whether a given learning problem requires domain-specific guidance from the learner, it is important to separate three ways that a learner can exhibit domain-specificity. Learners may be constrained in the representations of the domain, the data that they deem relevant, or in the procedures used for updating their knowledge. Any one of these by itself represents a kind of domain-specific constraint on learning, and solutions to learning problems may be only partially domain-specific, as we have seen in this case study.

The division of the learning theory into distinct components, that can in their own right be domain-specific or domain-general, is important. The debate about linguistic nativism typically takes it as an all or nothing proposition. Put bluntly, the standard arguments hold either that language learning is the consequence of general-purpose learning mechanisms or that it isn't. This is a false dichotomy. What we have shown here is that the solution to the language learning problem can be more nuanced, drawing on the strengths of both domain-specific and domain-general learning components.

In addition, we have tried to highlight the consequences associated with the existence of multiple, connected levels of representation in language. Because the levels of representation are linked to each other, conclusions drawn by the learner in one domain also ramify in other domains. When the learner used both syntactic and semantic information with no filters, the result was very poor learning. When the learner used both syntactic and semantic information, in concert with the no-type-II-data filter, the result was very good learning. However, when we disconnected the two domains, as when the learner learned only from semantic data, the result was learning that was not as good (though still much better than no filtering of the data at all). This was due to some of the available information – the syntactic implications of the data – being ignored. Thus, the connection between domains allows multiple analyses of a single data point across domains to each have an effect. This, in turn, magnifies the effect of a given data point, thus increasing the amount of information that can be extracted by the learner. This lesson should be generalized to learning in any situation involving multiple linked levels of representation.

It is important to recognize that we have simulated learning only for one very specific case of grammar acquisition. However, the inherent semantic compositionality of syntactic representations provides a severe hurdle for Bayesian learning techniques that are biased towards the most restrictive hypothesis. As we have noted, as the syntactic structure grows, the set of referents in the semantics shrinks. Consequently, the most restrictive hypothesis in the syntax

corresponds to the least restrictive hypothesis in the semantic interpretation, and vice versa. This makes it impossible to define a "most restrictive hypothesis" across both domains.

The existence of multiple, linked levels of representation in language, and presumably elsewhere in cognition, has important consequences for learning. A link between domains can amplify the positive effects that come from using data from multiple sources. Nonetheless, this link can structure the data in such a way as to nullify the essential advantage of learning via indirect negative evidence.

Finally, we have emphasized the efficacy of data intake filtering on learners. Filtering the data is, in some sense, a counterintuitive approach to learning because it discards potentially informative data. Moreover, eliminating data can lead to a data sparseness problem. However, in order to find the correct generalizations in the data in our case, we found that eliminating some data was more effective than using it all. The right generalizations are hiding in the data, but paying attention to all of the data will make them harder to find. Finding them can be easier if the data considered relevant are restricted to a highly informative subset.

Acknowledgement

This paper has benefited from discussion with Amy Weinberg, Norbert Hornstein, Colin Phillips, Sandy Waxman, Bill Idsardi, Terry Regier, Charles Yang, Gerry Altmann, the Psycholinguistics-Acquisition group at the University of Maryland, the Cognitive Neuroscience of Language lab at the University of Maryland, the Center for Language Sciences at the University of Rochester, and three anonymous reviewers. Needless to say, any remaining errors are our own. This work was supported by an NSF graduate fellowship to LP, NSF grant BCS-0418309 to JL and NIH grant R03-DC006829 to JL.

A.1. Appendix A

We demonstrate in this section how we derive the update functions for the hypotheses in the syntactic and semantic domains. The syntactic update function changes $p_{N'}$, which is the probability that the N' hypothesis is correct (that the linguistic antecedent of *one* is an N' constituent). The semantic update function changes $p_{N'\text{-prop}}$, which is the probability that the N'-property hypothesis is correct (that the referent of *one* has the property mentioned in the linguistic antecedent).

A.1.1. Syntax

The update function for $p_{N'}$ depends on the data type observed: unambiguous, type I ambiguous, or type II ambiguous. We derive the update function for each data type below.

A.1.1.1. Unambiguous Data

Because there are only 2 hypotheses in the syntactic domain (N' and N^0), we use a binomial distribution to approximate a learner's expectation of the distribution of the data to be observed. The binomial distribution is centered at $p_{N'}$, so the learner's expectation is about how many unambiguous N' data points should be observed.

The binomial distribution is normally used to represent the likelihood of seeing *r* data points out of *t* total with some property. There are only two choices for each data point: the

property is either present or absent. For the syntactic domain, the "property" is being an unambiguous N' data point (as opposed to being an unambiguous N^0 data point). The highest confidence is assigned to the distribution where *r* unambiguous N' data points are observed out of *t* unambiguous data points total. We calculate *r* by multiplying *t* by $p_{N'}$, the probability that the binomial distribution is centered at: $r = t * p_{N'}$.

As an example, suppose $p_{N'}$ is 0.5, as it is in the initial state where the learner assigns equal probability to both the N' and the N^0 hypothesis. The binomial distribution is centered at 0.5, and the learner is most confident that $r = t*0.5$ data points of those observed will be unambiguous N' data points. Thus, with $p_{N'} = 0.5$, the learner expects half the total data points to be unambiguous N' data points.

To update $p_{N'}$ after seeing a single unambiguous data point *u*, we adapt Manning & Schütze's (1999) Bayesian updating algorithm and calculate the maximum of the a posteriori (MAP) probability.[21] We begin with the a posteriori probability of $p_{N'}$, which is the probability of $p_{N'}$ after seeing an unambiguous data point *u*. We represent this as $\text{Prob}(p_{N'}|u)$, and calculate it using Bayes' rule:

$$\text{(A1)}\quad \text{Prob}(p_{N'} \mid u) = \frac{\text{Prob}(u \mid p_{N'}) * \text{Prob}(p_{N'})}{\text{Prob}(u)}$$

We now describe the individual pieces of the right hand side of the equation in (A1). $\text{Prob}(u \mid p_{N'})$ is the probability of observing the unambiguous N' data point *u*, given the expected probability of observing an unambiguous N' data point. The expected probability is $p_{N'}$, so the probability of observing *u* is simply $p_{N'}$. Therefore, $\text{Prob}(u \mid p_{N'}) = p_{N'}$.

$\text{Prob}(p_{N'})$ is the probability that $p_{N'}$ is the correct probability to center the binomial distribution at, i.e. that $p_{N'}$ is the correct probability that an unambiguous N' data point will be observed in a distribution that consists entirely of unambiguous data points. Recall that the binomial distribution centered at $p_{N'}$ will assign the highest confidence to the situation where $r = (p_{N'} * t)$ unambiguous N' data points are seen out of *t* unambiguous data points total. We instantiate $\text{Prob}(p_{N'})$ as the probability of observing *r* unambiguous N' data points out of *t* total in a binomial distribution for *all* values of *r*, from 0 to *t*.

$$\text{(A2)}\quad \text{Prob}(p_{N'}) = \binom{t}{r} * p_{N'}^{\,r} * (1 - p_{N'})^{t-r} \quad (\text{for each } r,\ 0 \le r \le t)$$

Substituting these pieces back into equation (A1) for the a posteriori probability, we obtain the equation in (A3).

$$\text{(A3)}\quad \text{Prob}(p_{N'} \mid u) = \frac{p_{N'} * \binom{t}{r} * p_{N'}^{\,r} * (1 - p_{N'})^{t-r}}{\text{Prob}(u)} \quad (\text{for each } r,\ 0 \le r \le t)$$

We can now calculate the MAP probability, by finding the maximum of this equation. To do this, we take the derivative with respect to $p_{N'}$, set it equal to 0, and solve for $p_{N'}$.

[21] We note that this implementation assumes the learner only extracts information from the data point at hand, rather than storing the data points individually and conducting a collective analysis on them later.

(A4) Calculating the MAP probability

$$\frac{d}{dp_{N'}}(\text{Prob}(p_{N'} \mid u) = \frac{d}{dp_{N'}}(\frac{p_{N'} * \binom{t}{r} * p_{N'}^{r} * (1 - p_{N'})^{t-r}}{\text{Prob}(u)}) = 0$$

$$\frac{d}{dp_{N'}}(\frac{p_{N'} * \binom{t}{r} * p_{N'}^{r} * (1 - p_{N'})^{t-r}}{\text{Prob}(u)}) = 0 \text{ (since Prob}(u) \text{ is a constant w.r.t. } p_{N'})$$

$$p_{N'} = \frac{r+1}{t+1}$$

Recall that *r* is the previous expected number of unambiguous N' data points observed out of *t* unambiguous data points total. Hence, $r = p_{N' \text{ old}} * t$. Therefore, we write the update function for $p_{N'}$ after observing unambiguous N' data point *u* as (A5).

(A5) Unambiguous data update function

$$p_{N'} = \frac{p_{N' \text{ old}} * t + 1}{t + 1}$$

An intuitive interpretation of the update function is that the numerator represents the learner's confidence that the observed unambiguous N' data point *u* is a result of the N' hypothesis being correct; the denominator represents the total data observed so far. Thus, 1 is added to the numerator because the learner is fully confident that *u* indicates the N' hypothesis is correct; 1 is added to the denominator because a single data point has been observed.

A.1.1.2. Type I Ambiguous Data

The derivation for the type I ambiguous data update function is identical, since we allow the Bayesian learner to treat these data as unambiguous for the N' hypothesis even though they are, in fact, ambiguous. This will lead to an overestimation of the probability an EO Bayesian learner would assign the N' hypothesis. As we mentioned before, even with this overestimation, the learner will fail to assign sufficient probability to the N' hypothesis.

A.1.1.3. Type II Ambiguous Data

The type II ambiguous data update function is quite similar, with the exception that a value smaller than 1 is added to the numerator. Intuitively, this smaller value represents the learner's smaller confidence that the ambiguous data point *a* indicates that the N' hypothesis is correct. We call this smaller value the partial confidence value, and represent it as $p_{N' \mid a}$. Note that this is where information from the hypothesis space layout is layered into the binomially-based model just described. The coin-flip analogy associated with the binomial distribution no longer transparently applies since the data used by the learner are ambiguous, rather than unambiguous.

(A6) Type II ambiguous data update function

$$p_{N'} = \frac{p_{N' \text{ old}} * t + p_{N' \mid a}}{t + 1}$$

The partial confidence value is the probability that *one* is anaphoric to N' in *a*. This is equivalent to the probability that *one* is anaphoric to N' in general, given that *a* has been observed. We write it as Prob(N' | *a*) and calculate it by using Bayes' rule.

$$\text{(A7)} \quad \text{Prob}(\text{N'}|\, a) = \frac{\text{Prob}(a \,|\, \text{N'}) * \text{Prob}(\text{N'})}{\text{Prob}(a)}$$

We now describe the individual pieces of the right hand side of the equation in (A7). Prob(*a* | N') is the probability of observing a type II ambiguous data point *a*, given that the N' hypothesis is true. Recall that a type II ambiguous data point has an utterance with a noun-only antecedent, such as "…ball…one…". The N' hypothesis states that the linguistic antecedent of *one* must be an N' constituent.

It is possible for a noun-only string to be an N' constituent: this is the situation where a noun-only string is chosen from the set of N' constituents, which consists of both noun-only strings ("ball", "bottle", etc.) and other strings that include modifiers ("red ball", "bottle in the corner", etc.). The probability we want is the probability of choosing a noun-only linguistic antecedent for *one* (such as in type II ambiguous utterance *a*), given the entire set of N' constituents. Suppose there are *n* noun-only strings and *o* other strings in the N' constituent set. We refer to the probability of choosing a noun-only string (such as "ball") as $p_{\text{n from N'}}$, and it is calculated below in (A8).

$$\text{(A8)} \ \text{Prob}(a \mid \text{N'}) = \frac{n}{n+o} = p_{\text{n from N'}}$$

Prob(N') is the current probability that the N' hypothesis is correct. This is simply $p_{\text{N'}}$.

Prob(*a*) is the probability of observing a type II ambiguous utterance *a*, no matter which hypothesis is correct. To calculate this value, we sum the conditional probabilities of observing *a* for each hypothesis. If N' is the correct hypothesis, the probability of observing *a* is Prob(*a* | N') from above. If N^0 is the correct hypothesis, then the linguistic antecedent of *one* is an N^0 constituent, which is always a noun. Thus, the probability of observing a noun-only linguistic antecedent (such as in *a*) is 1. We calculate Prob(*a*) in (A9).

$$\begin{aligned}\text{(A9) Prob}(a) &= \sum_{\textit{hypotheses}} p_{\textit{hypothesis}} * p(a \mid p_{\textit{hypothesis}}) \\ &= p_{\text{N'}} * p(a \mid p_{\text{N'}}) + p_{\text{N0}} * p(a \mid p_{\text{N0}}) \\ &= p_{\text{N'}} * \frac{n}{n+o} + (1 - p_{\text{N'}}) * 1\end{aligned}$$

Substituting these pieces back into the right hand side of the equation in (A7), we obtain (A10).

$$\text{(A10)} \ \text{Prob}(\text{N'}|\, a) = \frac{(\frac{n}{n+o}) * p_{\text{N'}}}{p_{\text{N'}} * (\frac{n}{n+o}) + (1 - p_{\text{N'}}) * 1} = \frac{p_{\text{n from N'}} * p_{\text{N'}}}{p_{\text{N'}} * p_{\text{n from N'}} + (1 - p_{\text{N'}}) * 1} = p_{\text{N'} \mid a}$$

As we can see, the partial confidence value $p_{\text{N'} \mid a}$ depends only on $p_{\text{n from N'}}$ and the current $p_{\text{N'}}$. This partial confidence value, which will be less than 1, is added to the numerator of the

type II ambiguous data update function instead of 1. The larger $p_{n \text{ from } N'}$ is, the less biased the learner's confidence is towards the subset N^0 hypothesis when a type II ambiguous data point is observed. This is because a higher $p_{n \text{ from } N'}$ signals that the superset N' is not much larger than the subset N^0. So, the learner is not heavily biased towards the subset because the likelihood of choosing data point *a* from the subset is not much higher than the likelihood of choosing data point *a* from the superset. Thus, the more likely it is that a noun-only string could be chosen from the N' constituent set, the less the N' hypothesis is penalized when this type of data is seen.

A.1.2. Semantics

The update function for $p_{N'\text{-prop}}$ also depends on the data type observed: same-property, different-property, or unknown-property. We derive the update function for each data type below.

A.1.2.1. Same-Property Data

Because there are only 2 hypotheses in the semantic domain (N'-property and any-property), we again use a binomial distribution to approximate a learner's expectation of the distribution of the data to be observed. The binomial distribution is centered at $p_{N'\text{-prop}}$, so the learner's expectation is about how many unambiguous N'-property data points should be observed.

Again, the binomial distribution is normally used to represent the likelihood of seeing *r* data points out of *t* total with some property. In the semantic domain, the "property" is being an unambiguous N'-property data point (as opposed to being an unambiguous any-property data point). The highest confidence is assigned to the distribution where *r* unambiguous N'-property data points are observed out of *t* unambiguous data points total. We calculate *r* by multiplying *t* by $p_{N'\text{-prop}}$, the probability that the binomial distribution is centered at: $r = t * p_{N'\text{-prop}}$.

As an example, suppose $p_{N'\text{-prop}}$ is 0.5, as it is in the initial state where the learner assigns equal probability to both the N'-property and the any-property hypothesis. The binomial distribution is centered at 0.5, and the learner is most confident that $r = t*0.5$ unambiguous data points of those observed will be unambiguous N'-property data points. Thus, with $p_{N'} = 0.5$, the learner expects half the total unambiguous data points to be N'-property data points.

To update $p_{N'}$ after seeing a single same-property data point *s*, we again follow an adapted version of Manning & Schütze's (1999) Bayesian updating algorithm and calculate the maximum of the a posteriori (MAP) probability. Like the type II ambiguous data update function in the syntactic domain, however, we will add a value smaller than 1 to the numerator. Intuitively, this smaller value represents the learner's smaller confidence that the same-property data point *s* indicates that the N'-property hypothesis is correct. We call this smaller value the partial confidence value, and represent it as $p_{N'\text{-prop} \mid s}$. Again note that this is where information from the hypothesis space layout is layered into the binomially-based model just described. And again, the coin-flip analogy associated with the binomial distribution no longer transparently applies since the data used by the learner are ambiguous, rather than unambiguous.

(A11) Same-property data update function

$$p_{N'\text{-prop}} = \frac{p_{N'\text{-prop - old}} * t + p_{N'\text{-prop} \mid s}}{t+1}$$

The partial confidence value is the probability that the referent of *one* has the N'-property mentioned in *s*. This is equivalent to the probability that the referent of *one* has the N'-property

in general, given that *s* has been observed. We write it as Prob(N'-prop | *s*) and calculate it by using Bayes' rule.

$$\text{(A12)}\quad \text{Prob(N'-prop}|\, s) = \frac{\text{Prob}(s \mid \text{N'-prop}) * \text{Prob(N'-prop)}}{\text{Prob}(s)}$$

We now describe the individual pieces of the right hand side of the equation in (A12). Prob(*s* | N'-prop) is the probability of observing a same-property data point *s*, given that the N'-property hypothesis is true. Recall that in a same-property data point, the referent of the antecedent of *one* has the same salient property that the referent of *one* has. The N'-property hypothesis states that the referent of the antecedent of *one* must have the property described by the linguistic antecedent of *one*. Therefore, if the N'-property hypothesis is true, the probability of observing a same-property data point is 1.

(A13) Prob(*s*|N'-prop) = 1

Prob(N'-prop) is the current probability that the N'-property hypothesis is correct. This is simply $p_{\text{N'-prop}}$.

Prob(*s*) is the probability of observing a same-property utterance *s*, no matter which hypothesis is correct. To calculate this value, we sum the conditional probabilities of observing *s* for each hypothesis. If N'-property is the correct hypothesis, the probability of observing *s* is Prob(*s* | N'-prop) from above. If any-property is the correct hypothesis, then there is no restriction on what property the referent of the linguistic antecedent of *one* has. The probability of that referent having the same property as the referent of *one* is simply 1/*c*, where there are *c* properties in the world that the learner is considering. We calculate Prob(*s*) in (A14).

$$\begin{aligned}\text{(A14) Prob}(s) &= \sum_{hypotheses} p_{hypothesis} * p(s \mid p_{hypothesis}) \\ &= p_{\text{N'-prop}} * p(s \mid p_{\text{N'-prop}}) + p_{\text{any-prop}} * p(s \mid p_{\text{any-prop}}) \\ &= p_{\text{N'-prop}} * 1 + (1 - p_{\text{N'-prop}}) * \frac{1}{c}\end{aligned}$$

Substituting these pieces back into the right hand side of the equation in (A12), we obtain (A15).

$$\text{(A15) Prob(N'-prop}|\, s) = \frac{1 * p_{\text{N'-prop}}}{p_{\text{N'-prop}} * 1 + (1 - p_{\text{N'-prop}}) * \frac{1}{c}} = \frac{p_{\text{N'-prop}}}{p_{\text{N'-prop}} * + \frac{(1 - p_{\text{N'-prop}})}{c}} = p_{\text{N'-prop} \mid s}$$

As we can see, the partial confidence value $p_{\text{N'-prop} \mid s}$ depends only on *c* and $p_{\text{N'-prop}}$. This partial confidence value, which will be less than 1, is added to the numerator of the same-property data update function instead of 1. The larger *c* is, the higher the learner's confidence in the N'-property hypothesis when a same-property data point is observed. Thus, the more properties there are in the learner's world, the more the N'-property hypothesis is rewarded when this type of data is seen. As for the denominator of the update function, we add 1 because a single data point has been observed.

A.2. Appendix B

The model implemented by R&G is quite liberal about shifting probability to the superset hypothesis: a *single* piece of data for the exclusive superset is enough to shift *all* the probability to that hypothesis. However, as we have seen, the correct hypothesis for English anaphoric *one* is in the subset in the semantic domain: the learner should prefer the larger N' constituent, e.g. *red ball*, and thus restrict referents to those that have the N'-property, e.g. red balls. The success of this learner for converging on the correct semantic hypothesis for anaphoric *one* relies on the assumption that there will never be unambiguous data for the semantic superset.

Recall that the semantic superset hypothesis is that *one* refers to an object that does not need to have the property mentioned in the linguistic antecedent. This is the any-property hypothesis. Unambiguous data for the superset would be an utterance where *one* refers to an object that does *not* have the property mentioned in the antecedent. For instance, if the utterance is "…red ball…one…", unambiguous superset data would be the situation where the referent of *one* does not have the property 'red', e.g. it is a purple ball.

It is crucial for R&G's model that this type of data never occur, though it is entirely possible that the learner might encounter this type of data as noise or in a very specific pragmatically biased situation. If the referent of *one* in the above utterance was a purple ball (perhaps by accident), the new probability for the subset hypothesis (the N'-property hypothesis) in the semantic domain would be 0. We detail why this occurs below.

Suppose that we refer to the probability that the N'-property hypothesis is correct as $p_{N'\text{-}prop}$. Suppose the learner initially has no bias for either semantic hypothesis, and so the initial probability of $p_{N'\text{-}prop}$ is 0.5 before any data is encountered. This probability will increase as each piece of ambiguous (subset) data is observed, due to the size principle which biases the learner to favor the subset hypothesis if ambiguous data is observed.

Let *u* be a piece of unambiguous data for the superset hypothesis, where the utterance is "…red ball….one…" and the referent of *one* is a non-red ball. The learner now calculates the updated probability that the N'-property hypothesis is correct, using Bayes' rule. The updated $p_{N'\text{-}prop}$ given the observation of *u* is represented as the conditional probability p(N'-prop| *u*). To calculate this probability, we use Bayes' rule.

(A16) Calculating the conditional probability p(N'-prop| *u*) using Bayes' rule

$$p(\text{N'-prop} \mid u) \propto p(u \mid \text{N'-prop}) * p(u)$$

The probability p(*u*|N'-prop) is the likelihood of observing the unambiguous superset data *u*, given that the N'-property hypothesis is true. In this case, the referent of *one* in *u* specifically doesn't have the N'-property ('red'). Therefore, it could not possibly be generated if the N'-property hypothesis was true, since the N'-property hypothesis requires the referent of *one* to have the property mentioned in the linguistic antecedent. So, the probability of observing *u* if the N'-property hypothesis is true (p(*u*|N'-prop)) is 0.

We substitute this value into the equation in (A17) to get $p(\text{N'-prop} \mid u) \propto 0 * p(u) = 0$. Therefore, the updated probability for $p_{N'\text{-prop}}$ after seeing a single piece of unambiguous superset data *u* is 0, no matter what the previous probability of $p_{N'\text{-prop}}$ was.

Since this is not terribly robust behavior for a learner, we have adapted the Bayesian updating approach described by Manning & Schütze (1999) to generate a more conservative Bayesian updating approach, detailed in Appendix A. Unlike the liberal model implemented in R&G, the learner using this more conservative approach shifts probability much more slowly

between hypotheses. Only after observing a vast majority of evidence for one hypothesis would a conservative Bayesian learner shift the vast majority of the probability into that hypothesis.[22]

A.3. Appendix C

Recall that our model contains a parameter, *t*, which represents the amount of change the learner can undergo in the course of learning. We quantify this parameter as the number of data points the learner can use to update its probabilities. In our simulation, this was 4017. However, one might be concerned that the value of *t* might play a critical role in determining the final probability of converging on the correct interpretation of anaphoric *one* for the EO Bayesian learner. Below, we show the EO Bayesian Learner's final probability of converging on the correct interpretation of anaphoric *one* as a function of the size of *t*. As we can see, the final value does not appreciably alter based on the size of *t*.

The reason for this stability is that the behavior of the learner is dependent on the probability distribution of the data. In case *t* is small, each data point has a larger impact. In case *t* is large, each data point has a smaller impact. But, because the probability distribution is always the same, the learner always ends up with the same value so long as *t* is equal to the number of data points in the learning period. Moreover, if the learner encounters data after having seen *t* amount of data, this data cannot be used to update the probabilities.

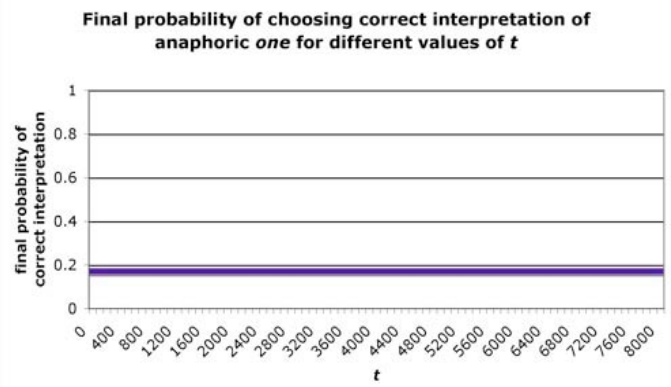


Figure A1. Final probability of the correct grammar, given different values of *t*. All values are approximately 0.171.

References

Akhtar, N., Callanan, M., Pullum, G., & Scholz, B. (2004). Learning antecedents for anaphoric *one*. *Cognition, 93,* 141-145.

Baker, C. L. (1979). *Syntactic theory and the projection problem*. Linguistic Inquiry, 10, 533-81.

Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Berwick, R. and Weinberg, A. (1984). *The Grammatical Basis of Linguistic Performance:*

[22] However, note that the conservative Bayesian updating model implemented here, while protecting against the impact of unambiguous superset data, will never reach either endpoint (0.0 or 1.0), except in the limit.

Language Use and Acquisition. Cambridge, MA: MIT Press.

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

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement and uncertainty, *Cognition, 58,* 1-73.

Dale, P.S. & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127.

Fodor, J.D. (1998). Parsing to Learn. *Journal of Psycholinguistic Research, 27*(3), 339-374.

Foraker, S., Regier, T., Khetarpal, A., Perfors, A., and Tenenbaum, J. (2007). Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Gallistel, C.R. (2001). Mental Representations, Psychology of. In *Encyclopedia of the social and behavioral sciences*. New York: Elsevier.

Gerken, L. (2006). Decision, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67-B74.

Goldsmith, J. & O'Brien, J. (2006). Learning Inflectional Classes. *Language Learning and Development*, *2*(4), 219-250.

Golinkoff, R.M., Hirsh-Pasek, K., Cauley, K.M., Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language, 14,* 23-45.

Hornstein, N., & Lightfoot, D. (1981). *Explanation in linguistics: the logical problem of language acquisition*. London: Longmans.

Legate, J. & Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review, 19,* 151-162.

Lidz, J. & Waxman, S. (2004). Reaffirming the poverty of the stimulus argument: a reply to the replies. *Cognition, 93,* 157-165.

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. (1991). *How to Set Parameters: arguments from language change*, Cambridge, MA: MIT Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The MIT Press.

Manzini, R. and Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry* 18.3, 413-444.

Pearl, J. (1996). Decision making under uncertainty. *ACM Computing Surveys (CSUR)*, 28.1, 89-92.

Pearl, L. (2005). The Input to Syntactic Acquisition: Solutions from Language Change Modeling, proceedings of *Second Workshop on Psychocomputational Models of Human Language Acquisition*, Ann Arbor, Michigan.

Pearl, L. and Weinberg, A. (2007). Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling, *Language Learning and Development, 3(1)*, 43-72.

Pierce. A. (1992). *Language Acquisition and Syntactic Theory: A Comparative Analysis of French and English Child Grammars*. Boston, MA: Kluwer Academic.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7*, 217-283.

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

Sakas, W.G. & Fodor, J.D. (2001). The structural triggers learner. In S. Bertolo (ed.) *Language Acquisition and Learnability*, Cambridge University Press, Cambridge, UK.

Sakas, W. & Nishimoto, E. (2002). Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Ms., CUNY: New York.

Shannon, C. (1948). A mathematical theory of communication,' *Bell System Technical Journal, 27*, 379-423 and 623-656.

Spelke, E. S. (1979). Perceiving Bimodally Specified Events in Infancy. *Developmental Psychology, 15 (6)*, pp. 626-636.

Staddon, J.E.R. (1988). Learning as Inference. In Bolles, R. and Beecher, M. (eds.), *Evolution and Learning*, Hillside, NJ: Lawrence Erlbaum.

Tenenbaum, J. & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-640.

Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309-318.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40,* 21-82.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.

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

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