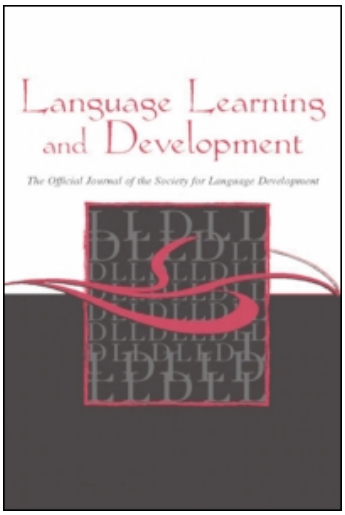

Language Learning and Development
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653671

When Domain-General Learning Fails and When It Succeeds: Identifying the Contribution of Domain Specificity

Lisa Pearl; Jeffrey Lidz [a]
[a] Linguistics Department, University of Maryland,

Online Publication Date: 01 October 2009

To cite this Article Pearl, Lisa and Lidz, Jeffrey(2009)'When Domain-General Learning Fails and When It Succeeds: Identifying the Contribution of Domain Specificity',Language Learning and Development,5:4,235 — 265
To link to this Article: DOI: 10.1080/15475440902979907
URL: http://dx.doi.org/10.1080/15475440902979907

Language Learning and Development, 5: 235–265, 2009

ISSN: 1547-5441 print / 1547-3341 online
DOI: 10.1080/15475440902979907

Psychology Press
Taylor & Francis Group

When Domain-General Learning Fails and When It Succeeds: Identifying the Contribution of Domain Specificity

Lisa Pearl
Department of Cognitive Sciences, University of California at Irvine

Jeffrey Lidz
Linguistics Department, University of Maryland

We identify three components of any learning theory: the representations, the learner's data intake, and the learning algorithm. With these in mind, we model the acquisition of the English anaphoric pronoun *one* in order to identify necessary constraints for successful acquisition, and the nature of those constraints. Whereas previous modeling efforts have succeeded by using a domain-general learning algorithm that implicitly restricts the data intake to be a subset of the input, we show that the same kind of domain-general learning algorithm fails when it does not restrict the data intake. We argue that the necessary data intake restrictions are domain-specific in nature. Thus, while a domain-general algorithm can be quite powerful, a successful learner must also rely on domain-specific learning mechanisms when learning anaphoric *one*.

1. INTRODUCTION: DOMAIN GENERALITY AND DOMAIN SPECIFICITY

Vast quantities of ink and hard feelings have been spilt and spawned 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 human 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 these two viewpoints can be successfully synthesized to explain language-learning phenomena. In the current paper, we bring this synthesis to the fore through a single, somewhat narrow, case study.

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

Correspondence should be addressed to Lisa Pearl, Department of Cognitive Sciences, 3151 Social Science Plaza A, University of California, Irvine CA 92697. E-mail: lpearl@uci.edu

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 input in terms of co-occurrence probabilities between properties of the acoustic signal, for example, 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 initially; instead, we would have domain-specific representations as the output of learning.

Second, learners must identify which data to learn from. For language learning, one might propose that only some data are used 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 was in fact due to learners having only a finite amount of working memory, causing them to use, say, only the first four words of an utterance, this would be a domain-general filter (though it operates over the domain-specific representation of words). 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, 1998a) 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 & Griffiths, 2001, among many others), then the learning algorithm is domain general.

In principle, it is an independent question for each aspect of the learning theory whether it (or a constraint on it) is domain-specific or domain-general. Although these subparts of the learning theory 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 can reconcile the opposing viewpoints on linguistic nativism. In the current paper, we provide a case study in which we show that a learner using a domain-general learning algorithm can succeed, but only when also employing domain-specific filters on the data intake. Thus, the successful learner is using both domain-specific and domain-general learning mechanisms—not just one or the other.

1.1. The Chosen Case Study

The phenomenon under investigation is the interpretation of the anaphoric element *one* in English. Two different levels of representation must be considered for interpreting anaphoric *one*: the syntactic level and the semantic level. At the syntactic level, the infant must identify the linguistic antecedent of *one*; at the semantic level, the infant must determine what object in the world a noun phrase (NP) containing *one* refers to. Both of these levels contribute to the information a learner would use when converging on the correct representation of *one*, since a

linguistic antecedent (syntax) can be translated into a reference to an object in the world (semantics), and vice versa. In terms of learning, the syntactic antecedent and the semantic reference are interconnected: the syntactic antecedent of *one* has semantic consequences on what referents are picked out, while the semantic referent has syntactic implications of what the linguistic antecedent is. Learners can thus use multiple sources of information to inform their hypotheses about the interpretation of anaphoric *one* in a given context, and use these conclusions to identify the appropriate grammatical representation underlying the use of *one*.

The case of anaphoric *one* learning has received considerable attention recently, as the linguistic knowledge determining its antecedent was previously considered unlearnable given the sparseness of unambiguous data (e.g., Baker, 1978; Hornstein & Lightfoot, 1981; Crain, 1991). However, computational modeling work has suggested that the correct representation of anaphoric *one* can be learned from a predefined hypothesis space by using a domain-general learning model that capitalizes on the relationship between the different hypotheses in the hypothesis space (Regier & Gahl, 2004, hereinafter referenced as Regier & Gahl). Using this hypothesis space information, the model is able to gain information from ambiguous data as well, and in fact can converge on the correct hypothesis for anaphoric *one* even if no unambiguous data are encountered.

Although this is a substantial achievement, the data used by Regier and Gahl's model were restricted to a subset of the available data—and so there was an implicit filter on the learner's data intake. As Regier and Gahl note (and we agree), "learning of any sort is impossible without constraints of some kind." The key question then is which constraints are necessary and whether the necessary constraints are domain-specific or domain-general. Here, we demonstrate that this particular data intake filter is, in fact, a necessary constraint for successful learning of anaphoric *one*. Further, we argue that it can be effectively implemented only by appealing to domain-specific knowledge. Anaphoric *one* is a single case study, but this general line of approach to identifying the necessary constraints on language acquisition, as well as the nature of those constraints, is one computational modeling is particularly suited for, and one that is increasingly being pursued (for recent examples, see Perfors, Tenenbaum, & Regier, 2006; Pearl, in press; Foraker, Regier, Khetarpal, Pefors, & Tenenbaum, 2009).

1.2. This Article's Plot

The remainder of the article proceeds as follows. First, we briefly describe the syntactic and semantic representations of anaphoric *one* and review the behavioral evidence indicating that 18-month-olds have acquired the adult representation of anaphoric *one*. We then assess the data an 18-month-old would have encountered in the input and note the infrequency of the available unambiguous data. Following this, we review Regier and Gahl's model that uses a domain-general learning algorithm, and we also observe that this model includes implicit constraints on data intake. To gauge the learning algorithm's performance without these implicit constraints, we extend their model so that it can learn from all available data, thereby removing the data intake constraints. We find that such a model fails to converge on the adult representations of anaphoric *one* with high probability, unlike what we see in 18-month-olds. We subsequently identify the source of the new model's failure and find that it is due to the abundance of a particular kind of ambiguous data available in the input. When this subset of ambiguous data is ignored by the learner, the model succeeds with high probability, as Regier and Gahl found in their original model. Because the unconstrained data intake is the problem, this suggests that successful learning

of anaphoric *one* from these same data requires some bias on the data intake. We conclude with some speculation on how a learner might know to ignore the troublesome data, highlighting that such behavior is likely to derive from domain-specific knowledge.

2. ANAPHORIC *ONE*

2.1. Adult Knowledge: Syntactic Representations

For English-speaking adults, the element *one* is anaphoric to strings that are classified as N' (i.e., *one*'s antecedent is an N' string) and not N^0, as in example (1) below. The structures for the N' strings are represented in Figure 1. We note that the precise labels of the constituents are immaterial; it is important only that there is a distinction between the relevant constituents, represented here as N' and N^0.

(1a) *One* is anaphoric to N' (*ball* is antecedent)

"Jack likes this *ball*. Lily likes that *one*."

(1b) *One* is anaphoric to N' (*red ball* is antecedent)

"Jack likes this *red ball*. Lily likes that *one*."

These representations encode two kinds of information: constituency structure and category structure. The constituency structure tells us that in a noun phrase containing a determiner (det: *this*), adjective (adj: *red*), and noun (N^0: *ball*), the adjective and noun (*red ball*) form a unit within the larger NP. The fact that *one* can be interpreted as a replacement for those two words, as in (1b), 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. Because *one* can replace both *ball* and *red*

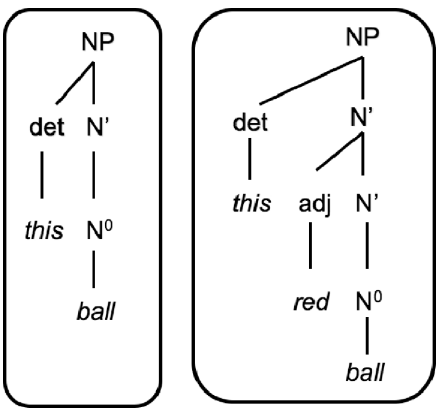

FIGURE 1 Structures for the N' strings *this ball* and *this red ball*.

ball, both *ball* and *red ball* are of the same category: type N'. See Appendix A for a more detailed argument about the category structure of *one*.

2.2. Adult Knowledge: Pragmatics and Semantic Reference

In addition, when there is more than one N' to choose from, as in (1b) above, adults prefer the N' corresponding to the longer string (*red ball*). For example, in (1b) 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, Callanan, Pullum, & Scholz, 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*, then 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. Note, however, that this preference is not categorical. It is straightforward to find cases in which it is overridden, as in (2):

(2) "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*.

2.3. Children's Knowledge: Behavioral Evidence

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 include the modifier (e.g., *red*), and hence the referent of *one* should include the properties mentioned by modifiers in the antecedent. Lidz, Waxman, and Freedman (2003, hereinafter referred to as Lidz et al.) conducted an intermodal preferential looking paradigm experiment (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987; Spelke, 1979) to see if 18-month-olds did, in fact, show this behavior.

The 18-month-olds demonstrated a significant preference for looking at the referent that had the same property mentioned in the N' string (e.g., looking at a bottle that was red when the N' string *red bottle* was a potential antecedent). Lidz et al. explained this behavior as a semantic consequence of the syntactic preference that *one* be anaphoric to the larger N' string (*red bottle*). Because infants preferred the larger N' string (as adults do) and this larger N' string could not be classified as N^0, Lidz et al. concluded that the 18-month-olds have the syntactic knowledge that *one* is anaphoric to N' strings in general. In addition, they have the adult pragmatic preference to choose the referent corresponding to the larger N' string when there is more than one N' antecedent, though this was not the focus of that work.

2.4. Data for Learning the Antecedent of Anaphoric *One*

To estimate the data available to children by 18 months, we should consider that learning the syntactic category and related semantic interpretations of *one* can commence only once the child has some repertoire of syntactic categories. We posit that the anaphoric *one* learning period begins at 14 months, based on 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 children hear approximately 1,000,000 sentences from birth until 18 months (Akhtar et al., 2004), then they should hear approximately 278,000 sentences between

14 months and 18 months. We can use the data frequencies found in Lidz et al.'s corpus analysis to estimate the expected distribution of anaphoric *one* data during this learning period. Of the 278,000 data points heard, 4,017 are likely to be anaphoric *one* data points. The expected distribution of these 4,017 data points is shown in Table 1.

All data are defined by a pairing of utterance and environment. Unambiguous data have properties similar to (3). 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*. Because *red ball* can be classified only as N', these data are unambiguous evidence that *one* is anaphoric to N'. Moreover, because the antecedent is *red ball*, and not simply *ball*, this data point highlights that the larger N' constituent should be chosen when multiple N's are available.

(3) Unambiguous example:

Utterance: "Jack wants a red ball. Lily doesn't have one for him."

Environment: Jack wants a red ball, but Lily doesn't have a red ball; she has a ball with different properties.

Type I ambiguous data have properties similar to (4a) or (4b). For data similar to (4a), Lily has a ball, so the antecedent of *one* could be *ball*. However, the ball Lily has 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. In addition, they are also ambiguous between the smaller N' constituent (*ball*) and the larger one (*red ball*), even if *one* is known to be anaphoric to N'. For data similar to (4b), 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.

(4a) Type I ambiguous example:

Utterance: "Jack wants a red ball. Lily has one for him."

Environment: Lily has a ball for Jack, and it has the property red.

TABLE 1
The Expected Distribution of Utterances in the Input to Children Between 14 and 18 Months

Data Type	# of data points
Unambiguous	10
"*Jack wants a red ball. Lily doesn't have one for him.*"	
(Lily doesn't have a ball with the property red, but she does have a ball.)	
Type I Ambiguous	183
"*Jack wants a red ball. Lily has one for him.*"	
(Lily has a red ball for Jack.)	
Type II Ambiguous	3805
"*Jack wants a ball. Lily has one for him.*"	
(Lily has a ball.)	
Ungrammatical	19
"*. . . you must be need one*."	

(4b) Another type I ambiguous example

Utterance: "Jack wants a red ball. Lily doesn't have one for him."

Environment: Lily doesn't have a ball at all.

Type II ambiguous data have properties similar to (5a) or (5b). For both forms of type II ambiguous data, the antecedent of *one* must be *ball*. However, because *ball* can be classified as either N' or N^0, such data are ambiguous with respect to what *one* is anaphoric to.

(5a) Type II ambiguous example:

Utterance: "Jack wants a ball. Lily has one for him."

Environment: Lily has a ball for Jack.

(5b) Another type II ambiguous example:

Utterance: "Jack wants a ball. Lily doesn't have one for him."

Environment: Lily does not have a ball at all.

Ungrammatical data involve a use of anaphoric *one* that is not in the adult grammar, such as in (6). Because 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 are therefore noise in the input and are presumably ignored by the child.

(6) Ungrammatical example (taken from CHILDES, MacWhinney, 2000):

Utterance: ". . . you must be need one."

3. LEARNING ANAPHORIC *ONE*

The vast majority of the anaphoric *one* input consists of type II ambiguous data (3,805 of 4,017, 94.7%), while type I ambiguous data make up a much smaller portion (183 of 4,017, 4.5%). Ungrammatical data are quite rare (19 of 4,017, 0.5%) and unambiguous data rarer still (10 of 4,017, 0.25%). Because Lidz et al. considered unambiguous data as the only informative data (as did many linguists before them), they concluded that such data were far too sparse to definitively signal to a child that *one* is anaphoric to N'.

The argument has the form of a proof by contradiction. If we assume that the hypothesis space of possible antecedents for anaphoric forms includes both heads (N^0, in this case) and phrasal categories (N' or NP), then there is not sufficient unambiguous information telling against the N^0 hypothesis. Given that no learner acquires N^0, the assumption that it was included in the hypothesis space should be rejected.

This argument based on data sparseness seems in line with theory-neutral estimations of the quantity of data required for acquisition by a certain age (Legate & Yang, 2002; Yang, 2004). Specifically, other linguistic knowledge acquired by 20 months required at least 7% unambiguous 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

child would be able to acquire the correct representation of anaphoric *one* by 18 months solely from the unambiguous data. In sum, the conclusion of this argument is that the child's learning mechanism includes a constraint against taking heads (such as N^0) as possible antecedents for anaphoric elements (such as *one*). It is this conclusion that Regier and Gahl's model calls into question.

3.1. Using Ambiguous Data

Regier and Gahl's model offers a solution to the problem posed by the sparseness of unambiguous data: learners can use a domain-general learning algorithm that extracts information from ambiguous data as well, based on the relationships between the different hypotheses the child considers for anaphoric *one*. This gives the child significantly more data to learn from. Given the utility of the ambiguous data, Regier and Gahl argue that the hypothesis space can include the hypothesis that heads such as N^0 can serve as antecedents to anaphors such as *one* after all. Thus, learners do not need a specific constraint against taking heads as antecedents as a part of an innate universal grammar. The power of Regier and Gahl's model comes specifically from learning from the type I ambiguous data. This is an attractive strategy, because 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 child up to 4.8% (193 of 4,017), which seems more in line with the amount required for acquisition as early as 18 months (Yang, 2004).

To understand how Regier and Gahl's model works, consider a type I ambiguous data point, such as the utterance "Jack wants a red ball, and Lily has one for him," paired with the situation in which Lily has a red ball. *One* could have either *red ball* (an N' string) or *ball* (an N' or N^0 string) as its syntactic antecedent. Even though the referent is the red ball Lily is holding, it is unclear whether the property red is important without knowing which string is the antecedent. If the antecedent is *red ball*, then it matters that the ball has the property red; if the antecedent is *ball*, it does not. If the antecedent is *red ball*, then the antecedent's category must be N'; otherwise, the category can be either N' or N^0. Thus, all the relevant knowledge for anaphoric *one* can be derived from knowing whether the property red is important for the referent to have.

A Bayesian learner can observe the distribution of referents of *one* for all the data points of this type in order to determine if the property red is important. If it is always the case that when there is an observable referent that referent has the property mentioned in the larger N' string (*red* in our example), then the Bayesian learner will believe it increasingly likely that the referent must have that property. This then implicates the larger N' string (*red ball*) as the antecedent, and *one*'s category as N'.

This learning strategy is captured implicitly in the Size Principle (Tenenbaum & Griffiths, 2001) that a Bayesian learner can use, and which children also seem to use (Gerken, 2006; Xu & Tenenbaum, 2007). Crucially, the Size Principle operates only when there is a certain relationship between the hypotheses the learner considers—specifically, when one hypothesis is a subset of the other. We can demonstrate the necessary subset relationship by considering the possible referents of anaphoric *one*.

For the type I ambiguous example given above, one hypothesis is that the property mentioned in the antecedent (*red*) is not important. Given this, the potential referent does not need to have that property. The set of potential referents is the set of all balls, as shown in the shaded portion of Figure 2a. The alternative hypothesis is that the property mentioned in the antecedent is

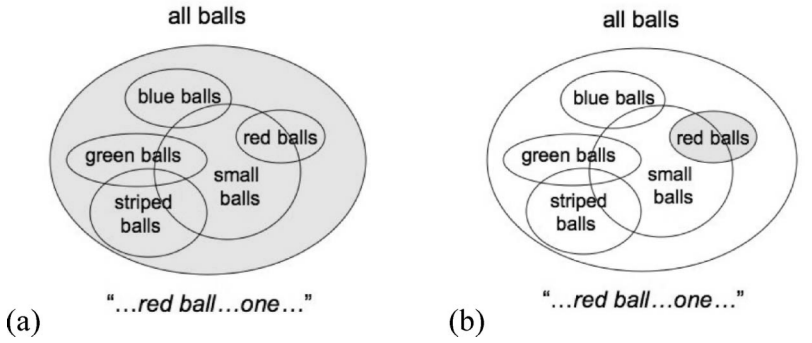

FIGURE 2 (a) The set of potential referents for *one* if the property red is not important. (b) The set of potential referents for *one* if the property red is important.

important and the potential referent must have that property. The set of potential referents is then the subset of red balls, as shown in the shaded portion of Figure 2b.

These are the Bayesian learner's expectations of what will be observed in the data, given each hypothesis. If the property red is not important, then referents both with and without the property red should be observed. However, if only referents with the property red keep being observed, then this is a conspicuous coincidence if the property red is not important. Even though it is possible that a red ball is the referent when any ball will do (the red ball is ambiguous between indicating the set of red balls vs. the set of balls), it grows increasingly improbable that only red balls are ever observed. Why wouldn't some other kind of ball appear as the referent once in a while?

More formally, it is highly unlikely that the referent of *one* is only ever a member of the subset of red balls if the referent could be any member of the superset of balls. We would expect to see at least one example of a non-red ball, which would unambiguously indicate the superset of balls is correct, and not the subset of red balls. If this continues to not happen, and only red ball referents are observed, the Bayesian learner considers a restriction to the subset of red balls to be more and more probable. This is the more precise instantiation of the size principle of Tenenbaum and Griffiths (2001): if there is a choice between a subset and the superset, and only data from the subset are seen, then the learner will be most confident that the subset is the correct hypothesis (often represented as giving the subset hypothesis a higher probability). Thus, the learner uses the lack of data for the superset as indirect evidence that the subset hypothesis is correct.

The amount the subset hypothesis is rewarded when a subset data point is observed depends on the relative sizes of the subset and superset. If the superset (BALLS) has many more members than the subset (RED BALLS), then 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 (Figure 3a). In other words, it is a conspicuous coincidence. The amount the subset hypothesis (RED BALLS) is rewarded given a subset data point (a RED BALL) in this case is then larger. In contrast, if the superset (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 (Figure 3b). This is not nearly so conspicuous

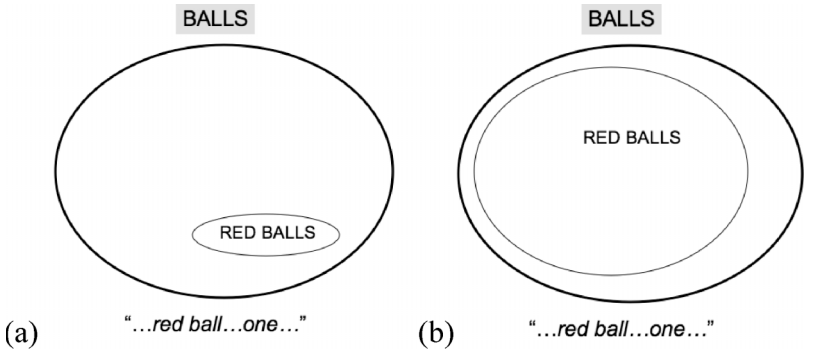

FIGURE 3 Comparison of different ratios of superset to subset hypotheses.

a coincidence. The amount the subset hypothesis (RED BALLS) is rewarded given a subset data point (RED BALL) in this case is then smaller. This effect of the size principle will become important in sections 3.4 and 4, when we consider different variants of a Bayesian learner.

By learning from both unambiguous and type I ambiguous data points, Regier and Gahl's model can indeed learn the correct referent for anaphoric *one* (ex: the red ball), which then translates to knowing the antecedent string (ex: *red ball*) and thus the category of the antecedent (N'). And it can do so without stipulating a priori that N^0 is not a possible antecedent. A great strength of Regier and Gahl's model is that the preference for the subset when observing only subset data points 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 and Griffiths (2001). This model therefore draws on a domain-general learning strategy to learn the representations for anaphoric *one*.

3.2. Using All the Ambiguous Data

Regier and Gahl's model does harbor an implicit bias about data filters on the learner's intake, however. Because Regier and Gahl's model was designed only to illustrate the utility of ambiguous data in learning, it made use only of type I ambiguous data, in other words, the data in which the semantic consequences of the alternative syntactic hypotheses stand in a subset–superset relation. In this sense, the model artificially restricted the learner's intake by ignoring the type II ambiguous data (i.e., the data in which the syntactic consequences of the alternative syntactic hypotheses stand in a subset–superset relation). These ambiguous data are potentially informative to a learner trying to determine the syntactic category of the antecedent of *one*. Regier and Gahl's decision to ignore these data was appropriate for illustrating the utility of ambiguous data. But was it also the right decision for accurately modeling the learning problem? If the model's task is to acquire the correct syntactic antecedent, then a truly unbiased learner would consider not only data that use the semantic consequences of the alternatives as information but also data that use the syntactic consequences of the alternatives as information.

To understand the importance of the syntactic consequences of the alternative hypotheses, consider Figure 4, which illustrates the set of strings covered by the two alternatives.

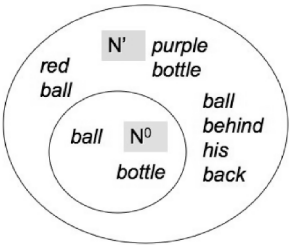

FIGURE 4 The set of strings compatible with the syntactic hypotheses of N^0 and N'.

Of note, the set of strings included in the N^0 hypothesis is a subset of the set of strings included in the N' hypothesis. For data points in which the only possible antecedent is a single word (e.g., *ball*), the size principle dictates that the N^0 hypothesis be rewarded by an increase in probability. This then also has the effect of rewarding the semantic consequence of the N^0 hypothesis, that is, not believing the property mentioned in the potential antecedent (e.g., *red*) is important. Thus, notice that the effect of the size principle in the syntactic hypothesis space works directly against the effect of the size principle we just observed in the semantic hypothesis space, which was the key factor in the success of the model discussed by Regier and Gahl.

In what follows, we extend Regier and Gahl's model and construct a Bayesian model that removes this implicit data intake filter by considering both the semantic and syntactic consequences of alternative syntactic hypotheses. Consequently, our model learns from type II ambiguous data, in addition to type I ambiguous data and unambiguous data. We can then see if this "equal opportunity" (EO) model, which gives equal treatment to all data and to all of the potential consequences of each hypothesis, fares as well when learning anaphoric *one* from the available data.

The EO model's structure reflects the generative process that could lead to each of the observable data types: unambiguous, type I ambiguous, and type II ambiguous. In this sense, the model reflects a learning strategy known as "analysis by synthesis": it learns by simulating the grammar that produced the data (Halle & Stevens, 1962; Townsend & Bever, 2001). In this model, all observable data can be generated by having the antecedent of *one* be either of category N^0 or N'. In addition, if a given data point is generated by N' and there are multiple N's to choose from, then it could be generated by either the upper N' constituent (ex: *red ball*) or the lower one (ex: *ball*). This second choice has specific semantic consequences, as mentioned previously. The learner is thus trying to determine two probabilities: (a) the probability that the current observable data point was generated by the antecedent of *one* being category N' (model parameter θ_N), and (b) if there are multiple N's to choose from, the probability that the current observable data point was generated by the upper N' constituent, which corresponds to the larger N' string (model parameter θ_U).

For unambiguous and type I ambiguous data, the syntactic information is the utterance containing anaphoric *one* and an antecedent that potentially includes a modifier (e.g., ". . . red ball . . . *one* . . .). Unambiguous data indicate that the antecedent is definitely the upper N' constituent (e.g., *red ball*), while type I ambiguous data leave open the options N^0 (*ball*), lower N' (*ball*),

and upper N' (*red ball*). The semantic information is the referent in the world; it is the object described by the noun in the antecedent (e.g., a BALL). For unambiguous and type I ambiguous data, the object will be observed to have the property mentioned in the modifier of the antecedent's noun (e.g., a RED BALL).

The decision tree in Figure 5 shows how to generate different data types from an utterance in which the potential antecedent contains a noun and a modifier (". . . red ball . . . *one* . . ."). Note that the model is capable of generating both data types we may observe in the input, corresponding to unambiguous (7e) and type 1 ambiguous (7b), (7d), (7e), as well as data types we will not observe, (7a), (7c). We follow each of the branches of the decision tree (left to right) to generate one of the following data types:

(7) Possible data types for an utterance whose potential antecedent has a modifier

(a) [antecedent = N^0, "ball", object = BALL (non-red)]
(b) [antecedent = N^0, "ball", object = RED BALL]
(c) [antecedent = N', lower N', "ball", object = BALL (non-red)]
(d) [antecedent = N', lower N', "ball", object = RED BALL]
(e) [antecedent = N', upper N', "red ball", object = RED BALL]

We first use θ_N to decide the category of *one*'s antecedent. With probability $1\text{-}\theta_N$, the category of *one*'s antecedent is N^0. The antecedent then can be only a noun (e.g., *ball*), since the strings that are of category N^0 consist only of nouns. We then move on to *one*'s referent. Suppose c represents the number of potential properties the learner is aware of (e.g., RED, PURPLE, STRIPED, BIG). The likelihood of choosing an object that, by coincidence, just happens to have the property mentioned in the potential antecedent (Mod-Prop OBJECT) when any object could have been chosen is $1/c$. The likelihood of

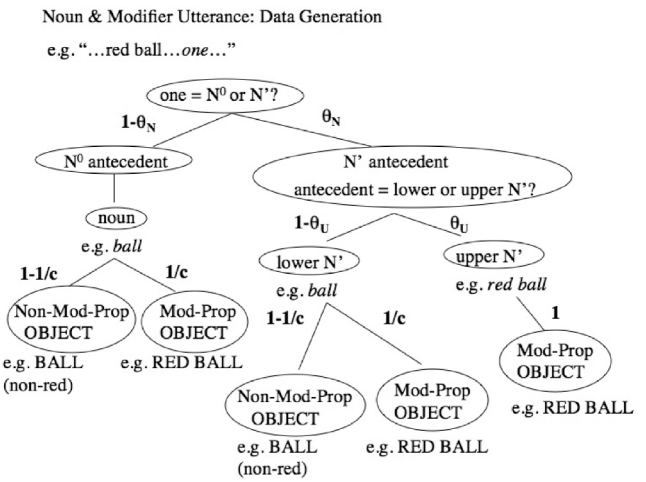

FIGURE 5 A decision tree for generating data whose utterance contains a potential antecedent with a noun and modifier (e.g., ". . . red ball . . . *one* . . .").

choosing an object that does not have the property mentioned in the potential antecedent (Non-Mod-Prop OBJECT) is then 1-1/*c*. To generate data type (7a), the Non-Mod-Prop OBJECT is the referent for *one*; to generate data type (7b), the Mod-Prop OBJECT is the referent for *one*.

Going back to the top of the decision tree, the category of *one*'s antecedent is N' with probability θ_N. There is then a choice between the upper N' constituent (*red ball*) and the lower N' constituent (*ball*) as *one*'s antecedent. We use θ_U to determine which one is chosen. With probability 1-θ_U, the lower N' constituent corresponding to the smaller N' string (*ball*) is chosen. Because the antecedent is simply the noun, 1/*c* is again the likelihood of just happening to choose the object with the property mentioned in the antecedent's modifier (Mod-Prop OBJECT); 1 – 1/*c* is the likelihood of choosing an object with some other property (Non-Mod-Prop OBJECT). To generate data type (7c), the Non-Mod-Prop OBJECT is the referent for *one*; to generate data type (7d), the Mod-Prop OBJECT is the referent for *one*. Returning to the upper/lower constituent decision point, with probability θ_U, the upper N' constituent (*red ball*) is chosen. The object must then have the property mentioned in the modifier (*red*), and so the referent for *one* will be a Mod-Prop OBJECT, data type (7e).

For type II ambiguous data, the syntactic information is the utterance containing anaphoric *one* and a potential antecedent that has no modifier (e.g., ". . . ball . . . *one* . . ."). The antecedent's category is either N^0 or N'. The semantic information is the referent in the world; it is again the object described by the noun in the antecedent (e.g., BALL).

The decision tree in Figure 6 shows how to generate different data types from an utterance where the potential antecedent contains no modifier. Both data types generated, (8a) and (8b) correspond to type II ambiguous data. We follow each of the branches of the decision tree (left to right) to generate one of the following data types:

(8) Possible data types for an utterance whose potential antecedent has no modifier

(a) [antecedent = N^0, "ball", object = BALL]
(b) [antecedent = N', "ball", object = BALL]

We first use θ_N to decide the category of *one*'s antecedent. With probability 1-θ_N, the category of *one*'s antecedent is N^0. The antecedent then can only be a noun (e.g., *ball*), because the strings that are of category N^0 consist only of nouns. The referent must simply be an OBJECT described by the noun in the antecedent, data type (8a). Returning to the top of the decision tree, the category of *one*'s antecedent is N' with probability θ_N. Because strings of category N' can contain modifiers (e.g., *red ball*), the probability of choosing only a noun (e.g., *ball*) is less than 1. More specifically, let *n* be the number of noun-only strings of category N' and *m* be the number of strings with nouns and modifiers. The likelihood of choosing a noun-only string is then *n*/*n*+*m*. The referent of the noun-only string must then be an OBJECT described by the noun, data type (8b).

These generative trees can then be used to create precise update equations that define how a learner will change beliefs in various hypotheses about anaphoric *one* (represented by parameters θ_N and θ_U) based on the data type actually observed. We note that the equations below shift probability between hypotheses more gradually than Regier and Gahl's model did. See Appendix B for details on how Regier and Gahl's probability update equations work and why it may be

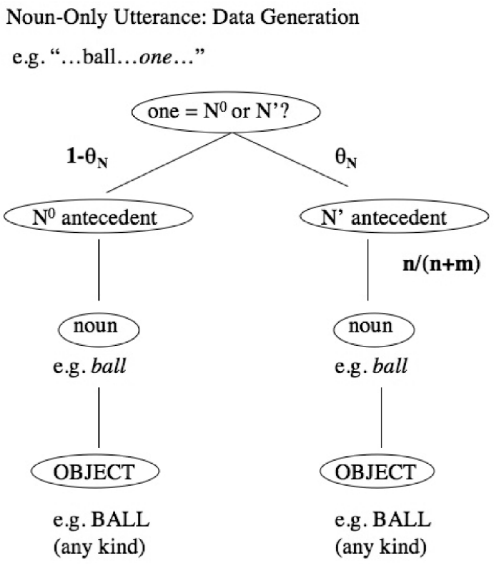

FIGURE 6 A decision tree for generating data whose utterance contains a potential antecedent without a modifier (e.g., ". . . ball . . . *one* . . .").

preferable for a learner to update more gradually. The update equations used for each data type in the EO model are in (9). See Appendix C for their derivation from the generative decision trees in Figures 5 and 6.

(9) Update equations for each observable data type

(a) Unambiguous data

$$\theta_N = \frac{\alpha + (data_N + 1)}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

$$\theta_U = \frac{\alpha + (data_U + 1)}{\alpha + \beta + (totaldata_U + 1)}, \alpha = \beta = 0.5$$

(b) Type I ambiguous data

$$\theta_N = \frac{\alpha + (data_N + l_{N-Type1})}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

$$\theta_U = \frac{\alpha + (data_U + l_{U-Type1})}{\alpha + \beta + (totaldata_U + 1)}, \alpha = \beta = 0.5$$

(c) Type II ambiguous data

$$\theta_N = \frac{\alpha + (data_N + l_{N-Type2})}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

θ_U not updated

To demonstrate these update equations, suppose the very first data point the learner sees is an unambiguous data point. Suppose that $c = 5$ (the learner knows 5 possible properties that could have been mentioned). Before updating, $\theta_N = \theta_U = 0.5$. After updating, $\theta_N = \theta_U = 0.75$. As we can see, unambiguous data are highly informative, especially if they are encountered early in the learning period.

Suppose instead that the very first data point the learner sees is a type I ambiguous point. Suppose that $c = 5$. Before updating, $\theta_N = \theta_U = 0.5$. After updating, $\theta_N = 0.625$ and $\theta_U = 0.666$. We can see that type I ambiguous data are not quite as informative as unambiguous data, but are still quite informative.

Suppose instead that the very first data point the learner sees is a type II ambiguous point. Suppose that $n = 1$ and $m = 1$. The learner knows 1 other possible N' item besides the noun-only element (e.g., *ball*), perhaps the adjective + noun element (e.g., *red ball*). Before updating, $\theta_N = \theta_U = 0.5$. After updating, $\theta_N = 0.417$ and $\theta_U = 0.5$. We can see that type II ambiguous data are potentially quite damaging to θ_N. Of course, the impact of any data point—unambiguous, type I ambiguous, or type II ambiguous—will lessen as more data come in, for instance, if the second data point the learner in this example encounters is an unambiguous one, $\theta_N = 0.611$ and $\theta_U = 0.75$.

3.3. What Good Learning Would Look Like

In the model, the learner initially sets the parameters θ_N and θ_U equal to 0.5. It is thus not biased toward either of the syntactic category hypotheses (N^0 or N'). It is also not biased toward either of the options when multiple N's are available (e.g., ". . . red ball . . . *one* . . ."), the lower N' or the upper N'. Again, these options have specific semantic consequences—the lower N' corresponds to the more general object (i.e., BALL), while the upper N' corresponds to the modifier-property object (i.e., RED BALL). The probability of choosing the preferred adult interpretation, given an utterance with two potential antecedents, depends on choosing the correct option in each case. So, if the learner hears, "Jack wants a red ball, and Lily has one for him," then the interpretation of *one* is calculated as in (10), using the generative tree in Figure 5.

(10) Interpreting *one* in "Jack wants a red ball, and Lily has one for him"

(a) Determine if the antecedent of *one* should be N^0 or N', using θ_N.
(b) If the antecedent is N^0, then the referent is any object described by the noun (e.g., BALL).
(c) If the antecedent is N', use θ_U to determine if the upper N' interpretation (referent is a modifier-property object) or the lower N' interpretation (referent is any object) should be used.

The probability of choosing the preferred adult interpretation (the larger N' constituent is the antecedent of *one* and the referent has the modifier-property) is the product of the

probability of choosing the correct category hypothesis (N') and that of choosing the correct hypothesis for the semantic interpretation (larger N' constituent = modifier-property object). Initially, this probability is 0.500 * 0.500 = 0.250. Given that the end state should be a probability near 1, a good learning model should steadily increase the probability of choosing the preferred adult interpretation.

3.4. 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 classified as unambiguous, type I ambiguous, type II ambiguous, or ungrammatical (noise). The parameters θ_N and θ_U are then updated accordingly.

3.4.1. Syntactic Category

The update equations for θ_N sometimes require the parameters *n* and *m* (type II ambiguous) or the parameter *c* (type I ambiguous). As noted previously, *n* and *m* correspond to the number of noun-only strings and noun-and-modifier strings in the N' set. The N^0 set, by definition, contains only the *n* noun-only items. The parameter *c* corresponds to the number of properties the learner believes a speaker could have chosen to utter. There are several reasonable ways to estimate what values the learner might assign to these parameters. We outline each of these below.

Type II ambiguous data are more damaging to the N' category hypothesis the larger the N' set is compared to the N^0 set (see Appendix C for details why). So, as we consider the different ways to estimate *n* and *m*, we will also consider which one generates sets that are closer in size so as to maximize the performance of the EO Bayesian Learner.

One way to estimate the relative sizes of the sets in the syntactic hypothesis space is to assume the learner is comparing the elements that could possibly be in the N^0 set against those that could possibly be in the N' set, given the generative system of language. For example, suppose the child knows that N^0 contains only nouns (*ball*) while N' contains nouns (*ball*), adjective + nouns (*red ball*), adjective + adjective + nouns (*big red ball*), and nouns + prepositional phrases (*ball behind his back*). The child might then estimate the relative sizes of these sets by what elements in the generative system of the language could possibly belong to them (N^0 contains all possible nouns; N' contains all possible nouns, all possible adjective + noun combinations, etc.). There are two reasonable ways a learner might do this. First, we could allow a string to consist of individual vocabulary items (*bottle*, *ball*, *ball behind his back*, etc.) that have been encountered previously. Alternatively, we could allow a string to consist of individual types ("Noun", "Adjective Noun", etc.) comprised of syntactic categories the child has already learned.

Data from the MacArthur CDI (Dale & Fenson, 1996) suggest that 14- to 16-month-olds know on average at least 49 adjectives and at least 247 nouns. Constructing N^0 and N' sets from individual vocabulary items, a child could posit the N^0 set has 247 elements while the N' set has (at the very least) 247 + 49*247 elements for the nouns and adjective + noun combinations. Parameter *n* would be 247, while *m* would be 49*247, for a subset-to-superset ratio of 0.020.

If we construct the N^0 and N' sets from individual types such as "Noun" rather than individual vocabulary items such as *ball*, then the N^0 set consists only of the type "Noun". The N' set contains "Noun", and perhaps (at the least) "Adjective Noun", "Adjective Adjective Noun", and "Noun Preposition Phrase", based on the input the child has encountered. Parameter *n* would be 1, while *m* would be 3, for a subset-to-superset ratio of 0.250. This ratio is considerably larger than the estimate by individual vocabulary items, and so will cause the type II ambiguous data to penalize the N' hypothesis less.

Still, there is another option for estimating the relative set sizes[1]: the learner uses only the actual data encountered, rather than all the elements that could be generated by the linguistic system. More specifically, as input comes in, the learner tracks how often the noun phrases spoken are simply determiners followed by nouns versus determiners followed by strings involving modifiers. To estimate this, we examined the Alice files in the Bernstein corpus of the CHILDES database (MacWhinney, 2000), which are transcriptions of speech to a child at ages 1;1, 1;3, and 1;5. Given 4,999 words of child-directed speech, there were 465 noun phrases, 346 of which contained only nouns. This leads to a ratio of 346/465 = 0.744. This ratio is much higher than the ratios generated by the other options and will cause the type II ambiguous data to penalize the N' hypothesis even less. Therefore, to be generous and maximize the model's estimate of θ_N, we choose the third option that makes the set size difference the smallest.

The update equation for type I ambiguous data requires the variable *c*, which corresponds to the number of potential modifiers (or properties) a speaker might use (or refer to). Again, there are a few ways to estimate this value. The learner may use the set of all possible potential modifiers the learner knows (e.g., *red*, *pretty*, *big*, *behind my back*). Alternatively, the learner may use the set of actual potential modifiers in the environment (e.g., if there is a red ball in environment, perhaps *red*, *round*, *bouncy*). The more suspicious a coincidence it is that the object indicated just happens to have the property mentioned in the potential antecedent (e.g., RED if the utterance was ". . . red ball . . . *one* . . ."), the more type I ambiguous data rewards the N' hypothesis (see Appendix C for details why). So, to maximize θ_N, we choose the first option mentioned above, which assumes the superset contains all possible properties the speaker might have mentioned, given the learner's vocabulary. As there will always be many more possible modifiers in the learner's vocabulary than the actual environment provides, this allows type I ambiguous data to reward the suspicious coincidence of a mod-property object the most. Given the data from the MacArthur CDI (Dale & Fenson, 1996), we estimate that an 18-month-old learner should be aware of at least 49 properties in the world, and the generous estimate of *c* is set to 49.

3.4.2. Lower or Upper N' (semantic consequences)

The update equations for θ_U sometimes require the variable *c* (type I ambiguous). The considerations remain the same for the choice of *c* as in the previous section. Therefore, a generous estimate of *c* that will maximize θ_U will use the size of the set of possible potential modifiers. As above, we have roughly estimated this to be 49.

[1]Many thanks to an anonymous reviewer for pointing this out.

4. THE EO MODEL'S PERFORMANCE

As we can see in Figure 7, the learning trajectory as a function of the amount of data seen does not match our ideal learning outcome of a steadily increasing trajectory for the probability of the correct interpretation of anaphoric *one*. In fact, as the learner encounters more data, the probability of the adult interpretation raises only slightly to a final value of 0.260 (1,000 runs, $SD = 0.050$). However, as is apparent in Figure 7, this is entirely the fault of θ_N, the learner's belief that the antecedent category is N' (M final value over 1,000 simulations = .261, $SD = 0.050$). When the learner thinks the category is N', the probability of choosing the upper N' (θ_U) is quite high (M final value over 1,000 simulations = 0.997, $SD = 0.000207$). Unfortunately, a learner using these probabilities will be unlikely to converge on the correct interpretation, in which *one*'s antecedent is an N' and, in fact, the upper N' constituent. As that is the gauge of success, the EO learner appears lacking.

This failure is especially striking because of how generous we were regarding the data available to the EO model and how the learner interpreted those data. There were several places in the construction of the model where we biased the learner toward the correct interpretation of anaphoric *one*, concerning the parameters n, m, and c that determine the relative sizes of the sets in both the syntactic and semantic hypothesis space. As we will demonstrate below, choosing alternative ways of estimating these set sizes only decreases the EO model's performance further.

The variables n and m concern the size of the N' set compared to the size of the N^0 set. The greater the difference, the more type II ambiguous data penalizes the N' hypothesis. We previously estimated these from the data a learner actually encounters, rather than from the generative capacity of the linguistic system. However, if we estimate these set sizes using the generative capacity of the system, we get far greater disparities in set size (cf. section 3.4.1). Using string types such as "Noun" and "Adjective Noun" leads to a subset-to-superset ratio of 0.25; using individual vocabulary items leads to a ratio of 0.020.

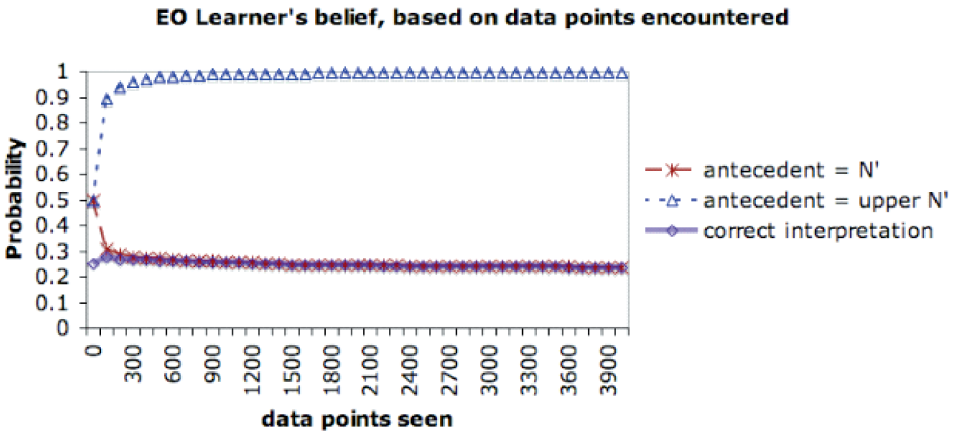

FIGURE 7 The EO learner's beliefs as a function of the amount of data encountered.

We could also alter the value of *c*, allowing it to refer only to the potential modifiers in the learner's environment at the time of the utterance. The larger *c* is, the more type I ambiguous data reward both the N' category hypothesis and the upper N' hypothesis, since it is a suspicious coincidence that the property mentioned in the utterance as a modifier just happens to be the property the referent has, even though the antecedent does not include that modifier. One could imagine that there are a limited number of potential modifiers in the available environment at any given time, perhaps five (e.g. *red*, *round*, *bouncy*, *next to Mommy*, and *rubber* for a red, round, bouncy rubber ball that is next to Mommy.) We might allow *c* to be 5, for instance.

Table 2 shows the results of the EO model's performance with these parameters. As we can see, the probability of converging on the correct interpretation of anaphoric *one* takes a sharp decline. At best, the learner converges on the correct interpretation of *one* with less than a 1 in 20 chance; at the worst, the chance is less than 1 in 250. Note, however, that just as before the real culprit is the category of *one*'s antecedent (as determined by θ_N).

To summarize, even with generous estimates of different variables, the EO model 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 about 1 in 4 (0.260). Altering values for variables in the model only reduces this probability further.

We are of course aware that this is the result of only one unconstrained probabilistic model. However, the important factor in this model's failure has to do with its sensitivity to the type II ambiguous data. These data will be lurking in the input no matter what the probabilistic model used, so any probabilistic model that uses all informative data will be subject to these datas' influence. Thus, we believe our result to be generalizable to all probabilistic models that learn from all the available data: unambiguous, type 1 ambiguous, and type 2 ambiguous.[2] Given this, we conclude that unconstrained (and specifically, unfiltered) Bayesian learning by itself is not sufficient to model human learning or behavior in this domain.

TABLE 2
The EO Model's Performance With Less Generous Estimates of *n, m,* and *c.*

N	*M*	*Ratio: n/(n+m)*	*c*	θ_N	θ_U	*Correct Interpretation*
1	3	0.250	49	0.0489	0.997	0.0487
1	3	0.250	5	0.00917	0.991	0.00909
49	49*247	0.020	49	0.0307	0.997	0.0306
49	49*247	0.020	5	0.00391	0.991	0.00387

[2]It is important to note that this claim is based on realistic estimates of the variables involved. If we alter *n* and *m* such that the ratio is unrealistically high (e.g., above 0.85), then this model can in fact converge on the correct interpretation (e.g., for a ratio of 0.85, just over 50% of the time). However, we feel that it is reasonable to focus on realistic variables estimates as we are attempting to explain realistic child learning behavior.

5. ADDING BACK THE CONSTRAINTS 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 model attempted to solve the problem of anaphoric *one* using a prespecified representational format and domain-general learning procedure but no other constraints on the data intake or learning procedure. In contrast, the model presented by Regier and Gahl, which also used a prespecified representational format and a domain-general learning procedure, implicitly used a filter on data intake. That model succeeded. To make sure that the generative model proposed here is a fair extension of Regier and Gahl's model, we will examine the result of imposing filters on data intake. If the generative model is fair, it will succeed when an appropriate data intake filter is applied. All results reported below use the generous estimates of $n = 346$, $m = 465$, and $c = 49$.

The filter implicit in Regier and Gahl's model was to systematically exclude type II ambiguous data, where the potential antecedent had no modifiers (e.g., . . . ball . . . *one* . . .). We dub this model the Ignore–Type-II model. The trajectory of this model is shown in Figure 8, and shows dramatic improvement. The probability of converging on the adult interpretation is the product of the probability of the correct syntactic category hypothesis (0.997, 1,000 simulations with $SD = 0.000225$) and the probability of the correct N' constituent when there are multiple options (0.997, 1,000 simulations with $SD = 0.000210$), which is 0.994. The Ignore–Type-II learner now has a nearly perfect chance of converging on the preferred adult interpretation of anaphoric *one*. This change can be attributed to θ_N, as θ_U was very near 1 in the EO model as well.

Another filter we might try is to systematically exclude *all* ambiguous data, learning from only unambiguous data (e.g., Fodor, 1998b; Pearl & Weinberg, 2007). Unambiguous data are maximally informative data and will never lead the learner astray. However, an

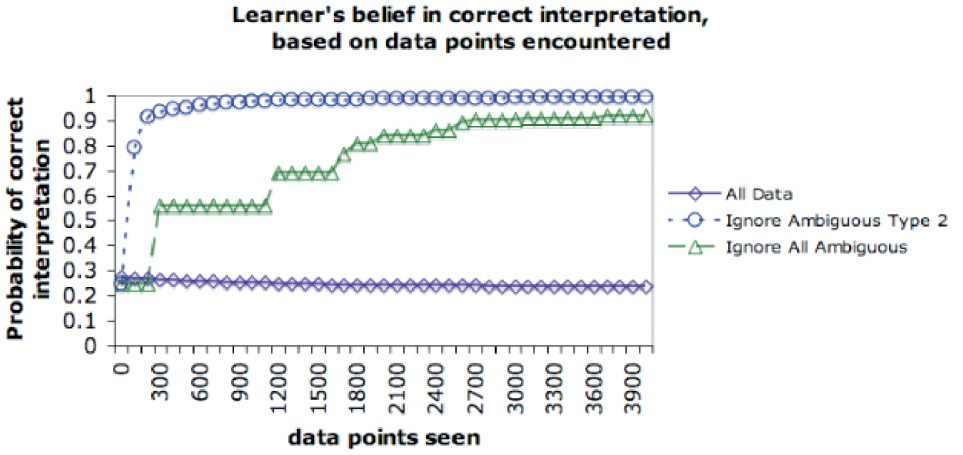

FIGURE 8 The trajectories of the filtered learners (Ignore–Type-II and Ignore–Ambiguous) compared against the EO learner (all data).

unambiguous data learner will also not benefit from any helpful information in the ambiguous data (particularly the type I ambiguous data), and may face data sparseness problems. We dub this model the Ignore–Ambiguous model. The trajectory of this model is also shown in Figure 8 and is far superior to the EO model.[3] Notably, however, it is not quite as good as the Ignore–Type-II learner precisely because it cannot use the helpful information in the ambiguous data. Still, the Ignore–Ambiguous learner also has a very high chance of converging on the preferred adult interpretation of anaphoric *one*. The probability of converging on the adult interpretation is the product of the probability of the correct syntactic category hypothesis (0.950, 1,000 simulations with $SD = 0.0184$) and the probability of the correct N' constituent when there are multiple options (0.950, 1,000 simulations with $SD = 0.0184$), which is 0.903.

To summarize, the EO model shows us that a learner not equipped with filters on data intake cannot converge on the correct interpretation for anaphoric *one*. However, if that same generative model uses a data intake filter that ignores type II ambiguous data, then the model succeeds, just as Regier and Gahl's did. This data intake filter is in fact more beneficial than a constraint to learn only from unambiguous data, as it lessens the data sparseness problem and using type I ambiguous data is more helpful than ignoring them.

6. ON THE NATURE OF THE NECESSARY CONSTRAINT

We have seen that some kind of data intake filter seems to be required in order for a domain-general learning algorithm to learn the correct interpretation of anaphoric *one*. Moreover, a filter to ignore only type II ambiguous data seems more beneficial than a filter that ignores all ambiguous data. Given that this filter requires the learner to single out a specific type of potentially informative data to ignore, and the property of this ignored data involves whether the potential linguistic antecedent has a modifier, we consider this filter to be specific to language learning. As such, it seems reasonable to consider it a domain-specific filter. However, while the content of the filter may be linguistic in nature, it is possible that such a filter can be derived from more general learning–theoretic concerns.

We now consider the question of where this filter comes from, even given that its content is domain specific. It seems fairly obvious that the learner cannot (and probably should not) come equipped with a filter that says "ignore type II ambiguous data" without some procedure for identifying these data. What we really want to know is whether there is a principled way to derive 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 domain-specific filter would become feasible for a learner to implement.

One possibility is that there is a domain-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). So, then, we can ask what counts as uncertainty in this domain and, more specifically, by what measures of uncertainty do type II ambiguous data not introduce uncertainty.

[3]The sudden jumps in the trajectory result from the learner actually encountering an unambiguous data point in the input. The rest of the time, the learner does not learn anything from the input.

Consider again the syntactic hypotheses: *one* = N^0 and *one* = N'. In the case of a type II ambiguous data point such as "Jack wants a ball. Lily has one for him", the choice between N^0 and N' does not yield an observable difference in the linguistic antecedent string: it is always *ball* ([N^0 ball] or [N'[N^0 ball]]). It also does not yield an observable difference in the semantic interpretation: the referent is always a BALL. Hence, there is no uncertainty with respect to either the antecedent for that data point or the interpretation of the sentence relative to the choice of syntactic antecedent. Consequently, one could treat this kind of example as locally uninformative. That is, for these examples, there is no information conveyed, and so no reduction in uncertainty associated with the choice of antecedent. Hence, such sentences should not contribute to learning.[4]

It should be noted, however, that at a more global level the type II ambiguous data do have the capacity to convey information. Because these data involve a choice between N^0 and N', the effect of that choice can have consequences for the appropriate interpretation of subsequent examples. The EO model demonstrates that these consequences are pernicious. Thus, if the learner is driven by a general concern for reducing uncertainty and so learns only in cases of uncertainty, then the calculation of uncertainty must be done at a local level. More specifically, for a given data point, if the choice of value for the syntactic category has no consequences for the item currently being analyzed, then the learner does not use that data point to update the probabilities of the various hypotheses.

To review, we have posited that the origin of the necessary domain-specific filter for learning anaphoric *one* may originate from a learner who uses a domain-general learning strategy, such as learning only in cases of local uncertainty. The intake will consist only of the data points for which there is an observable difference in the local interpretation that results from choosing the syntactic category of *one*. These data points include the unambiguous and type I ambiguous data points only. The syntactocentric learner will not be led astray by the type II ambiguous data points because they will be locally uninformative and hence will not function as input to the learning mechanism. Crucially, the linguistic content of the filter does not need to be built explicitly into the learner's knowledge because it can be derived from the principle of learning only from cases in which the choice has a local consequence. 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 local uncertainty).

7. CONCLUSION

The case of anaphoric *one* demonstrates the interplay between domain specificity and domain generality in language learning. What we have seen is that a domain-general learning procedure can be successful in this case 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-general learning strategy.

[4]Thanks to an anonymous reviewer for highlighting the relevance of local versus global uncertainty.

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, in the data that they deem relevant, or in the procedures used for updating their knowledge. Any one of these by itself may represent 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 is not. This is a false dichotomy. What we have shown here is that the solution to particular problems in language acquisition can be more nuanced, drawing on the strengths of both domain-specific and domain-general learning components.

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 is far easier if the data considered relevant are restricted to just the right subset.

ACKNOWLEDGMENTS

This paper has benefited from discussion with Amy Weinberg, Norbert Hornstein, Colin Phillips, Sandy Waxman, Bill Idsardi, Terry Regier, Charles Yang, Gerry Altmann, Susan Goldin-Meadow, 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 six anonymous reviewers. Special thanks to LouAnn Gerken whose comments improved the explication of these ideas immeasurably. Needless to say, any remaining errors are our own. This work was supported by an NSF graduate fellowship to LP, NSF grant BCS-0843896 to LP, NSF grant BCS-0418309 to JL, and NIH grant R03-DC006829 to JL.

REFERENCES

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

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

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*, 533–581.

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.

Chew, V. (1971). Point estimation of the parameter of the binomial distribution. *American Statistician*, *25*(5), 47–50.

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

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

Fodor, J. D. (1998a). Parsing to learn. *Journal of Psycholinguistic Research*, *27*(3), 339–374.

Fodor, J. D. (1998b). Unambiguous triggers. *Linguistic Inquiry*, *29*, 1–36.

Foraker, S., Regier, T., Khetarpal, A., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Cognitive Science*, *33*, 287–300.
Gallistel, C. R. (2001). Mental representations, psychology of. In *Encyclopedia of the social and behavioral sciences* (pp. 9691–9695). New York: Elsevier.
Gerken, L. (2006). Decision, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67–B74.
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.
Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research, *IRE Transactions on Information Theory*, *IT-8*, 155–159.
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., & 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.
Pearl, L. (in press). Learning English metrical phonology: When probability distributions are not enough. *Proceedings of GALANA 3*. Cascadilla Press.
Pearl, L., and Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development*, *3*(1), 43–72.
Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 417–422). Vancouver, British Columbia.
Pierce, A. (1992). *Language acquisition and syntactic theory: A comparative analysis of French and English child grammars*. Boston: Kluwer Academic.
Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, *93*, 147–155.
Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.
Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626–636.
Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.
Townsend, D., & Bever, T. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: The MIT Press.
Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition*, *40*, 21–82.
Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.
Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Science*, *8*(10), 451–456.

A.1. APPENDIX A

The following argument, due to Baker (1979), explains how we know the English element *one* is anaphoric to strings that are classified as N', and not to a simple noun of category N^0. Consider the following examples in which *one* cannot be anaphoric to a noun:

(A1) a. I met the member of Congress . . .
b. * . . . and you met the one of the Ballroom Dance Team.
c. $[_{NP}$ the $[_{N'}$ $[_{N0}$ member] $[_{PP}$ of Congress]]]

(A2) 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 the following cases in which *one* can substitute for what appears to be only the head noun.

(A3) 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]]]

(A4) 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]]]

These cases differ with respect to the status of what follows the noun. In (A1) and (A2) what follows the noun is a complement, but in (A3) and (A4) 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 such as "Jack has a ball, and Lily has *one*, too", it must be the case that *ball* = N', as in the structure in Figure 1. If it were not, then we would have no way to distinguish this case from one in which *one* cannot substitute for a single word, as in (A1) and (A2).

A.2. APPENDIX B

The model implemented by Regier and Gahl is quite liberal about shifting probability to the superset hypothesis: a single unambiguous data point for the 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, for example, *red ball*, and thus restrict referents to those that have the mod-property, for example, the set of 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 (e.g., a member of the non-red BALLS set is the referent when the utterance is ". . . red ball . . . *one* . . .").

It is crucial for Regier and Gahl'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 (see section 2.2). If the referent of *one* for the utterance ". . . red ball . . . *one* . . ." was a purple ball (perhaps by accident), the new probability for the subset hypothesis in the semantic domain (the mod-property hypothesis, and so the upper N' hypothesis) would be 0. We detail why this occurs below.

Suppose that we refer to the probability that the mod-property hypothesis (equivalent to the upper N' hypothesis) is correct as θ_U. Suppose the learner initially has no bias for either the mod-property hypothesis or the any-property hypothesis, and so the initial probability of θ_U is 0.5 before any data are 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 (mod-property in the semantic domain) if ambiguous data are 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 mod-property hypothesis is correct, using Bayes's rule. The updated θ_U given the observation of *u* is represented as the conditional probability p(mod-property| *u*). To calculate this probability, we use Bayes's rule.

(A5) Calculating the conditional probability p(mod-property| *u*) using Bayes' rule

$$p(\text{mod-prop}|u) \propto p(u|\text{mod-prop})*p(u)$$

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

We substitute this value into the equation in (A5) to get $p(\text{mod-prop}|u) \propto 0*p(u) = 0$. Therefore, the updated probability for θ_U after seeing a single unambiguous superset data point *u* is 0, no matter what the previous probability of θ_U was.

Because this is not terribly robust behavior for a learner, we have adapted the Bayesian updating approach described by Manning and Schütze (1999) to generate a more conservative Bayesian updating approach in the context of a generative model, detailed in section 3 and Appendix C. Unlike the model implemented in Regier and Gahl, the learner using this more conservative approach shifts probability more slowly between hypotheses. A single data point cannot reduce the probability of a hypothesis to 0.

A.3. APPENDIX C

The two parameters to be updated in the model via the input are θ_N and θ_U. The update equations for each use the following equation (Chew, 1971), which assumes parameter θ comes from a binomial distribution and the beta distribution is used to estimate the prior:

(A6) General form of the update equation for θ_N and θ_U

$$\theta_x = \frac{\alpha + data_x}{\alpha + \beta + totaldata_x}, \alpha = \beta = 0.5$$

The parameters α and β determine the initial bias in the learner. When $\alpha = \beta$, each hypothesis is initially equiprobable (θ centered at 0.5). When $\alpha, \beta < 1$, the learner is biased to move near one endpoint or the other (0.0 or 1.0) for θ, rather than remain near the center. The variable $data_x$ refers to how many data points the learner has seen that were "successes" for the parameter (N' data points for θ_N, upper N' data points for θ_U). The $totaldata_x$ variable represents the total

number of data points seen that are relevant for this parameter. For θ_N, all data types (except ungrammatical data points) are relevant since all must choose between N^0 and N'. For θ_U, only unambiguous and type I ambiguous data points are relevant, since they give an option for the lower versus the upper N', if the N' category is chosen.

It is reasonable to think of both θ_N and θ_U as parameters in binomial distributions, given the generative trees in Figures 5 and 6. The classic example of a binomial distribution is for a coin, which has two options: heads or tails. Given the generative trees, two options are available for both θ_N and θ_U. For data generation with θ_N, either the data were generated by N^0 or by N'. For data generation with θ_U, either the data were generated by the lower N' or by the upper N'.

We can adapt the general Bayesian updating equation for each observable data type. Unambiguous data have an utterance in which the potential antecedent of *one* contains a modifier, for example, ". . . red ball . . . *one* . . .". Given the utterance and the environment, the learner knows that the antecedent must be *red ball* and that the object referred to is a ball with the property red (RED BALL).

To update θ_N, the learner must determine what the new value of $data_N$ is, which represents the quantity of N'-generated data points seen so far. This will simply be the previous value of $data_N$ plus the likelihood that the current data point was generated by *one* having an N' antecedent. Using the generative tree in Figure 5 and Bayes's equation, we can calculate the probability that the antecedent was N', given that the antecedent *a* is known to be *red ball* and the object *o* is known to be a RED BALL. This is shown in (A7). The probability that the antecedent is *red ball* and the object is a RED BALL, given that the antecedent is N', is simply the probability of choosing the upper N' constituent, which is θ_U. The prior probability of N' is θ_N. So, the numerator of (A7) is $\theta_U * \theta_N$. The denominator is the probability that the antecedent is *red ball* and the object is RED BALL. Only one branch of the generative tree in Figure 5 will yield this combination: the one in which N' is chosen as the antecedent category and then the upper N' is chosen. This probability is $\theta_N * \theta_U$. The likelihood of N' having generated this data point is therefore 1. This is intuitively satisfying, as unambiguous data (by definition) indicate with perfect certainty that N' is the correct category for *one*'s antecedent.

(A7) Probability that data point was N'-generated for unambiguous data

Known: antecedent of *one* = *red ball*, OBJECT = RED BALL

Value to be added to $data_N$ = p(N' | *a* = *red ball*, *o* = RED BALL)

$$= \frac{p(a = red\ ball, o = RED\ BALL \mid N\ ) * p(N\ )}{p(a = red\ ball, o = RED\ BALL)}$$

(using Bayes's equation)

$$= \frac{\theta_U * \theta_N}{\theta_N * \theta_U} = 1$$

The $totaldata_N$ value is incremented by 1, as a single informative data point has been seen for θ_N. The θ_N update equation is thus as follows:

(A8) Update equation for θ_N, for unambiguous data

$$\theta_N = \frac{\alpha + (data_N + 1)}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

To update θ_U, the learner must determine what the new value of $data_U$ is, which represents the quantity of upper N'-generated data points seen so far. This will simply be the previous value of $data_U$ plus the likelihood that the current data point was generated by *one* having an upper N' antecedent. Using the generative tree in Figure 5 and Bayes's equation, we can calculate the probability that the antecedent was the upper N', given that the antecedent *a* is known to be *red ball* and the object *o* is known to be a RED BALL. This is shown in (A9). The probability that the antecedent is *red ball* and the object is a RED BALL, given that the antecedent is the upper N', is simply 1. The prior probability of the upper N' is θ_U. So, the numerator of (A9) is $1*\theta_U$. The denominator is the probability that the antecedent is *red ball* and the object is RED BALL. Only one branch of the generative tree in Figure 5 will yield this combination: the one in which the upper N' is chosen as the antecedent category. This probability is θ_N*1. The likelihood of the upper N' having generated this data point is therefore 1. This is again intuitively satisfying, as unambiguous data (by definition) indicate with perfect certainty that the upper N' is the correct N' for *one*'s antecedent, with the semantic consequence that the object chosen should have the mod-property.

(A9) Probability that data point was upper N'-generated for unambiguous data

Known: antecedent of *one* = *red ball*, OBJECT = RED BALL

Value to be added to $data_U$ = p(Upper N' | *a* = *red ball*, *o* = RED BALL)

$$= \frac{\text{p}(a = red\ ball, o = \text{RED BALL} \mid \text{Upper N}) * \text{p}(\text{Upper N})}{\text{p}(a = red\ ball, o = \text{RED BALL})}$$

(using Bayes's equation)

$$= \frac{1 * \theta_U}{\theta_U * 1} = 1$$

The $totaldata_U$ value is incremented by 1, as a single informative data point has been seen for θ_U. The θ_U update equation is thus as follows:

(A10) Update equation for θ_U, for unambiguous data

$$\theta_U = \frac{\alpha + (data_U + 1)}{\alpha + \beta + (totaldata_U + 1)}, \alpha = \beta = 0.5$$

Type I ambiguous data have an utterance in which the potential antecedent of *one* contains a modifier, for example, ". . . red ball . . . *one* . . .". Given the utterance and the environment, the

learner knows that the antecedent could be either *ball* or *red ball* and that (in the cases we consider) the object referred to is a ball with the property red (RED BALL).

To update θ_N, the learner must determine what the new value of $data_N$ is, which represents the quantity of N'-generated data points seen so far. This will again be the previous value of $data_N$ plus the likelihood that the current data point was generated by *one* having an N' antecedent. Using the generative tree in Figure 5 and Bayes's equation, we can calculate the probability that the antecedent was N', given that the object *o* is known to be a RED BALL. This is shown in (A11). The probability that the object is a RED BALL given that the antecedent is N' is the probability of choosing the upper N' (θ_U) plus the probability of choosing the lower N' and a RED BALL by chance ($(1-\theta_U)*1/c$). The prior probability of N' is θ_N. So, the numerator of (A11) is $(\theta_U + (1-\theta_U)*(1/c)) *\theta_N$. The denominator is the probability that the object is RED BALL. Three branches yield this result in the generative tree in Figure 5. The first is where N' is chosen and the upper N' is chosen ($\theta_N * \theta_U$). The second is where N' is chosen, the lower N' is chosen, and a RED BALL is chosen by chance ($\theta_N*(1-\theta_U)*1/c$). The third is where N^0 is chosen and a RED BALL is chosen by chance $(1-\theta_N)*1/c$. All three of these possibilities are summed up into the denominator of the equation in (A11). The value added to $data_N$ will be termed $l_{N\text{-}Type1}$ (intuitively, the likelihood of the type 1 ambiguous data point being generated by N'). This value will be less than 1, which is intuitively satisfying, as ambiguous data (by definition) indicate some uncertainty that N' is the correct category for *one*'s antecedent.

(A11) Probability that data point was N' generated for type I ambiguous data

Known: OBJECT = RED BALL

Value to be added to $data_N$ = p(N' | *o* = RED BALL)

$$= \frac{\text{p}(o = \text{RED BALL} \mid \text{N}') * \text{p}(\text{N}')}{\text{p}(o = \text{RED BALL})}$$

(using Bayes's equation)

$$= \frac{(\theta_U + (1-\theta_U) * \frac{1}{c}) * \theta_N}{(\theta_U + (1-\theta_U) * \frac{1}{c}) * \theta_N + \frac{1}{c} * (1-\theta_N)} = l_{N-Type1}$$

The $totaldata_N$ value is incremented by 1, as a single informative data point has been seen for θ_N. The θ_N update equation is thus as follows:

(A12) Update equation for θ_N, for type I ambiguous data

$$\theta_N = \frac{\alpha + (data_N + l_{N-Type1})}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

To update θ_U, the learner must determine what the new value of $data_U$ is, which represents the quantity of upper N'-generated data points seen so far. This will again be the previous value

of $data_U$ plus the likelihood that the current data point was generated by *one* having an upper N' antecedent. Using the generative tree in Figure 5 and Bayes's equation, we can calculate the probability that the antecedent was the upper N', given that the object *o* is known to be a RED BALL and the antecedent's category was N'. This is shown in (A13). The probability that the object is a RED BALL and the antecedent category is N' given that the antecedent is the upper N' (*red ball*) is 1 *θ_N. The prior probability of the upper N' is θ_U. The denominator is the probability that the object is RED BALL and the antecedent's category is N'. Two branches yield this result in the generative tree in Figure 5. The first is where N' is chosen and the upper N' is chosen (θ_N* θ_U). The second is where N' is chosen, the lower N' is chosen, and a RED BALL is chosen by chance (θ_N*(1-θ_U)*1/c). Both of these possibilities are summed up into the denominator of the equation in (A13). The value added to $data_U$ will be termed $l_{U\text{-}Type1}$ (intuitively, the likelihood of the type 1 ambiguous data point being generated by the upper N'). This value will be less than 1, which is intuitively satisfying as ambiguous data (by definition) indicate some uncertainty that the upper N' generated this data point. However, this does reward the suspicious coincidence of choosing an object that happens to have the property mentioned in the modifier (*red*) when the antecedent was really the lower N' ("ball").

(A13) Probability that data point was upper N'–generated for type I ambiguous data

Known: object = RED BALL, antecedent category = N'

Value to be added to $data_U$ = p(Upper N' | *o* = RED BALL, N')

$$= \frac{p(o = \text{RED BALL}, N' \mid \text{Upper N}') * p(\text{Upper N}')}{p(o = \text{RED BALL}, N')}$$

(using Bayes's equation)

$$\frac{1 * \theta_N * \theta_U}{\theta_N * \theta_U * 1 + \theta_N * (1 - \theta_U) * \frac{1}{c}} = l_{U-Type1}$$

The $totaldata_U$ value is incremented by 1, as a single informative data point has been seen for θ_U. The θ_U update equation is thus as follows:

(A14) Update equation for θ_U, for type I ambiguous data

$$\theta_U = \frac{\alpha + (data_U + l_{U-Type1})}{\alpha + \beta + (totaldata_U + 1)}, \alpha = \beta = 0.5$$

Type II ambiguous data have an utterance in which the potential antecedent of *one* contains only a noun, for example, ". . . ball . . . *one* . . . ". Given the utterance and the environment, the learner knows that the antecedent must be *ball* and that the object referred to is a ball (BALL).

To update θ_N, the learner must determine what the new value of $data_N$ is, which represents the quantity of N'-generated data points seen so far. This will again be the previous

value of $data_N$ plus the likelihood that the current data point was generated by *one* having an N' antecedent. Using the generative tree in Figure 6 and Bayes's equation, we can calculate the probability that the antecedent was N', given that the antecedent *a* is *ball* and the object *o* is known to be a BALL. This is shown in (A15). The probability that the object is a BALL and the antecedent is *ball* given that the antecedent category is N' is the probability of choosing a noun-only element from the N' category, which contains a number of noun-only elements (*n*) and a number of noun-and-modifier elements (*m*). This probability is *n*/(*n*+*m*). The prior probability of N' is θ_N. The denominator is the probability that the object is a BALL and the antecedent is *ball*. Both branches in Figure 6 yield this result. The first is when N' is chosen and a noun-only item is generated (θ_N* (*n*/(*n*+*m*))). The second is when N^0 is chosen and a noun-only item is generated (1-θ_N)*(*n*/*n*).[5] Both of these possibilities are summed up into the denominator of the equation in (A15). The value added to $data_N$ will be termed $l_{N\text{-}Type2}$ (intuitively, the likelihood of the type II ambiguous data point being generated by N'). This value will be less than 1, which is once more intuitively satisfying, as ambiguous data (by definition) indicate some uncertainty that N' is the correct category for *one*'s antecedent.

(A15) Probability that data point was N' generated for type II ambiguous data

Known: antecedent = *ball*, object = BALL

Value to be added to $data_N$ = p(N' | a = *ball*, o = BALL)

$$= \frac{\mathrm{p}(a = ball, o = \mathrm{BALL} \mid \mathrm{N}') * \mathrm{p}(\mathrm{N}')}{\mathrm{p}(a = ball, o = \mathrm{BALL})}$$

(using Bayes's equation)

$$= \frac{\frac{n}{n+m} * \theta_N}{\frac{n}{n+m} * \theta_N + \frac{n}{n} * (1-\theta_N)} = l_{N-Type2}$$

The $totaldata_N$ value is incremented by 1, as a single informative data point has been seen for θ_N. The θ_N update equation is thus as follows:

(A16) Update equation for θ_N, for type II ambiguous data

$$\theta_N = \frac{\alpha + (data_N + l_{N-Type2})}{\alpha + \beta + (totaldata_N + 1)}, \alpha = \beta = 0.5$$

Because type II ambiguous data do not have the choice between multiple N' antecedents, these data are not informative for updating θ_U.

[5]This is where the Size Principle comes into play. The subset is the N^0 category, and so it has a higher likelihood of having generated the observed data. The ratio of subset to superset determines how much the subset is favored.