# Chapter 3: The Case of Anaphoric *One*

## *3.1 Anaphoric One: The Necessity of Domain-Specific Constraints*

The phenomenon under investigation is the interpretation of the anaphoric element *one* in English. Previous work has argued that infants' knowledge of anaphoric *one* could not be derived from their experience with this form (Lidz & Waxman, 2004; Lidz, Waxman, & Freedman, 2003). Instead, it was argued that the learner must be equipped with prior constraints on the hypothesis space. Because of these constraints, certain interpretations are simply never considered as potential hypotheses – specifically for anaphoric *one*, the learner would not consider the hypothesis that *one* is anaphoric to $N^0$. These constraints were described as being part of the domain-specific representational format for language learning. However, subsequent work (Regier & Gahl, 2004) replied that a probabilistic learner could acquire this knowledge using the domain-general learning procedure of Bayesian updating. No constraints on the hypothesis space (or domain-specific constraints of any other kind) would then be required. Regier and Gahl (henceforth R&G) provided their learning model with a small set of hypotheses to choose from that were derived from domain-specific representational content. Because there are no constraints on the hypothesis space, R&G's model considers more hypotheses than the learner of Lidz, Waxman, and Freedman (henceforth, LWF).

The two sides are then set up. The LWF learner requires a hypothesis space defined over domain-specific representations, as well as domain-specific constraints that preclude certain hypotheses from being considered. No filters on data intake are posited, and the learner is compatible with a Bayesian updating procedure. The R&G learner also requires a hypothesis space defined over domain-specific representations, but does not require additional constraints on the hypothesis space. Instead, the R&G learner rules out the incorrect hypotheses using a particular implementation of Bayesian updating that exploits the layout of the hypothesis space. R&G also do not *explicitly* posit filters on data intake, and thus claim that no additional information beyond probabilistic updating is required to converge on the correct interpretation of anaphoric *one*.

However, I will argue that R&G's conclusion was too quick. In particular, the R&G learner considers only a restricted source of evidence, which inflates the estimate of the learner's success. By restricting the data intake this way, this model in fact *implicitly* implements two domain-specific filters on the learner's data intake, which will be discussed in detail later in the chapter. However, when a model of a learner that is in the true spirit of the R&G proposal is set up, i.e. one that has no filters on data intake, we will find that this unconstrained Bayesian learning model does not display the correct behavior. If the learner considers the full array of evidence in the input, the learner will fail to learn the correct interpretation of anaphoric *one*.

A Bayesian model without domain-specific constraints is plagued by a particularly pernicious problem in language learning. Specifically, representations

across domains are aligned (e.g. strings of words project to interpretations about referents in the world). In the case studied here, when we allow the learner to consider the correspondences across levels of representation (syntax and semantic reference), we find that an unrestricted Bayesian model fares very poorly. This conclusion casts doubt on Bayesian learning as the sole source of constraints on learners. In short, this case suggests that the overly general nature of domain-general learning must be reigned in by domain-specific representations and domain-specific filters on data intake.

### *3.2 Why Learning Anaphoric One Is Interesting*

To learn the correct interpretation of anaphoric *one*, it is believed that the learner must consider both the syntactic level of representation and the semantic level of representation. At the syntactic level, the learner must learn what the linguistic antecedent of *one* is; at the semantic level, the learner must determine what object in the world the noun phrase containing *one* refers to. Both of these levels contribute to the information a Bayesian learner would use when converging on the correct representation of *one.* A linguistic antecedent (syntax) can be translated into a reference to an object in the world (semantics) and so both syntactic and semantic representations are implicated in knowledge of *one*. As we will see below, the correct syntactic representation for English adults is that the linguistic antecedent of *one* is a string classified as the category N'. This syntactic knowledge has semantic consequences, which are what LWF used to determine if 18-month olds preferred that specific syntactic representation. In this way, we can see that the knowledge that *one* refers to N' strings traverses both the syntactic domain and the semantic domain.

Acquisition of anaphoric *one* is an interesting learning problem because the data that would lead a learner to the correct representation are quite sparse. In particular, LWF estimated that less than 0.3% of the child's input containing anaphoric *one* provided unambiguous evidence for the correct representation. Moreover, the rate of ungrammatical sentences containing anaphoric *one* was twice this amount, making the occurrence of useful (unambiguous) data below noise level. Given this pattern of data, LWF argued (following Baker (1979) and Hornstein & Lightfoot (1981)) that constraints on the representation of anaphoric *one* must be built into the learner's domain-specific representations. The learner should never consider hypotheses where *one* refers to categories smaller than N', such as $N^0$.

R&G countered that a learner using a domain-general Bayesian learning procedure could converge on this knowledge by using ambiguous data with certain properties. This particular class of ambiguous data functions as indirect negative evidence for the correct hypothesis[10]. Using this ambiguous data, they argued, would make the proposed constraint on the linguistic representations unnecessary. The appeal of a domain-general learning procedure without domain-specific filters resides in the lack of biases found inside the learner. However, R&G's model made use of only *some* of the available ambiguous data and of only *semantic* data to converge on the syntactic representation. This decision implements two domain-specific filters on

[10] But see Lasnik (1987) for comments about indirect negative evidence in language learning.

the learner's data intake. I will investigate the results of a probabilistic Bayesian learning procedure that removes these filters.

The procedure I develop uses all the available ambiguous data as well as both syntactic and semantic data to converge on the probabilities of competing representations. I will show that, even under the most generous estimates of the various parameters involved in such a model, a Bayesian learner lacking domain-specific filters on data intake will fail to converge on the syntactic knowledge that *one* is anaphoric to N' strings and fail to have the standard adult interpretation of what set of referents in the world *one* can refer to. In short, the unconstrained Bayesian learner will not learn the preferred adult interpretation of anaphoric *one*, in contrast to what real children do.

The chapter proceeds as follows. First, I will briefly describe the relevant parts of the grammar of anaphoric *one*. I will then review the behavioral evidence indicating 18-month olds have acquired the adult representation of anaphoric *one* and the argument by LWF that the input available to children is too sparse to support acquisition of this knowledge. Then, I address various proposals to circumvent the sparse data problem, and detail how the R&G proposal about a domain-general solution to this problem implicitly implements domain-specific filters on the data intake. Following that, I describe a Bayesian learning model that is truly domain-general, in that it removes all implicit filtering on the data. I show that such a model fails to acquire the adult interpretation of anaphoric *one*. In addition, I describe how under a less charitable assumption of a certain parameter value, the Bayesian learning model would perform even more poorly. Then, I identify the source of the model's failure. One contributing factor to the spectacular failure of the model derives from the link between syntax and semantics. A second contributor to this failure is the abundance of ambiguous data, which given Bayesian learning techniques, causes to the learner to misconverge. I argue that successful acquisition depends on a domain-specific filter on the data. Finally, I speculate on the origin of the necessary domain-specific filter, suggesting that its roots may lie in a syntactocentric approach to learning anaphoric *one*.

## *3.3 Anaphoric One*

### 3.3.1 Adult Knowledge: Syntactic Categories and Semantic Referents

For English adults, the element *one* is anaphoric to strings that are classified as N' (i.e., the antecedent for *one* is an N' string), as in example (1) below. The structures for the N' strings are represented in figure 11.[11]

(1a) *One* is anaphoric to N' (*ball* is antecedent)
"Jack likes this *ball* and Lily likes that *one.*"
(1b) *One* is anaphoric to N' (*red ball* is antecedent)
"Jack likes this *red ball* and Lily likes that *one*."

[11] Note that the precise labels of the constituents here are immaterial. If the structure is [$_{DP}$ this [$_{NP}$ red [$_{NP}$ ball]]], the conclusions reached in this chapter would not be changed.

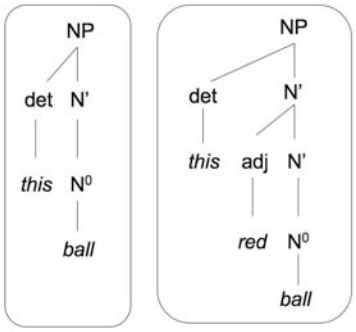


Figure 11. Structures for the N' strings *this ball* and *this red ball*.

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

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

(2i) a. Jack met the member of Congress…
b. * …and Lily met the one of the Society for Creative Anachronism.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ member] [$_{PP}$ of Congress]]]

(2ii) a. Jack reached the conclusion that syntax is innate…
b. * …and Lily reached the one that learning is powerful.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ conclusion] [$_{CP}$ that syntax is innate]]]

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

(2iii) a. Jack met the student from Peoria…
b. … and Lily met the one from Podunk.
c. [$_{NP}$ the [$_{N'}$ [$_{N'}$ [$_{N0}$ student]] [$_{PP}$ from Peoria]]]

(2iv) a. Jack met the student that Lily invited to the party
b. … and Lily 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 (2i) and (2ii) what follows the noun is a complement, but in (2iii) and (2iv) 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. I will represent this by saying that *one* must take N' as its antecedent and that in cases in which there is no complement, the noun by

itself is categorized as both $N^0$ and N'. In other words, in cases like (1a), it must be the case that *ball* = N', as in the structure in Figure 11. If it weren't, we would have no way to distinguish this case from one in which *one* cannot substitute for a single word, as in (2i) and (2ii).

3.3.2 The Pragmatics of Anaphoric *One*

In addition, when there is more than one N' to choose from (as in (1b) above), adults generally 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 et al. (2004)). This semantic consequence is the result of the syntactic preference for the larger N' *red ball*. If the adult preferred the smaller N' *ball*, the semantic consequence would be no preference for the referent of *one* to be red, but rather for it to have any property at all. Importantly, though, this preference is not categorical. It is straightforward to find cases where it is overridden, as in (3):

(3) Jack likes the yellow bean but Lily likes that one.

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

3.3.3 Children's Knowledge of Anaphori*c One*

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

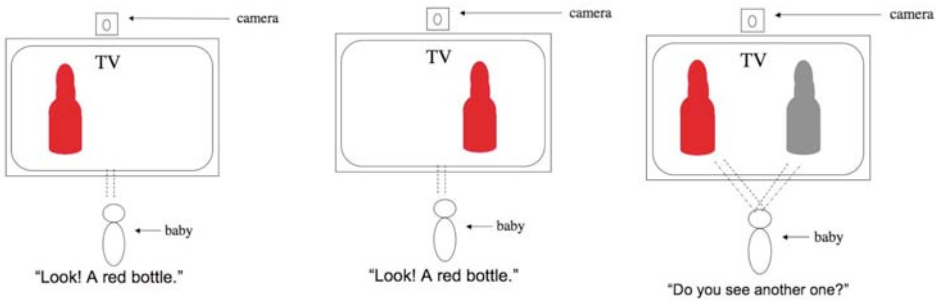



Figure 12. LWF experimental set up.

The infant in the LWF experiment is first shown a bottle of one color while several utterances of the form "Look! An *adjective* bottle." are played

simultaneously. Then, in the test stage, two bottles are shown – one of the *adjective* color and one of another color. The utterance "Do you see another one?" is played simultaneously and the infant's looking preferences are recorded.

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

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

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

### 3.3.4. Sparse Data for Anaphoric *One*

In order to determine whether children's knowledge could have been acquired on the basis of experience with the relevant forms and structures, LWF conducted a corpus analysis on child-directed speech. The important empirical question was how frequently data appeared in child-directed speech that signaled that *one* was anaphoric to N' instead of $N^0$. If the data were not frequent, learning this syntactic knowledge would be difficult. The distribution LWF found is displayed in table 3.1 below.

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

Table 3.1. The distribution of utterances in the corpus examined by LWF.

All data are defined by a pairing of utterance and environment. I will now elaborate on the pairings for each data type. Unambiguous antecedent data have the following form:

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

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

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

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

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

Type I ambiguous antecedent data have the following form:

(6a) Type I ambiguous antecedent example
Utterance: "Jack wants a red ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has the property red.

(6b) Type I ambiguous antecedent example
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him ."
Environment: Lily doesn't have another ball at all.

For data of the form in (6a), Lily has a ball, so the antecedent of *one* could be *ball.* However, Lily also has a ball that is red, so the antecedent of *one* could be *red ball.* Because *ball* could be classified as either N' or $N^0$, these data are ambiguous between *one* anaphoric to N' and *one* anaphoric to $N^0$.

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

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

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

For data of the form in (6b), Lily does not have a ball – but it is unclear whether the ball she does *not* have has the property red. For this reason, the antecedent of *one* is again ambiguous between *red ball* and *ball,* and *one* could be classified as either N' or $N^0$. There were no examples in either Adam or Nina's corpus of this form.

Type II ambiguous antecedent data have the following form:

(8a) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has various properties.

(8b) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, but Lily doesn't have one for him."
Environment: Lily does not have another ball.

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

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

(9) CHI: my pillow my
MOT: That's a good one to jump on.

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

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

Ungrammatical data involve a use of anaphoric *one* that is not in the adult grammar, such as in (10):

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

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

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

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

LWF's experimental results, however, suggested that 18-month olds know that *one* is anaphoric to N'. They therefore claimed that such knowledge does not need to be learned. Instead, the learner would have other innate biases that would allow this knowledge to be derived from the data available. One possibility (cf. Hornstein & Lightfoot (1981), Baker (1979)) would be that the child is constrained only to hypothesize phrasal antecedents for pronouns. Thus, once the child identified *one* as a pronominal form, the possibility that it was anaphoric to Nº would simply never be considered as a potential hypothesis.

## *3.4 Learning Anaphoric One*

### 3.4.1 Suggestions for Learning that *One* is Anaphoric to N'

Two replies to LWF made suggestions for how this syntactic knowledge could be learned from the available data. The first reply by Akhtar et al. (2004) noted that even if the percentage of unambiguous data is quite small, 18-month olds have still been exposed to an estimated 1,000,000 utterances; this should yield a larger quantity of unambiguous data than the LWF corpus analysis obtained. So, a learner using only unambiguous data would encounter more unambiguous examples by 18 months. Still, the overall percentage of unambiguous data remains quite small (0.25%).

However, it is unlikely that this is even a fair estimate of the amount of data that the child has been exposed to. This is because much of the first year of life is spent learning phonological and lexical properties of the language which would be prerequisites to learning syntax. To derive a fairer estimate of the amount of relevant data an 18-month old might have been exposed to, I assume that learning the syntactic and semantic properties of *one* can only commence once the child has some

repertoire of syntactic categories. Thus, I estimated that the learning period begins at 14 months because there is experimental data supporting infant recognition of the category Noun and the ability to distinguish it from other categories such as Adjective at this age (Booth & Waxman, 2003). If learners hear approximately 1,000,000 sentences from birth until 18 months, they should hear approximately 278,000 sentences of input between 14 months and 18 months. The adjusted expected distribution of anaphoric *one* data is displayed in table 3.2.

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

Table 3.2. The expected distribution of utterances in the input to learners between 14 and 18 months.

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

R&G offer a solution: make use of the type I ambiguous data as well, which gives 183 additional data points (on this estimate). Using a Bayesian learning model that implements the size principle of Tenenbaum & Griffiths (2001), R&G demonstrate how a learner could use both unambiguous and type I ambiguous data to converge on the correct representation. I review their learning model in the next section.

### 3.4.2 A Regier & Gahl Bayesian Learner

The power of R&G's model comes from using indirect evidence available in the type I ambiguous data. This is an attractive strategy, since there are nearly 20 times as many type I ambiguous data as there are unambiguous data (183 to 10). This raises the useable data for the learner up to 4.8% (193 of 4017), which seems more in

line with the amount required for acquisition as early as 18 months (Yang (2004)). The indirect evidence itself is derived solely from the environment in which type I ambiguous data are uttered – specifically, by the learner examining the distribution of the referents of *one*. For example, suppose the learner hears type I ambiguous data such as the example in (6a) (repeated below as (11)):

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

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

Now, the connection between the N' antecedent and a referent with the property mentioned in the N' will be true for some portion of the type I ambiguous data.[12] Crucially, for R&G's model, it is *never* true that the referent of *one* definitively lacks the property mentioned in the N' antecedent (i.e. the referent of *one* is definitively not red when the antecedent is *red ball*). A Bayesian learner using the size principle is very sensitive to this fact in the following way:

(12) Bayesian Learner Logic
(a) For type I ambiguous data, suppose that the referent of *one* could have any property, and not necessarily have the property mentioned in the larger N' antecedent. Suppose also that the set of potential referents for an utterance like (11) is represented in figure 13.

[12] This reasoning will not work for type I ambiguous data of the form in (2b): "Jack wants a red ball, but Lily doesn't have another one for him", where Lily does not have a ball. This is because the learner cannot tell whether or not the ball Lily doesn't have has the property red. These data are therefore not useful as indirect evidence. Such data did not occur in the Adam and Nina corpora from which my estimates are derived.

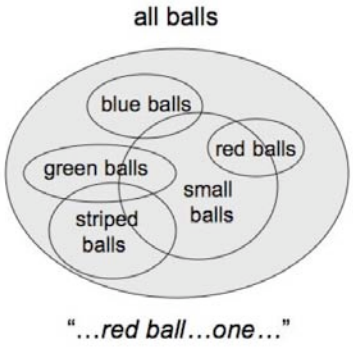



Figure 13. The set of potential referents for *one* in the world when an utterance such as "Jacks wants a red ball, and Lily has another one for him" is heard.

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

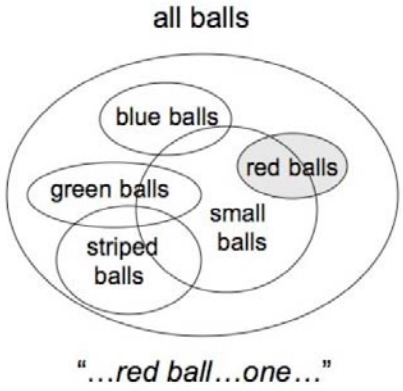



Figure 14. The observed set of referents for *one* when an utterance such as "Jack wants a red ball, and Lily has another one for him" is heard.

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

The specific instantiation of the bias for the subset (red balls) given a single subset data point is based on the likelihood of encountering that subset data point. The likelihood of choosing a specific member of the subset (a red ball) is higher if members can be drawn only from the subset (red balls), as opposed to if members can be drawn from the superset (all balls). This occurs because the superset necessarily has more members to choose from, and

therefore there is a lower probability of choosing a specific subset member.

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

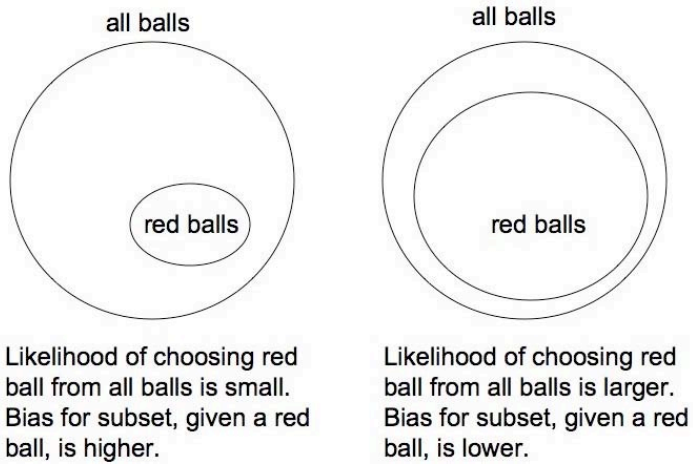



Figure 15. Comparison of different ratios of superset to subset, the likelihood of choosing a member of the subset, and the effect on subset bias.

(d) Once the learner is biased to believe that there is a restriction to the subset of referents described by the property mentioned in the N' (*red* in *red ball*), the learner then assumes that the correct antecedent is, in fact, the larger N'.[13] Since the larger N' cannot be classified as $N^0$, the learner then knows that *one* always has an N' antecedent.

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

A great strength of the R&G model is that the bias to choose the subset, given indirect evidence, does not need to be explicitly assumed. Instead, it falls out neatly from the mathematical implementation of the Bayesian learning procedure itself that uses the size principle of Tenenbaum & Griffiths (2001).

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

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

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

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

## *3.5 An Equal-Opportunity Bayesian Learner*

### 3.5.1 Introducing the Equal-Opportunity Bayesian Learner

I will refer to the unrestricted domain-general learning model as the Equal-Opportunity (EO) Bayesian Learner since it removes the two implicit biases of R&G's Bayesian learner and so gives equal treatment to all data. First, it denies privileged status to a subset of the data and instead uses all the data available: unambiguous, type I ambiguous, and type II ambiguous. Second, it denies privileged status to semantic data – syntactic and semantic data are both used to shift probability between opposing hypotheses. There is an intuitive logic to using both types of data, since one should presumably use syntactic data (among other kinds of data) to converge on syntactic knowledge.[14] This syntactic knowledge has semantic

[14] Note that even if we believed the knowledge about *one* was stated purely in semantic terms, the data that any grammar predicts will include both syntactic data (i.e. what the linguistic antecedent for *one* is) and semantic data (what the referent of *one* is). So, excluding either kind of data is an arbitrary restriction on the learner that would need to be justified. For this reason, the hypothesis to include both

consequences, which are displayed in the LWF experiment. If a Bayesian learning procedure, unconstrained by domain-specific filters, is to be an effective domain-general learning solution, it should correctly acquire knowledge that spans domains such as syntax and semantics as well as knowledge contained completely within these domains.

### 3.5.2 The Hypothesis Space

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

For each domain, the elements analyzable by one hypothesis are a subset of the elements analyzable by the other. For syntax, the hypotheses under consideration are (a) that *one* is anaphoric to strings that are classified as $N^0$ and (b) that *one* is anaphoric to strings that are classified as N'. Every string in $N^0$ can also be classified as N' but there are strings in N' that cannot be classified as $N^0$. Therefore, the strings that comprise the $N^0$ set are a subset of the strings that comprise the N' set.

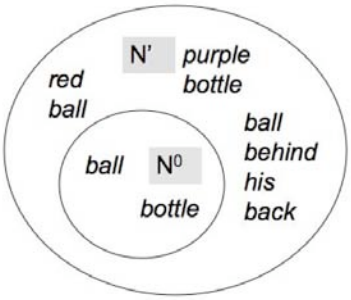



Figure 16. The syntax hypothesis space, $N^0$ vs. N'. All the elements in the sets are strings that are possible antecedents of *one*. Every string classified as $N^0$ can also be classified as N'. In addition, there are strings in N' that are not in $N^0$, and so the $N^0$ set is a subset of the N' set.

For the semantic interpretation, the referents of *one* could have the restriction that they must have the property named by the modifier; alternatively, the referents of *one* could have no restriction on what property they have. Since the modifier is linguistically not part of the $N^0$ (recall figure 11) and instead is part of the N' phrase,

---

syntactic and semantic data does not rely on a particular specification of knowledge about anaphoric *one*.

I will refer to the property named by the modifier as the N'-property. I will refer to referents with no restrictions as being any-property referents, since these referents can have any property (though of course they must still be instances of the noun in the antecedent, e.g. balls, if the antecedent is *red ball*). So, in the semantic domain, the two hypotheses under consideration are (a) that the referent of *one* is restricted to have the N'-property and (b) that the referent of *one* can have any property (is not restricted to have the N'-property).

Just as in the syntactic domain, the elements predicted by one hypothesis are a subset of the elements predicted by the other (see figure 17). Every referent that has the N'-property (say red for *red ball*) is a member of the N'-property set. By definition, every member of the N'-property set is also a member of the any-property set, since the N'-property is one of the properties available for objects to have. However, there are members of the any-property set (say green balls for the linguistic antecedent *red ball*) that do not have the N'-property (red). So, since all the members of the N'-property set are members of the any-property set, the N'-property set is a subset of the any-property set. Moreover, some members of the any-property set are *not* members of the N'-property set. So, the any-property set is a superset of the N'-property set in the semantic domain.

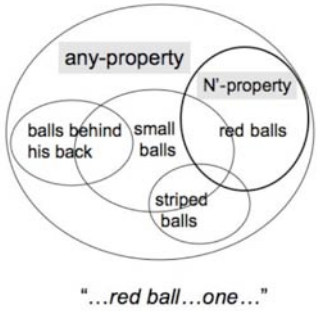



Figure 17. The semantic hypothesis space, N'-property vs. any-property. Any-property is a superset of N'-property. Note that in order to define the sets (N'-property vs. any-property), the utterance must be used to determine the salient property that the referent of the antecedent has. The salient property can be determined from the linguistic antecedent of *one*.

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

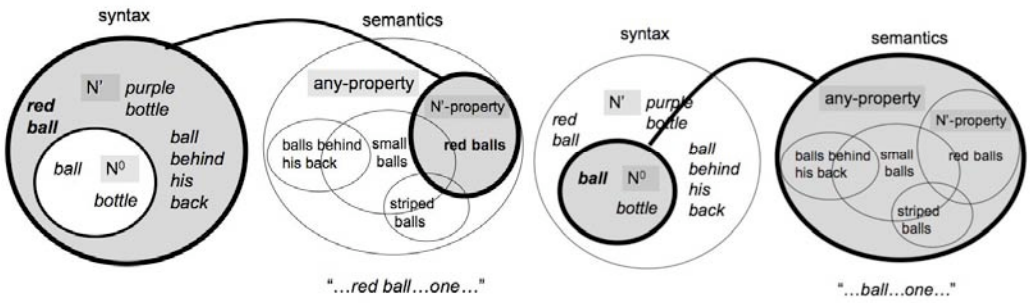



Figure 18. In the syntactic domain, the exclusive superset is linked to the subset in the semantic domain. The subset of the syntactic domain is linked to the superset in the semantic domain.

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

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

It is important to recognize that the problem of linked hypothesis spaces extends far beyond the particular case of anaphoric *one*. Because syntactic structures are semantically compositional, this problem will persist across the acquisition of any aspect of the grammar that depends on the link between syntax and semantic reference.

### 3.5.3 EO Bayesian Learning

The EO Bayesian learning model uses Bayesian reasoning to update the learner's confidence in each of two alternative hypotheses. The implementation I will use differs from the R&G learner by being more conservative about updating the probabilities of the competing hypotheses. I will first describe the R&G Bayesian implementation, and then describe the implementation I will use here. I detail the learning process independently for each of the two domains (syntax and semantics) that are relevant for determining the appropriate structure of anaphoric *one*. I then

describe how to implement the updating algorithm, given that these two domains are linked.

3.5.3.1 The R&G Bayesian Learner Implementation

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

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

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

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

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

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

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

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

linguistic antecedent. So, the probability of observing *u* if the N'-property hypothesis is true (p(*u*|N'-prop)) is 0.

We substitute this value into the equation in (13) to get $p(\text{N'-prop} \mid u) \propto 0 * p(u) = 0$. Therefore, the updated probability for $p_{\text{N'-prop}}$ after seeing a single piece of unambiguous superset data *u* is 0, no matter what the previous probability of $p_{\text{N'-prop}}$ was.

Since this is not terribly robust behavior for a learner, I have adapted the Bayesian updating approach described by Manning & Schütze (1999) to generate a more conservative Bayesian updating approach, detailed in the previous chapter. Unlike the liberal R&G model, the learner using this more conservative approach shifts probability much more slowly between hypotheses. Only after observing a vast majority of evidence for one hypothesis would a conservative Bayesian learner shift the vast majority of the probability into that hypothesis.

3.5.3.2 Updating the Syntax Hypotheses

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

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

The update function requires a single parameter *t*, which represents the total amount of data expected during the learning period, as described in the previous chapter, and can be thought of as the total amount of change the real learner's brain is allowed to undergo before settling into the final state. In the simulated learner here, I quantify that amount of change as the total estimated amount of useable data available during the learning period (4017 data points, if using all available data). Of course, the value of *t* is essentially arbitrary, but in order to model this learning process, it needs to be estimated. The model uses *t* to determine how much probability shifting should be done, given a single piece of data. If *t* is small, only a small number of changes are allowed and each piece of data shifts the probability quite a lot; conversely, if *t* is large, a large number of changes are allowed and each piece of data shifts the probability a smaller amount. The value of *t* I use here will allow the modeled learner to converge as close as possible to an endpoint (e.g. $p_{N'} \approx 1.0$). In this way, I hope to estimate the best-case scenario for this kind of learner. While the *t* estimate presented here seems fair, I present a range of possible *t*-values in the results section. What we will see there is that the size of t does not influence the final probability of the correct interpretation of anaphoric *one*.

The exact update functions for $p_{N'}$ depend on the data type observed – unambiguous, type I ambiguous, or type II ambiguous. Unambiguous and type I ambiguous data cause the learner to use the function in (14a), which is essentially an implementation of the indirect negative evidence update function used by the R&G

model. Type II ambiguous data, which were not considered by the R&G learner, cause the EO Bayesian learner to use the function in (14b).

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

Utterance: "…red ball…one…"

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

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

(14b) Update function for type II ambiguous data

Utterance: "…ball…one…"

World: referent has various properties (type II ambiguous)

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

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

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

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

However, I will allow the Bayesian learner to treat the type I ambiguous data with full confidence in the N' hypothesis. I make this allowance for two reasons. First, I know of no principled way to reasonably estimate how much confidence should be associated with a type I ambiguous data point. Second, this allowance is generous towards the Bayesian learner because it allows the model to overestimate the confidence the learner has in the N' hypothesis. If I was less generous and lessened the confidence in the type I ambiguous data, the probability of N' would

only be lower than what I present here. As we will see below, even with this generous estimate, the learner will fail to assign sufficient probability to the N' hypothesis.

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

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

$$(15) \quad \text{Prob}(N' \mid a) = \frac{\text{Prob}(a \mid N') * \text{Prob}(N')}{\text{Prob}(a)}$$

I now describe the individual pieces of the right hand side of the equation in (15). Prob(*a* | N') is the probability of observing a type II ambiguous data point *a*, given that the N' hypothesis is true. Recall that a type II ambiguous data point has an utterance with a noun-only antecedent, such as "…ball…one…". The N' hypothesis states that the linguistic antecedent of *one* must be an N' constituent.

It is possible for a noun-only string to be an N' constituent: this is the situation where a noun-only string is chosen from the set of N' constituents, which consists of both noun-only strings ("ball", "bottle", etc.) and other strings that include modifiers ("red ball", "bottle in the corner", etc.). The probability we want is the probability of choosing a noun-only linguistic antecedent for *one* (such as in type II ambiguous utterance *a*), given the entire set of N' constituents. Suppose there are *n* noun-only strings and *o* other strings in the N' constituent set. I refer to the probability of choosing a noun-only string (such as "ball") as $p_{N \text{ from } N'}$, and it is calculated below in (16).

$$(16) \text{ Prob}(a \mid N') = \frac{n}{n+o} = p_{N \text{ from } N'}$$

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

Prob(*a*) is the probability of observing a type II ambiguous utterance *a*, no matter which hypothesis is correct. To calculate this value, we can sum the conditional probabilities of observing *a* for each hypothesis (Prob(*a* | N') + Prob(*a* |

$N^0$)) . If N' is the correct hypothesis, the probability of observing *a* is Prob(*a* | N') from above. If $N^0$ is the correct hypothesis, then the linguistic antecedent of *one* is an $N^0$ constituent, which is always a noun. In that case, the probability of observing a noun-only linguistic antecedent (such as in *a*) is 1. We can calculate Prob(*a*) in (17).

(17) $\text{Prob}(a) = \sum_{hypotheses} p_{hypothesis} * p(a \mid p_{hypothesis})$

$= p_{N'} * p(a \mid p_{N'}) + p_{N0} * p(a \mid p_{N0})$

$= p_{N'} * \frac{n}{n+o} + (1-p_{N'}) * 1$

Substituting these pieces back into the right hand side of the equation in (15), we obtain (18).

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

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

The likelihood value $p_{N \text{ from } N'}$ is what allows the learner to retrieve information from the type II ambiguous data. The more unbalanced the ratio of noun-only strings to other strings in the N' set, the stronger the effect of the size principle will be that biases the learner towards the subset $N^0$ hypothesis. Example (19) displays how much biasing occurs after a single piece of type II ambiguous data, assuming a current $p_{N'}$ of 0.5, a ratio of noun-only strings to total N' strings of 0.25, and a *t* of 4017.

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

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

3.5.3.3 Updating the Semantics Hypotheses

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

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

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

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

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

(20a) Example of same-property data (syntax: type I ambiguous)
Utterance: "Jack wants a red ball, and Lily has another one for him."
World: Lily has another red ball for Jack.

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

The data analyzed as unambiguous in the syntactic domain are also same-property data in the semantic domain. There are 10 data points of this form. Because these data necessarily include negation, seeing why they are same-property data is a bit complicated. Consider the example in (20b).

(20b) Example of same-property data (syntax: unambiguous)
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."
World: Lily has a non-red ball for Jack.

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

Because the N'-property hypothesis depends on matching the property overtly mentioned in the modifier (e.g. *red* of *red ball*), type II ambiguous data are not informative for choosing between the two hypotheses. This is simply because there is no overtly mentioned modifier, as shown in (20c). Therefore, the semantic interpretation projection from the syntactic hypothesis space is a single hypothesis (the any-property hypothesis). Since the semantic domain only has one hypothesis for type II ambiguous data, no information can be inferred about the correct hypothesis when there is more than one semantic interpretation to choose. The learner therefore ignores the semantic hypothesis space when encountering type II ambiguous data.

(20c) Example of same-property data (syntax: type II ambiguous)
Utterance: "Jack wants a ball, and Lily has another one for him."
World: Lily has a ball with some property for Jack.

A different-property example is given in (21), when the potential antecedent has a property mentioned in the modifier (e.g. red of *red ball*), but the referent of *one* does not have this property. This situation would occur in rare cases, perhaps as noise or perhaps because of a pragmatic bias.

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

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

Finally, we come to the unknown-property data, as in (22).

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

In the examples in (22), the speaker is asserting the absence of a ball. The referent of *one* is a ball, with some unknown properties, that is not present in the

situation. Thus, the referent of *one* may or may not include the property (red) mentioned in the potential antecedent.

A portion of type I ambiguous data consists of unknown-property data. Such data cannot be used for updating the probabilities of the opposing semantic hypotheses. However, I will be generous and allow R&G's assumption to hold true: none of the type I ambiguous data are of this form. Therefore, I will allow all type I ambiguous data to be of the form in (20a), which is an example of same-property data. This gives an overestimation of $p_{N'\text{-prop}}$, which is the subset in the semantic hypothesis space. Consequently, this will bias the learner towards the superset in the syntactic hypothesis space, N'. Thus, the model here will again overestimate the amount of probability the learner will assign to the correct hypothesis for the structure and interpretation of anaphoric *one*, given an utterance with more than one potential antecedent.

Table 3.3 represents the expected distribution of data for updating the semantic hypotheses in this model.

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

Table 3.3. The expected distribution of utterances in the input to the Bayesian learner for updating the semantics hypotheses. Note that the type II ambiguous data points are uninformative in the semantic interpretation domain, so those 3805 data points are ignored.

The exact update functions for $p_{N'\text{-prop}}$ depend on the data type observed. However, the only update function relevant for this model is the same-property update function (23), which is similar to its syntactic counterpart in (14b). In both cases, the subset hypothesis is favored upon encountering an ambiguous data point.

(23) Update function for same-property data

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

The same-property update function is derived using the same reasoning as the type II ambiguous update function in the syntactic domain. We again have two hypotheses (N'-property and any-property), and so we can use a binomial distribution to approximate the learner's expectation of the distribution of data to be encountered. The binomial distribution is centered at $p_{N'\text{-prop}}$, so the learner's expectation is about how many N'-property data points should be observed. To update $p_{N'}$ after seeing a single same-property data point *s*, we again follow the framework laid out in the previous chapter and calculate the maximum of the a posteriori (MAP) probability.

Like the type II ambiguous data update function in the syntactic domain, however, we will add a value smaller than 1 to the numerator. Intuitively, this smaller value represents the learner's smaller confidence that the same-property data point *s* indicates that the N'-property hypothesis is correct. I call this smaller value the partial confidence value, and represent it as $p_{N'\text{-prop} \mid s}$.

The partial confidence value $p_{N'\text{-prop} \mid s}$ is the probability that the referent of *one* must have the N'-property mentioned in *s*. This is equivalent to the probability that the referent of *one* must have the N'-property in general, given that *s* has been observed. I write it as Prob(N'-prop | *s*) and calculate it by using Bayes' rule.

$$(24) \quad \text{Prob(N'-prop}| s) = \frac{\text{Prob}(s \mid \text{N'-prop}) * \text{Prob(N'-prop)}}{\text{Prob}(s)}$$

I now describe the individual pieces of the right hand side of the equation in (24). Prob(*s* | N'-prop) is the probability of observing a same-property data point *s*, given that the N'-property hypothesis is true. Recall that in a same-property data point, the referent of the antecedent of *one* must have the same mentioned property that the referent of *one* has. The N'-property hypothesis states that the referent of the antecedent of *one* must have the property described by the linguistic antecedent of *one*. Therefore, if the N'-property hypothesis is true, the probability of observing a same-property data point is 1.

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

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

Prob(*s*) is the probability of observing a same-property utterance *s*, no matter which hypothesis is correct. To calculate this value, we sum the conditional probabilities of observing *s* for each hypothesis (Prob(*s* | N'- prop) + Prob(*s* | any-prop)) . If N'-property is the correct hypothesis, the probability of observing *s* is Prob(*s* | N'-prop) from above. If any-property is the correct hypothesis, then there is no restriction on what property the referent of the linguistic antecedent of *one* has. I estimate the probability of that referent having the same property by chance as the referent of *one* as simply 1/*c*, where there are *c* properties in the world. I calculate Prob(*s*) in (26).

(26) $\text{Prob}(s) = \sum_{hypotheses} p_{hypothesis} * p(s \mid p_{hypothesis})$

$= p_{N'\text{-prop}} * p(s \mid p_{N'\text{-prop}}) + p_{any\text{-prop}} * p(s \mid p_{any\text{-prop}})$

$= p_{N'\text{-prop}} * 1 + (1 - p_{N'\text{-prop}}) * \frac{1}{c}$

Substituting these pieces back into the right hand side of the equation in (24), we obtain (27).

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

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

### 3.5.4 The Updating Algorithm for Linked Domains

Recall that there is an inherent connection between the syntax and the semantic interpretation. In particular, the subset hypothesis in the syntax ($N^0$, or the smaller N' constituent) corresponds to the superset hypothesis in the semantics (any-property), and the exclusive subset in the syntax (larger N' constituents) corresponds to the subset (N'-property) in the semantics (figure 18). Given this arrangement of hypothesis spaces, any piece of data impacting a hypothesis in one domain should impact the corresponding hypothesis in the other domain by the same amount. I now provide a description of how I model this process.

First, suppose the learner receives an unambiguous or type I ambiguous data point (which have two strings as potential antecedents, e.g. *red ball* or *ball*). This data point can be analyzed in either domain, syntax or semantics. So, the learner chooses which one to analyze it in first. Then, the update functions described above are employed to determine the amount the probability that should be shifted within that domain. Next, the probability is shifted in the other domain by the same amount. See figure 19, which shows the learner analyzing the data in syntax and updating both syntax and semantics. Now, the learner analyzes the data point in the other domain, applies the update functions described previously to determine the amount the probability that should be shifted within this domain. Next, the probability is shifted

in the other domain by the same amount. See figure 20, which shows the learner analyzing the data in the semantics and updating both semantics and syntax.

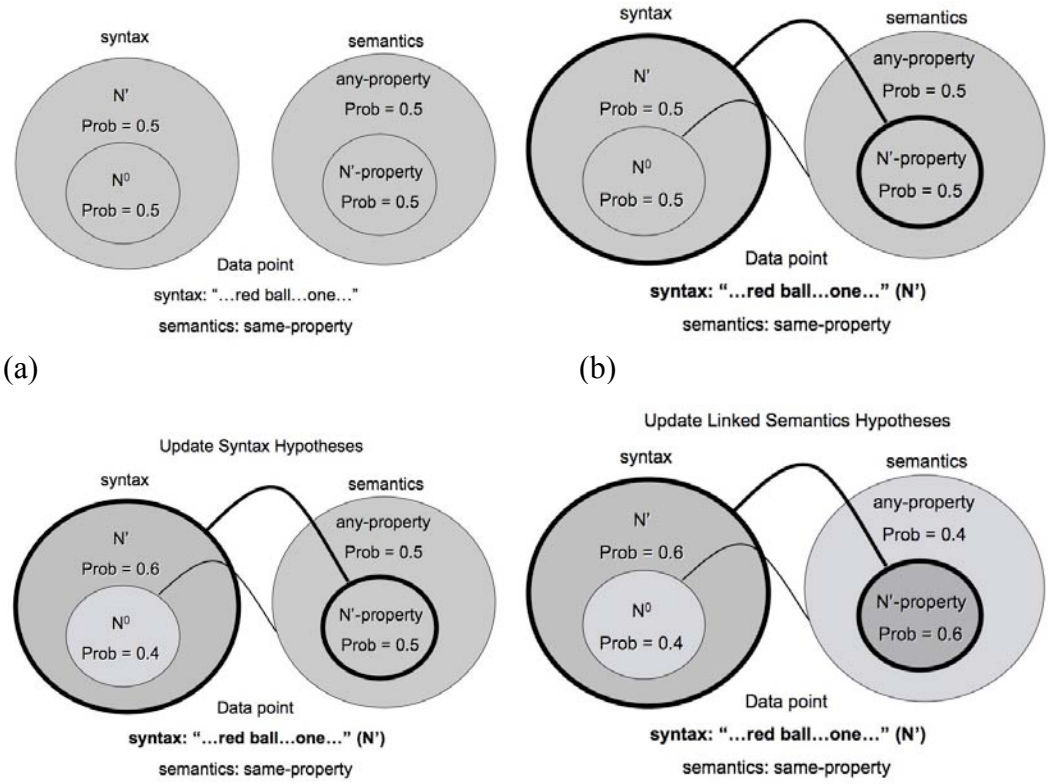
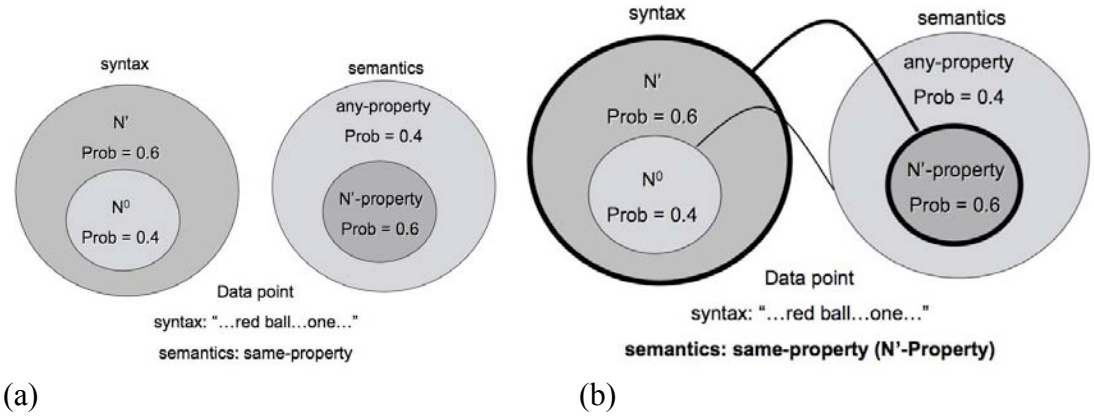
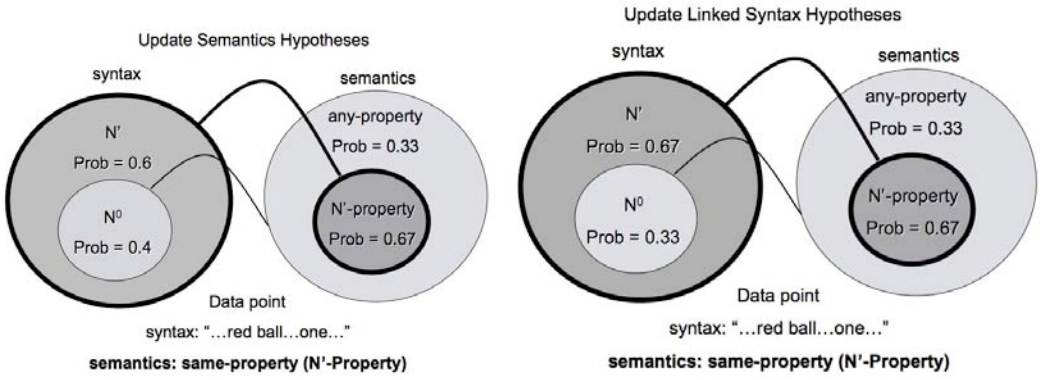



Figure 19. The learner encounters an unambiguous data point (a) and analyzes it first in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c) and the probability of the linked semantics hypotheses (d).

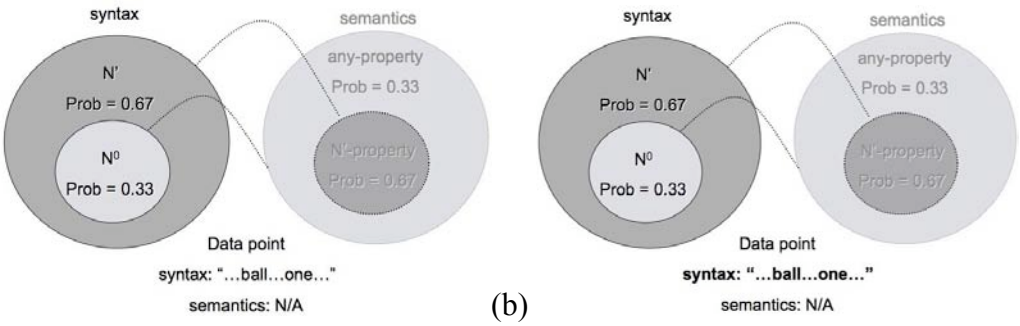




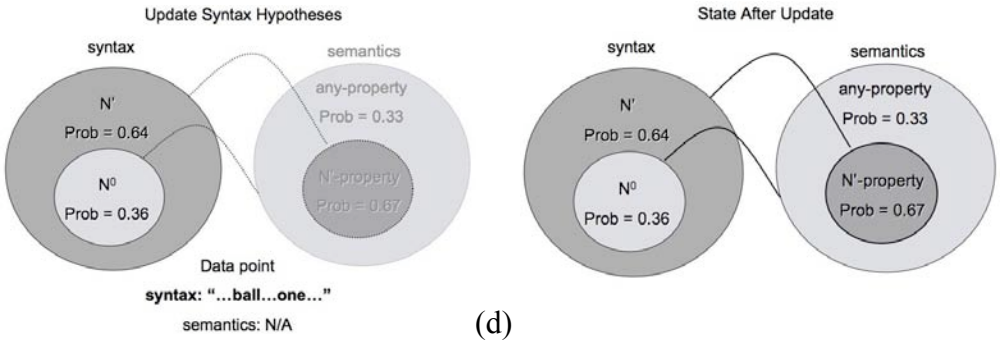

(c) (d)

Figure 20. After analyzing the data point in the syntax domain and updating the probabilities across the domains, the learner then starts at the state in (a) and analyzes the data point in the semantics domain (b). Then, the learner updates the probability of the semantics hypotheses (c) and the probability of the linked syntax hypotheses (d).

The update process differs for a type II ambiguous data point, however. This is because there is only one string that is the potential antecedent (e.g. *ball*), and the projection from the syntax to the semantics leaves only one interpretation (any-property). Type II ambiguous data points are thus uninformative for the semantic interpretation domain. So, the learner simply updates in the syntax domain alone, as shown in figure 21. The semantic interpretation domain is ignored for this type of data.




(a) (b)



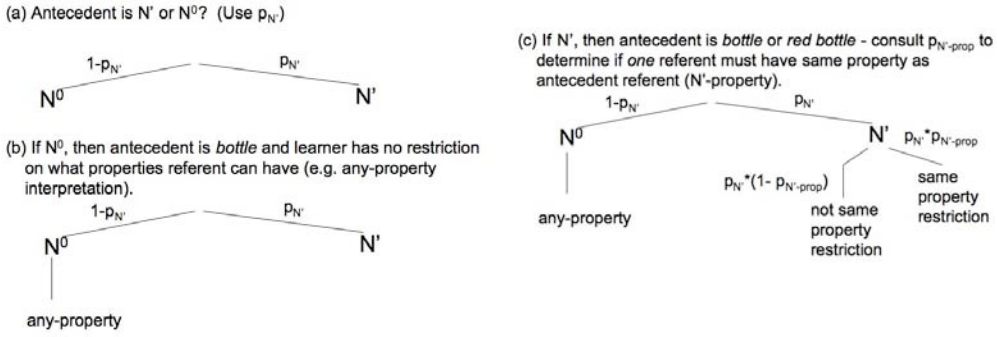

(c) (d)

Figure 21. The learner encounters a type II ambiguous data point (a) and analyzes it in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c). The final state after update is show in (d). Importantly, the semantic domain is not influenced by the type II ambiguous data point because there is only one semantic interpretation available for an antecedent with no modifiers (e.g. *ball*), the any-property hypothesis. The semantic domain is only influenced when there is more than one potential antecedent, leading to more than one semantic interpretation.

3.5.5 What Good Learning Would Look Like

In the model, the learner initially assigns equal probability to the two hypotheses in each of the two domains: in the syntax, $N^0$ and N', and in the semantics, N'-property (corresponding to the larger N' constituent interpretation, e.g. *red ball*) and any-property (corresponding to the smaller N' constituent interpretation, e.g. *ball*). The probability of choosing the preferred adult interpretation, given an utterance with two potential antecedents, depends on choosing the correct hypothesis in each domain. So, if the learner hears, "Look! A red bottle! Do you see another one?" (as in the LWF experiment), the interpretation of *one* is calculated as in (28), which is schematized in the decision tree in figure 22.

(28) Interpreting *one* in "Look! A red bottle! Do you see another one?"
(a) Determine if the antecedent of *one* should be $N^0$ or N', using $p_{N'}$.
(b) If the antecedent is $N^0$, then the referent can have any-property.
(c) If the antecedent is N', use $p_{N'\text{-prop}}$ to determine if the smaller N' constituent interpretation (any-property) or the larger N' constituent interpretation (N'-property) should be used.



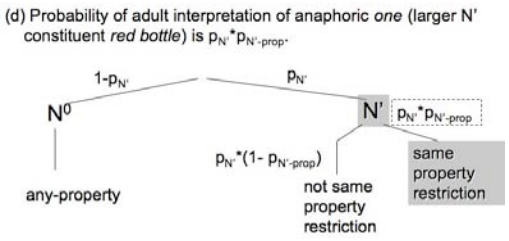



Figure 22. Decision tree to interpret anaphoric *one* in utterances with more than one potential antecedent, such as "Look! A red bottle! Do you see another one?" The probability of having the adult interpretation (*one* = *red bottle*) is $p_{N'}*p_{N'\text{-prop}}$.

The probability of choosing the preferred adult interpretation (the larger N' constituent is the antecedent of *one*) is the product of the probability of choosing the correct hypothesis in the syntax (N') and that of choosing the correct hypothesis in the semantic interpretation (N'-property = larger N' constituent): 0.500 * 0.500 = 0.250. Given that the end state should be a probability near 1, a good learning algorithm should have a trajectory like that illustrated in figure 23. In short, the learner should steadily increase the probability of choosing the preferred adult interpretation.

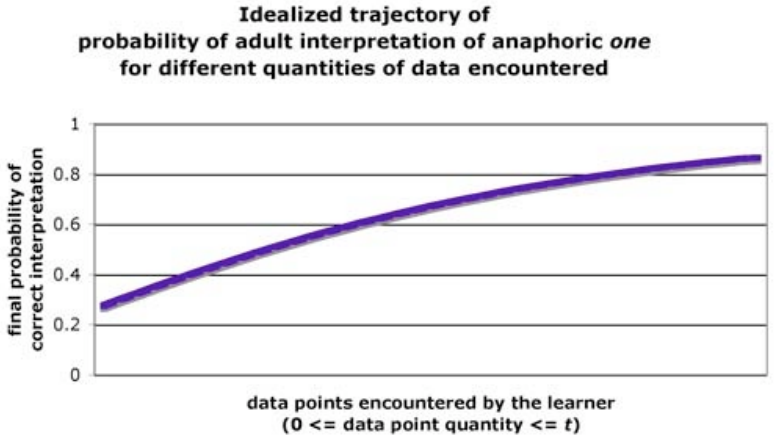



Figure 23. The idealized trajectory of the probability of the correct interpretation for anaphoric *one* as a function of the data points encountered by the learner.

3.5.6 Simulating an EO Bayesian Learner

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

#### 3.5.6.1 Syntax

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

Let the set of strings in N' = {Noun, Adjective Noun, Noun PrepositionalPhrase, Adjective Noun PrepositionalPhrase}.[15] The probability of producing a Noun string from this N' string set is 1/4 or 0.25. We can now look at the semantic domain.

#### 3.5.6.2 Semantics

The probability $p_{N'\text{-prop}}$ is updated as each data point is observed. The model requires a value for *c*, the number of properties in the learner's world. Recall that as *c* increases, the ratio between the superset (any-property) and subset (N'-property) increases; the higher this ratio, the more the subset hypothesis (N'-property) is rewarded whenever a same-property data point is encountered. Data from the MacArthur CDI (Dale & Fenson, 1996) suggest that 14-16 months olds know at least 49 adjectives. Therefore, I estimate that an 18-month old learner should be aware of at least 49 properties in the world.[16]

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

---

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

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

probability of choosing the more restricted referent set in the semantics, and thus an overestimation of the probability of the learner choosing the correct interpretation.

3.5.6.3 Linked Domain Updating

Recall that the update algorithm analyzes each data point in two domains and shifts the probability between the opposing hypotheses within each domain and across domains accordingly, as relevant. As we can see in figure 24, the learning trajectory as a function of the amount of data seen does not match our ideal learning outcome. In fact, as the learner encounters more data, the probability of the adult interpretation steadily drops to a final value of 0.171. This final value represents the product of the probability of the correct syntactic hypothesis ($p_{N'}$), which is 0.310 (1000 simulations, sd = .00377) and that of the correct semantic interpretation hypothesis ($p_{N'\text{-prop}}$), which is 0.551 (1000 simulations, sd = .00382).[17] Thus, based on the data observed, the learner is extremely unlikely to access the preferred adult interpretation for *one* (i.e., that *one* is anaphoric to strings described by N', and that the referent of *one* must have the N' property) in an utterance with two potential antecedents.

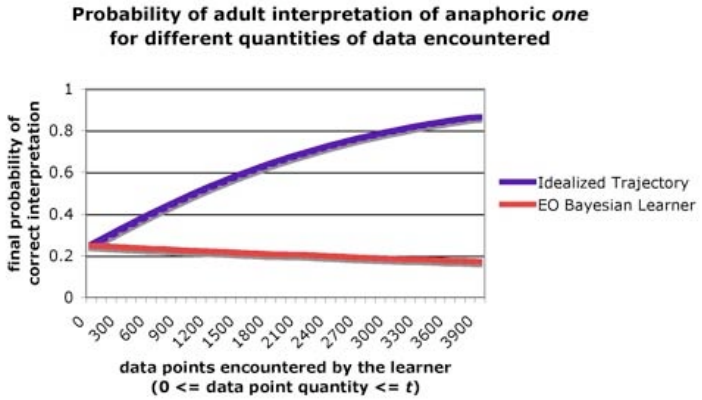



Figure 24. The EO Bayesian Learner's trajectory as a function of the amount of data encountered compared against the idealized trajectory for a learner.

3.5.6.4 Changing *t*

Recall that this model contains a parameter, *t*, which represents the amount of change the learner can undergo in the course of learning. I quantify this parameter as the number of data points the learner can use to update its probabilities. In my simulation, this was 4017, the number of data estimated during the learning period for an 18-month old. However, one might be concerned that the value of *t* might play a critical role in determining the final probability of converging on the correct

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

interpretation of anaphoric *one*. In figure 25, I show the final probability of converging on the adult interpretation of anaphoric *one* as a function of the size of *t*.

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

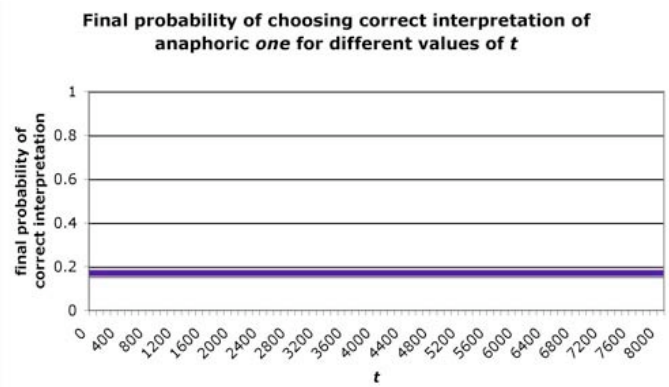



Figure 25. Final probability of the adult interpretation, given different values of *t*. All values are approximately 0.171.

However, suppose the learner encounters *fewer* data points than *t*. For instance, if the *t* for 18-month olds was actually larger than 4017, then the final probability would vary with respect to *t*. Below, I show the final probability for *t* greater than 4017 data points.

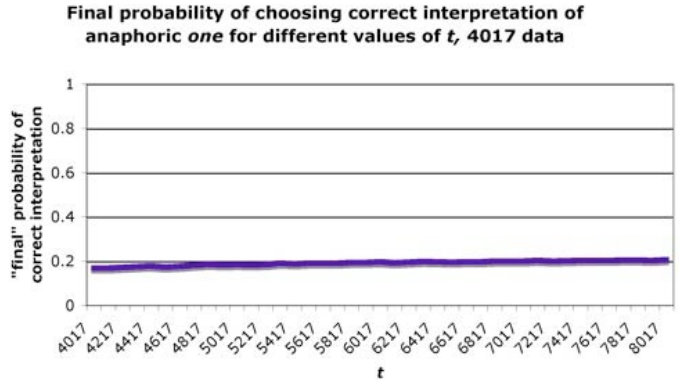



Figure 26. "Final" probability of the adult interpretation, given different values of *t* and a learning period of 4017 data points. Values approach the initial probability of 0.250, reaching 0.206 if *t* is 8017 data points (roughly twice the *t* assumed in the EO Bayesian learner).

Here, we see that the larger *t* is, the less the final probability deviates from the 0.250 initial probability. This is because each data point shifts the probability less, as *t* is larger. What we effectively see is the result of the 4017 data point cut-off (assumed for 18-month olds) not being at the end of the learning period. Thus, the learner (or learner's brain) expects to encounter more data points before settling onto the final probability; the "final" probability at 4017 data points is higher than the ultimate final probability at the end of the learning period. If the learner encounters *t* data points, the final probability will be 0.171, as we saw in figure 25 above.

### 3.5.7 The Outcome of an EO Bayesian Learner

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

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

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

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

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

In the first case, I was unable to determine a fair estimate of the amount of informative data in the environment – for example, the confidence a learner had in the type I ambiguous data (section 3.5.3.2), the quantity of type I ambiguous data that were informative (section 3.5.3.3), and the quantity of data indicating the non-

preferred adult interpretation (section 3.5.3.3). Consequently, I maximized the size of the informative data set in order to get an upper bound on the probability of converging on the correct interpretation. In what follows, I leave these assumptions as is.

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

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

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

### *3.6 On the Necessity of Domain-Specific Filters on Data Intake*

We began our discussion with the observation that a learning theory can be divided into three components: the representational format, the filters on data intake, and the learning procedure. The EO Bayesian learner attempted to solve the problem of anaphoric *one* using a prespecified representational format[19], but no domain-specific filters or learning procedures. In contrast, the model presented by R&G,

---

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

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

which also used a prespecified representational format and a domain-general learning procedure, used two domain-specific filters on data intake. This model succeeded. We can now examine (a) whether both of these filters are necessary to converge on the preferred interpretation of anaphoric *one*, and (b) whether we can derive the necessary filters in a principled fashion.

The first filter R&G's learner considers is to use only semantic data. That is, alternative syntactic hypotheses were evaluated only with respect to the predictions they made about the referents of phrases containing anaphoric *one*. These are the semantic consequences of the syntactic hypotheses. However, these hypotheses were not evaluated with respect to the predictions they made about the set of possible strings that would be available as antecedents for anaphoric *one*. So, the syntactic implications of the syntactic hypotheses were not considered. The second filter R&G's learner used was to systematically exclude type II ambiguous data. These are examples in which the antecedent for anaphoric *one* is an NP containing no modifiers (e.g. *…ball…one…*).

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

If I simulate an EO Bayesian learner that only learns via the semantic analysis of the data, the final probability for $p_{N'}$ and $p_{N'\text{-prop}}$ is 0.810. There is no deviation, since the data points consist of the 10 unambiguous data points, which are maximally informative for the N' and N'-property hypotheses, and the 183 type I ambiguous data points, which I generously assumed were maximally informative for the N' and N'-property hypotheses. Moreover, there are no countervailing data points for the alternative hypotheses ($N^0$ in the syntax and any-property in the semantics). Thus, the probability for the correct hypotheses is continually increased. Because only data with semantic consequences is considered, the type II ambiguous data is ignored and so its effect on the final probability is nullified. The final probability of converging on the correct interpretation is the product of the two probabilities, which is 0.656. This is a marked improvement over the unfiltered Bayesian learner; the semantics-only filtered Bayesian learner is nearly four times as likely to converge on the preferred adult interpretation of anaphoric *one*. However, this probability is still significantly below the ideal probability of 1.0, which would indicate absolute certainty of choosing the preferred adult interpretation. Analyzing the data only in terms of its semantic interpretation can generate significant improvement, but seems

to still fall short of leading the learner to the correct interpretation with high probability.

The second filter that R&G's model used was the exclusion of type II ambiguous data. We can now ask what happens if I follow R&G in excluding this data. This variant of the model will, like the original EO Bayesian learner, take into account both the semantic and syntactic consequences of its hypotheses, but ignore the type II ambiguous data. Note that ignoring the type II ambiguous data is an explicit filter that specifies the exclusion of this type of data, rather than having the exclusion result from a restriction on the semantic interpretation (as in the semantics-only filter we just examined).

To simulate this no-type-II-data filter, I considered only the unambiguous and type I ambiguous data points (193, by my estimate), as in the previous filter. However, both the syntactic data and semantic data was used for updating, thus making use of the link across the two domains and the fact that there are multiple sources of information. When I run the model on this data set, the final probability for the N' hypothesis in the syntax and the N'-property hypothesis in the semantics is 0.930. The product of these two, which represents the probability of converging on the correct interpretation for anaphoric *one* is 0.865. This is again a sharp improvement over the filter-free variant of the model (over 5 times more likely to converge on the correct interpretation). Additionally, the no-type-II-data filter outstrips the semantics-only filter in performance (0.865 probability against 0.656 probability), and is far closer to the ideal probability of 1.0 that indicates certainty for choosing the preferred adult interpretation of anaphoric *one*.

I now consider the consequences of using both of these filters simultaneously. Recall that the effect of the semantics-only filter, which restricted the learner to using only the semantic analysis, was that only semantic data could impact the hypotheses. This results in the type II ambiguous data being excluded from consideration, as it is uninformative with respect to the alternate semantic interpretations since it has only one potential antecedent. The no-type-II-data filter explicitly excludes type II data. So, if the model use these two filters in concert, the result is *the same* as when it used the semantics-only filter alone; the type II ambiguous data is excluded (by the semantics-only filter, due to its lack of semantic consequences, and by the no-type-II-data filter explicitly) and only semantic data can impact the probabilities associated with the hypotheses (due to the semantics-only filter). Thus, the resulting probabilities for the N' hypothesis and N'-property hypothesis are 0.810 and the probability of the preferred adult interpretation of anaphoric *one* is 0.656. Since using both filters yields an identical result to using the semantics-only filter alone, the benefit gained from using the no-type-II-filter is lost. It is therefore in the interest of the learner to apply only the no-type-II-filter. That is, the learner should ignore type II ambiguous data, but still use both syntactic and semantics data equally to update the hypothesis spaces.

To summarize, the EO Bayesian learner shows us that a learner not equipped with domain-specific filters on data intake cannot converge on the correct interpretation for anaphoric *one*. Figure 27 displays the learning trajectories and outcomes for the full set of simulations: no filter, semantics-only filter, no-type-II-data filter, both filters. As we can see, using the no-type-II-data filter by itself yields

the highest probability for the correct interpretation. Moreover, the efficacy of this filter is negated when used with the semantics-only filter. In other words, the ideal learner must use both syntactic and semantic evidence, but be restricted in which sentences it takes as opportunities to learn from.

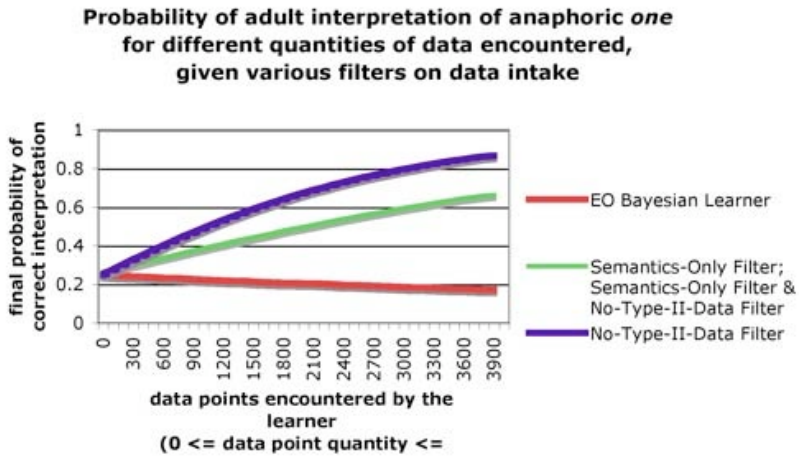



Figure 27. The Bayesian Learner's trajectory as a function of the amount of data encountered: no filters, semantics-only filter, no-type-II-data filter, and both semantics-only filter and no-type-II-data filter.

*3.7 Deriving the Necessary Domain-Specific Filter*

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

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

One suggestion would be to call on the semantics-only filter, arguing that interpreting anaphoric *one* is simply a semantic problem. This could be termed a semantocentric approach to learning, and so the syntactic implications are irrelevant for learning. The result of this strategy would be that the learner only uses the semantic consequences of the data to update the hypotheses. As we saw in the previous section, this would rule out type II ambiguous data (with a single string as potential antecedent, such as *ball*), because such data has only one semantic interpretation available (any-property)– thus, there is no uncertainty. However, as we also saw in the previous section, this causes the learner to lose the useful effect that the *syntactic* data can have. Specifically, if only semantic data are used, the benefit gained from having linked domains is lost. The learner uses only semantic data to

update the both hypothesis spaces; the learner does not also use the syntactic aspect of the data to update both hypothesis spaces. This leads to a lower probability of converging on the adult interpretation of anaphoric *one*.

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

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

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

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

### *3.8 Future Directions*

Learning anaphoric *one* is a case study that can be mined further still. For example, we can consider if learning success is possible in a hypothesis space that contains more than two hypotheses in a subset-superset relationship. Does the learner only consider two overlapping hypotheses at a time (small N' *ball* vs. larger N' *red ball*), or can the learner achieve success when, say, three hypothesis are considered concurrently (small N' *ball*, larger N' *red ball*, even larger N' *big red ball*)?

Moreover, we can open up the current hypothesis space containing only two possible N's even more if we allow the learner to entertain syntactic hypotheses

involving antecedents containing covert modifiers. Suppose, for example, that the learner hears, "Look, a bottle! Oh, and it's red! Jack doesn't have one like that." Suppose also that Jack has a non-red bottle, so it is clear that *one* refers to a red bottle in the world. The difficulty for the learner is that the antecedent of *one* in the available utterances is overtly *bottle*, but it is implicitly *red bottle* (as the bottle Jack doesn't have is a red bottle). Yet, *red bottle* does not appear overtly in the data. The learner might then need to entertain a hypothesis where the antecedent contains a covert modifier that corresponds to the property the referent in the world has, e.g. (*red*) *bottle* when the referent in the world is a red bottle. This would alter how the learner updates the probabilities associated with each hypothesis when considering information from both the potential syntactic antecedents and semantic referents in the world for anaphoric *one* data points.

I do note that before pursuing this it is worthwhile to determine via standard experimental techniques, such as those used by LWF (2003), how real learners interpret a data point of this kind. If they do interpret *one* as referring to a red bottle in the example above (and so having a linguistic antecedent of *red bottle*, even though it is not explicit in the utterance), then the question of how to expand the learner's syntactic and semantic hypothesis spaces appropriately becomes particularly relevant.

In addition, I have defined the hypothesis spaces by the number of data types that are compatible with each hypothesis (e.g. Noun, Adjective Noun, etc.). But we might also include frequency of data type, especially when considering the relative size of one hypothesis space against another. For instance, suppose the $N^0$ hypothesis space consists of data types{Noun} and the N' hypothesis space consists of data types {Noun, Adjective Noun}. The N' hypothesis space is twice as big as the $N^0$ hypothesis, under this definition. But suppose the learner has encountered 9 examples of Nouns and 1 example of Adjective Noun. Then the N' hypothesis space is only 1/10 larger than the $N^0$ hypothesis space, given the learner's current experience. This then influences the updating that occurs when encountering an ambiguous data point (Noun). The relative size of the hypothesis spaces alters over time, as the learner encounters more examples from the input. So, the impact of ambiguous data likewise alters over time. Under these conditions, is acquisition success possible without filtering the data intake? This is certainly a question worth exploring.

### *3.9 Conclusion*

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

In addition, I have tried to highlight the consequences associated with the existence of multiple, connected levels of representation in language. Because the levels of representation are linked to each other, conclusions drawn by the learner in

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

Finally, it is important to recognize that I have simulated learning only for one very specific case of grammar acquisition. However, the inherent semantic compositionality of syntactic representations provides a severe hurdle for Bayesian learning techniques that are biased towards the most restrictive hypothesis. As I have noted, as the syntactic structure grows, the set of referents in the semantics shrinks. Consequently, the most restrictive hypothesis in the syntax corresponds to the least restrictive hypothesis in the semantic interpretation, and vice versa. This makes it impossible to define a "most restrictive hypothesis" across both domains.

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