What Indirect Evidence Can Tell Us About Universal Grammar: Anaphoric One Revisited

Lisa S. Pearl and Benjamin Mis

June 9, 2011

Abstract

A controversial claim in linguistics is that children learning their native language face an induction problem: the data in their input are insufficient to identify the correct language knowledge as rapidly as children do. If this is true, children must bring some helpful learning biases to the problem, and the nature of these biases is often debated. In particular, induction problems are often used to motivate innate, domain-specific biases (sometimes called Universal Grammar). We examine the case study of English anaphoric one, an induction problem receiving recent attention in the computational modeling literature, and consider whether indirect evidence leveraged by an online probabilistic learner from a broader input set could be effective. We find our learner can reproduce child learning behavior, given child-directed speech. We discuss what is required for acquisition success, and how this impacts the larger debate about Universal Grammar.

1 Induction problems in language acquisition

One of the most controversial claims in developmental and theoretical linguistics is that children learning their native language face an induction problem, often called the "Poverty of the Stimulus" (Chomsky, 1980a, 1980b; Crain, 1991; Lightfoot, 1989), the "Logical Problem of Language Acquisition" (Baker, 1981; Hornstein & Lightfoot, 1981), or "Plato's Problem" (Chomsky, 1988; Dresher, 2003). Simply put, this is the claim that the data in children's input are insufficient to identify the correct language knowledge - or at least, to identify the correct knowledge as quickly as children seem to (Legate & Yang, 2002; Lightfoot, 1982).

If this is true, then children must bring something to the language acquisition problem - and the nature of this "something" is often debated (e.g., see Crain and Pietroski (2002); Fodor (1998a, 1998b); Foraker, Regier, Khetarpal, Perfors, and Tenenbaum (2009); Lidz, Waxman, and Freedman (2003); McMurray and Hollich (2009); Pearl (2007); Pearl and Lidz (2009); Perfors, Tenenbaum, and Regier (2011); Pullum and Scholz (2002); Regier and Gahl (2004); Scholz and Pullum (2002); Soderstrom, Conwell, Feldman, and Morgan (2009), among others). There at least three dimensions we can consider about the nature of children's learning biases:

(i) Are they *domain-specific* (and are only used for learning language) or *domain-general* (and are used when learning anything)?

(ii) Are they *innate* (and so part of the human biological endowment) or *derived* from prior experience (probably prior experience with language data)?

(iii) Are they about *what to learn* (and so may restrict the learner's hypotheses explicitly) or about *how to learn* (and so may restrict the learner's hypotheses implicitly)?

These questions are particularly important, as induction problems in language acquisition are often used to motivate innate, domain-specific knowledge about language (sometimes called *Universal Grammar* (Chomsky, 1965)). However, as we can see from the distinctions above, there are clearly other kinds of learning biases that might be used. So, if Universal Grammar is to be

supported, it is worth discovering if induction problems require innate, domain-specific learning biases.

1.1 Induction problems and data

Traditionally when identifying potential induction problems in language acquisition, it has been assumed that only directly related data are informative to the child. We might call this the *direct evidence assumption*. The basic intuition of the direct evidence assumption is that in order to learn some linguistic knowledge L, a learner observes examples of L in the linguistic input. It's also possible that a learner (particularly a statistical learner) can be sensitive to *indirect negative evidence* related to the directly informative data, and so will notice what direct evidence examples are missing from the input.

For example, when learning how to form complex yes/no questions in English, a learner pays attention to examples of complex yes/no questions like (1a) and potentially notices the absence of ungrammatical complex yes/no questions like (1b).

(1) Complex yes/no question examples

(a) Is the boy who is in the corner t_{is} happy?

(b) *Is the boy who t_{is} in the corner is happy?[1]

When learning the representation of English anaphoric one, a learner pays attention to examples of one being used anaphorically (2a) and potentially notices the absence of ungrammatical uses of one like (2b).

(2) Anaphoric one examples

(a) Look - a red bottle. Oh, look! Another one.

(b) *She sat by the side of the river, and he sat by the one of the road.

[1]The * will be used to indicate ungrammaticality.

When learning to form complex wh-questions in English, a learner pays attention to examples of complex wh-questions in English (3a-c) and potentially notices the absence of ungrammatical examples like (3d).

(3) Complex wh-question examples

(a) What did the teacher think t_{what} inspired the students?

(b) Who did the teacher think the letter from the soldier inspired t_{who}?

(c) Who t_{who} thought the letter from the soldier inspired the students?

(d) *Who did the teacher think the letter from t_{who} inspired the students?

However, there is another kind of data that could be informative to a learner: *indirect positive evidence*. This refers to observable data that may not be directly informative for the linguistic knowledge in question, but can nonetheless be informative if viewed the correct way by the learner (for example, due to a learner's helpful learning biases). If children can recognize and use indirect positive evidence, this broadens the set of informative data and may help solve some of the induction problems facing children. In fact, some recent computational modeling approaches have been exploring the utility of this kind of indirect evidence for different induction problems (e.g., see Foraker et al. (2009); Kam, Stoyneshka, Tornyova, Fodor, and Sakas (2008); Perfors et al. (2011); Reali and Christiansen (2005)).

Given this, there are two broad questions we can explore with respect to language acquisition. First, when induction problems exist, what does it take to solve them? We can examine not only the direct positive evidence and indirect negative evidence available, but also the indirect positive evidence available that a learner could recognize and use. Given this expanded data set, we can then explore the nature of the learning biases necessary to solve the induction problem.

Related to this is the second broad question: How can the necessary learning biases inform us about the process of acquisition? If we understand the data and the learning biases a child has to work with, then we have a clearer picture of the trajectory of acquisition, as defined by the

sequence of knowledge states a child passes through. More specifically, given the data and the learning biases we believe the child has available, we can predict what knowledge the child should have at a given point during learning. Different knowledge states predict different observable behavior, and we can then see whether the predicted behavior matches empirical observations of children's behavior. If it does, then we know more about the learning process that could lead to that observable behavior because we know how the child could use the available data to produce that observable behavior.

1.2 Case study: Anaphoric one

The potential induction problem presented by English anaphoric one (from example (2) above) has received considerable recent attention (e.g., Akhtar, Callanan, Pullum, and Scholz (2004); Foraker et al. (2009); Lidz et al. (2003); Lidz and Waxman (2004); Pearl (2007); Pearl and Lidz (2009); Pullum and Scholz (2002); Regier and Gahl (2004); Tomasello (2004); among others). The original proposal for learning anaphoric one required children to have innate domain-specific knowledge about the structure of language, as part of the child's Universal Grammar (Baker, 1978). However, more recent studies have suggested alternative solutions involving innate domain-general statistical learning abilities, usually coupled (either implicitly or explicitly) with input restrictions that arise from domain-specific learning constraints (Foraker et al., 2009; Pearl & Lidz, 2009; Regier & Gahl, 2004) and sometimes also with knowledge that is likely to be innate and domain-specific (Foraker et al., 2009). Here, we consider whether indirect evidence leveraged from a broader input set could lead children to the correct knowledge about anaphoric one. If so, we can then refine the current views on the learning biases required for successful acquisition - and specifically, the nature of those biases.

We first briefly discuss adult and child knowledge of anaphoric one, and then highlight what the learning problem is - in particular, why anaphoric one has been considered an induction problem for language acquisition. We then review previous proposals for how to learn the correct represen-

tation of anaphoric one from the available input. Following this, we motivate why a child might view a broader input set as informative for anaphoric one, and discuss the different kinds of information that are available in informative data points. We then present an online Bayesian learner adapted from Pearl and Lidz (2009) that uses this broader data set, and find that our learner is indeed capable of reproducing the child behavior associated with correct knowledge of anaphoric one - notably, without imposing any domain-specific input restrictions. In addition, we compare our learner's performance on the broader data set to performance on the restricted datasets previously proposed, and find that it is the broader data set that produces the correct learning behavior rather than something inherent in the probabilistic learning model.

Our model also provides a way to explicitly test the assumption in the behavioral study by Lidz et al. (2003) that correct behavior during an experiment testing children's interpretation of anaphoric one indicates the child has the correct representation for anaphoric one. We find that our modeled learner would both produce the correct behavior in that experiment and infer the correct representation at the time it produces that behavior - surprisingly, even if the learner does *not* generally have the correct representation for one. We conclude with discussion of what a child requires in order to solve the induction problem for anaphoric one, what this tells us about the acquisition trajectory for one, and how this impacts the larger debate about Universal Grammar.

2 English anaphoric one

2.1 Adult knowledge

An example of anaphoric one and the various components involved in its interpretation is in (4).

(4) Situation: Two red bottles are present.

Utterance: "Look - a red bottle! Oh, look - another one!"

Interpretation of one:

syntactic antecedent of one = "red bottle"

semantic referent of one = RED BOTTLE

The adult representation of English anaphoric one has both a syntactic and semantic component. In order to interpret an utterance like (4) ("Look - a red bottle! Oh, look - another one!"), the listener must first identify the syntactic antecedent of one, i.e., what string one is standing in for. In (4), adults generally interpret one's syntactic antecedent as "red bottle", so the utterance is equivalent to "Look - a red bottle! Oh, look - another *red bottle*!".[2]

Then, the listener uses this syntactic antecedent to identify the semantic referent of one, e.g., what object in the world one is referring to. Given the syntactic antecedent "red bottle", adults interpret the referent of one as a bottle that is red (RED BOTTLE), as opposed to just any bottle (BOTTLE). That is, the one the speaker is referring to is a bottle that specifically has the property red and this utterance would sound somewhat strange if the speaker actually was referring to a purple bottle.

According to standard linguistic practice, the string "red bottle" has the structure in (5), while "a red bottle" has the structure in (6). The bracket notation corresponds to the syntactic phrase structure tree in figure 1.

(5) $[_{N'}$ red $[_{N'}$ $[_{N^0}$ bottle]]

(6) $[_{NP}$ a $[_{N'}$ red $[_{N'}$ $[_{N^0}$ bottle]]]

The syntactic category N^0 can only contain noun strings (e.g., "bottle"), and the category NP contains any noun phrase (e.g., "a bottle", "a red bottle"). The syntactic category N' is larger than N^0 but smaller than NP, and can contain both noun strings (e.g., "bottle") and noun+modifier strings (e.g., "red bottle"). Note that the noun-only string "bottle" can be labeled both as syntactic category N' (7a) and syntactic category N^0 (7b) (this also can be seen in figure 1, where "bottle"

[2]There are cases where the "bottle" interpretation could become available (and so a purple bottle would be a valid referent since it is in fact a bottle), and these often have to do with contextual clues and special emphasis on particular words in the utterance (Akhtar et al., 2004). The default interpretation, however, seems to be "red bottle". We discuss these non-default interpretations more in section 6.2.

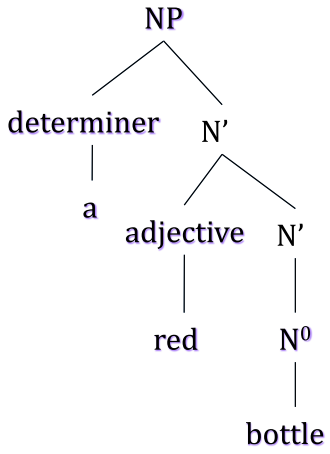

Figure 1: Phrase structure tree corresponding to the bracket notation in examples (5) and (6).

projects to both N^0 and N').[3]

(7a) $[_{N'}\ [_{N^0}$ bottle]]

(7b) $[_{N^0}$ bottle]

Linguistic theory posits that anaphoric elements (like one) can only have antecedents of the same syntactic category. Since one's antecedent can be "red bottle", then one should be category N' in these cases. Notably, if the syntactic category of one were instead N^0, one could not have "red bottle" as its antecedent; instead, it could only have noun-only strings like "bottle", and we would interpret (4) as "Look - a red bottle! Oh, look - another *bottle*!" In that case, we should be perfectly happy to have one's referent be a purple bottle. Since we do not have this interpretation in (4) and instead prefer one's antecedent to be "red bottle" (and its referent to be a RED BOTTLE), one's syntactic category must be N' here.

One way to represent adult knowledge is as in (8). On the syntax side, the syntactic category of one is N' and so one's antecedent is also N'. On the semantic side, the property mentioned in

[3]We note that while we use the labels N' and N^0, other theoretical implementations may use different labels to distinguish these hierarchical levels. The actual labels themselves are immaterial - it is only relevant for our purposes that these levels are distinguished the way we have done here, i.e., that "red bottle" and "bottle" are the same label (N' here), while "bottle" can also be labeled with a smaller category label (N^0 here). However, see discussion in section 6.2 for what happens with alternate theoretical representations that additionally differentiate "red bottle" from "bottle".

the potential antecedent (e.g., "red") is important for the referent to have. This has a syntactic implication for one's antecedent: the antecedent is the larger N' that includes the modifier (e.g., "red bottle", rather than "bottle").

(8) Adult anaphoric one knowledge in utterances like

"Look - a red bottle! Do you see another one?"

(a) Syntactic structure: category N'

(b) Semantic referent and antecedent: The mentioned property ("red") in the potential antecedent is relevant for determining the referent of one. So, one's antecedent is $[_{N'}$ red $[_{N'}$ $[_{N0}$ bottle]]] rather than $[_{N'}$ $[_{N0}$ bottle]].

2.2 Child knowledge

Behavioral evidence from Lidz et al. (2003) (henceforth **LWF**) suggests that young children also have this same interpretation for utterances like (4).[4] Using an intermodal preferential looking paradigm (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987; Spelke, 1979), LWF examined the looking behavior of 18-month-olds when hearing an utterance like "Look, a red bottle! Do you see another one?". The 18-month-olds demonstrated a significant preference for looking at the bottle that was red (as compared to a bottle that was some other color), just as adults would do. Thus LWF interpreted this to mean that by 18 months, children have acquired the same representation for anaphoric one that adults have. We note that it is an assumption that correct behavior in this experiment indicates the correct representation for one - it is possible that children could produce that behavior even if they have a different representation for one (as we will explore below in section 5.3). However, the empirical fact is that children's behavior appears adult-like at 18 months when interpreting anaphoric one utterances like these.

[4]Though see Tomasello (2004) for a critique of LWF's interpretation of their experiment and Lidz and Waxman (2004) for a convincing rebuttal.

3 Learning anaphoric one

3.1 The learning problem

Learning the correct representation for anaphoric one is difficult because many anaphoric one data are ambiguous with respect to what syntactic category one is, even if children already know that the choice is between N' and N^0. Moreover, as we see in figure 1, sometimes there is more than one N' antecedent to choose from (e.g., "red bottle": $[_{N'}$ red $[_{N'}$ $[_{N^0}$ bottle]]] vs. "bottle": $[_{N'}$ $[_{N^0}$ bottle]]), which means that there is also ambiguity with respect to the semantic referent (e.g., RED BOTTLE vs. any BOTTLE). Examples (9) and (10) demonstrate two kinds of ambiguous data, one which is ambiguous syntactically (9) and the other which is ambiguous both semantically and syntactically (10).

(9) Syntactic (**Syn**) Ambiguity

Situation: There are two bottles present.

Utterance: "Look, a bottle! Oh look - another one!"

(10) Semantic and Syntactic (**Sem-Syn**) Ambiguity

Situation: There are two red bottles present.

Utterance: "Look, a red bottle! Oh look - another one!"

Syn ambiguous data like (9) do not clearly indicate the category of one, even though the semantic referent is clear. In (9), the semantic referent must be BOTTLE since the antecedent can only be "bottle". But, is the syntactic structure $[_{N'}$ $[_{N^0}$ bottle]] or just $[_{N^0}$ bottle]? Notably, if the child held the mistaken hypothesis that one was category N^0, this data point would not conflict with that hypothesis since it is compatible with the antecedent being $[_{N^0}$ bottle].

Sem-Syn ambiguous data like (10) are unclear about both the referent and the category of one. In (10), if the child held the mistaken hypothesis that the referent is simply BOTTLE (unlike the adult interpretation of RED BOTTLE), this would not be disproven by this data point - there is in fact

another bottle present. That it happens to be a red bottle is merely a coincidence. The alternative hypothesis is that the referent is RED BOTTLE (this is the adult interpretation), and it's important that the other bottle present have the property red. Since both these options for semantic referent are available, this data point is ambiguous semantically. This data point is ambiguous syntactically for the same reason Syn data like (9) are: if the referent is BOTTLE, then the antecedent is "bottle", which is either N^0 or N'.

Fortunately, there are some unambiguous data available like (11), but these require a very specific conjunction of situation and utterance.

(11) Unambiguous (**Unamb**) data

Situation: Both a red bottle and a purple bottle are present.

Utterance: "Look - a red bottle! There doesn't seem to be another one here, though."

In (11), if the child mistakenly believes the referent is just BOTTLE, then the antecedent of one is "bottle" and it's surprising that the speaker would claim there's not "another bottle here", since another bottle is clearly present. Thus, in order to make sense of this data point, it must be that the property "red" is important, so the semantic referent must be RED BOTTLE (and indeed, there isn't another red bottle present, so the utterance is then a reasonable thing to say). The corresponding syntactic antecedent is "red bottle", which has the syntactic structure $[_{N'}$ red $[_{N'}$ $[_{N^0}$ bottle]]] and indicates one's category is N'.

Unfortunately, unambiguous data were presumed to be very rare. LWF discovered in their corpus analysis that a mere 0.25% of child-directed anaphoric one utterances were unambiguous data. For this reason, the debate has arisen about how children might solve this acquisition problem as rapidly as they do.

3.2 Innate, domain-specific knowledge

An early proposal (Baker, 1978) (henceforth, **Baker**) assumed that only unambiguous data were informative. Given the sparsity of these data, it was assumed that children could not learn the correct representation from the data available - there was an induction problem. Instead, it was proposed that children possess domain-specific knowledge about the structure of language. In particular, children innately know that anaphoric elements (like one) cannot be syntactic category N^0. Instead, children automatically rule out that possibility from their hypothesis space, and simply know that one is category N'.[5] Thus, this proposal assumes an innate, domain-specific learning bias concerning the knowledge being acquired.

3.3 Domain-general learning abilities and domain-specific knowledge

3.3.1 Regier & Gahl 2004

Regier and Gahl (2004) (henceforth **R&G**) noted that Sem-Syn data like (10) could be leveraged to learn the correct representation for anaphoric one. Specifically, a probabilistic learner could track how often a property that was mentioned was important for the referent to have (e.g., when "red" was mentioned, was the referent just a BOTTLE or specifically a RED BOTTLE?). If the referent keeps having the property mentioned in the potential antecedent (e.g., keeps being a RED BOTTLE), this is a suspicious coincidence unless one's antecedent actually does include the modifier describing that property (e.g., "red bottle"). If the antecedent includes the modifier, this then indicates that one's antecedent is N', since N^0 cannot include modifiers. One would then be N' as well, since it is the same category as its antecedent.

The R&G data set consisted of both unambiguous data and Sem-Syn ambiguous data, and their online Bayesian learner was able to learn the correct interpretation for anaphoric one. No innate,

[5] Note that this proposal only deals with the syntactic category of one and does not provide a solution for how to choose between two potential antecedents that are both N', such as "red bottle": [$_{N'}$ red [$_{N'}$ [$_{N^0}$ bottle]]] vs. "bottle": [$_{N'}$ [$_{N^0}$ bottle]]. It does, however, rule out the potential antecedent [$_{N^0}$ bottle].

domain-specific knowledge was required to converge on the correct representation for anaphoric one. Instead, once the hypothesis space was defined, a learner with innate domain-general statistical learning abilities could succeed by leveraging this particular set of ambiguous data.

3.3.2 Pearl & Lidz 2009

Pearl and Lidz (2009) (henceforth **P&L**) noted that if the child had to learn the syntactic category of one, then an "equal-opportunity" (**EO**) learner able to leverage ambiguous data (like R&G's learner) would view Syn ambiguous data like (9) as informative. Unfortunately, P&L found that Syn ambiguous data lead an online Bayesian learner to the wrong syntactic category for one (i.e., one=N^0), and in fact far outnumber the Sem-Syn ambiguous and unambiguous data combined (about 20 to 1 in their corpus analysis). Thus, a probabilistic learner like R&G proposed would need to explicitly filter out the Syn ambiguous data. P&L suggested that this kind of filter is domain-specific, since it involves ignoring a specific kind of linguistic data. However, they speculate how this restriction could be derived from innate domain-general learning preferences.[6] Thus, P&L find that a probabilistic learner using innate domain-general learning abilities also needs a domain-specific input restriction to succeed, though this input restriction may be derived from other innate domain-general learning biases.

3.3.3 Foraker et al. 2009

Foraker et al. (2009) (henceforth **F&al**) focused on identifying the syntactic category of one, the original problem considered by Baker, and applied an ideal Bayesian learner to the syntactic input alone. In order to leverage the distributional information in the syntactic input, their learner employed subtle conceptual knowledge to identify the likely syntactic category for one. Specifically, their learner was able to distinguish syntactic *complements* from syntactic *modifiers*, where a syn-

[6] In particular, they suggest that a learner who learns only in cases of uncertainty in the local context would ignore Syn ambiguous data while still heeding unambiguous and Sem-Syn ambiguous data (see Pearl and Lidz (2009) for more explicit discussion of this proposal).

tactic complement is "conceptually evoked by its head noun" and indicates the noun string is N^0, while a modifier is not and indicates the noun string is N'. Figure 2 shows the syntactic structure associated with modifiers and complements, where a modifier like "with dots" is sister to N' and a complement like "of the road" is sister to N^0.

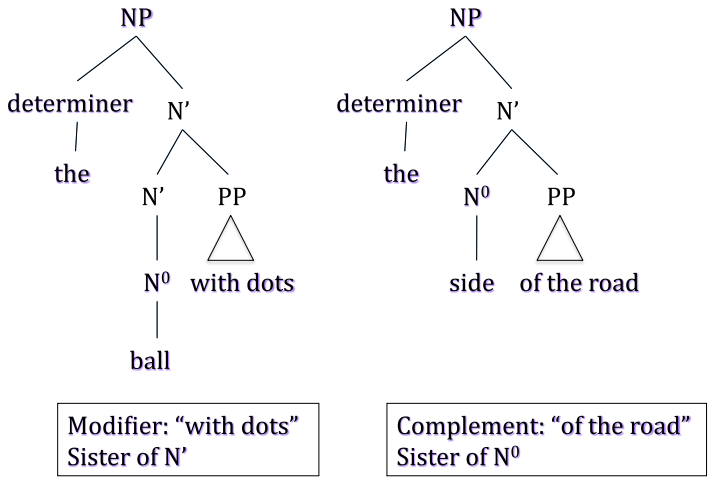

Figure 2: Phrase structure trees corresponding to a modifier and a complement.

Because of this, one (being N') cannot appear with complements, since complements adjoin with N^0. This is why "one of the road" is ungrammatical (12a), while "one with dots" is grammatical (12b).

(12a) *Lily waited by the side of the building while Jack sat by the one of the road.

(12b) Lily was fond of the ball with stripes while Jack preferred the one with dots.

Thus, simple nouns (known to be N^0 and project to N') can appear with both complements ("side of the road") when they are N^0 and modifiers ("ball with dots") when they are N', while one only occurs with modifiers ("one with dots"). F&al's learner uses this indirect negative evidence, and notes the absence of one being used with complements. This then indicates that one is not N^0, but rather the larger syntactic category N'.

While there were not many informative one data points in their data, F&al's ideal learner was

able to learn the correct syntactic category for one. In order to do this, their learner appears to require a domain-specific input restriction to syntactic data (rather than also using semantic information, as Baker, R&G, and P&L's learners do). In addition, the specific syntactic information their learner leverages appears to require (possibly innate) domain-specific knowledge in order to both realize the subtle semantic distinction between complements and modifiers and this distinction's implication for the syntactic category of the corresponding noun.

3.3.4 Comparison of previous proposals

Table 1 compares the learning biases required by previous proposals for how to learn anaphoric one, including a description of the bias, and where it falls on the different dimensions of innate vs. derived, domain-specific vs. domain-general, and what to learn vs. how to learn. Note that only two biases (the one proposed by Baker, and potentially one proposed by F&al) are innate and domain-specific.

Table 1: Learning biases required by previous proposals.

Proposal	**Bias**	**Innate**	**Derived**	**Dom-Spec**	**Dom-Gen**	**What To Learn**	**How To Learn**
Baker	know one is not N^0	*		*		*	
R&G, P&L filtered	recognize suspicious coincidence of antecedent property	*			*		*
R&G, P&L filtered	ignore Syn ambig data	*	*	*	*	*	*
F&al	know what syntactic complements vs. modifiers imply	?		*		*	
F&al	leverage complement vs. modifier distribution		*		*		*

4 A broader view of informative data

Instead of restricting the input set, we consider expanding it beyond Unambiguous (11), Sem-Syn Ambiguous (10), and Syn Ambiguous (9) data. Consider that there are other anaphoric elements in the language besides one, such as pronouns like it, him, her, etc. - thus, the ability for a linguistic element to stand in for something else is not unique to one. These other pronouns are category NP, since they replace an entire noun phrase (NP) when they are used (13):

(13) "Look at the cute penguin. I want to hug it/him/her."

≈ "Look at the cute penguin. I want to hug *the cute penguin*."

Here, the antecedent of the pronoun it/him/her is the NP "the cute penguin":

(14) [$_{NP}$ the [$_{N'}$ cute [$_{N'}$ [$_{N^0}$ penguin]]]]

In fact, it turns out that one can also have an NP antecedent:

(15) "Look! A red bottle. I want one."

≈ "Look! A red bottle. I want *a red bottle*."

We note that the issue of one's syntactic category only occurs when one is being used in a syntactic environment that indicates it is smaller than NP (such as in utterances (4), (9), (10), and (11)). However, since one is similar to other pronouns semantically (by being anaphoric) and shares some syntactic distribution properties with them (since it can appear as an NP), a learner could decide that information gleaned from other pronouns is relevant for interpreting one.

Following R&G's idea of tracking suspicious coincidences, a learner could track how often a property mentioned in the potential antecedent (e.g., "red" in "a red bottle" in (15)) is important for the referent to have. Crucially, we can apply this not only to data points where one is $<$NP ((9) and (11)), but also to data points where pronouns are used anaphorically and in an NP syntactic environment ((13) and (15)). When the potential antecedent mentions a property and the pronoun is used as an NP, the antecedent is necessarily also an NP, and so necessarily includes the men-

tioned property (e.g., "a red bottle"). Data points like (13) and (15) are thus unambiguous both syntactically (category=NP) and semantically (the referent must have the mentioned property). We will refer to them as unambiguous NP (**Unamb NP**) data points, and these are the additional data points our learner (the **P&M** learner) will learn from.

Like the R&G and P&L learners, our learner differs from the Baker learner by learning from data besides the unambiguous <NP data. However, our learner differs from the learners in R&G and P&L by learning from data containing anaphoric elements besides one.[7] Table 2 shows which learners use which data.

Table 2: Data sets used by learners.

Data type	Example	Learners
Unamb <NP	"Look - a red bottle! There doesn't seem to be another one here, though."	Baker, R&G, P&L's EO, P&M
Sem-Syn Ambig	"Look - a red bottle! Oh, look - another one!"	R&G, P&L's EO, P&M
Syn Ambig	"Look - a bottle! Oh, look - another one!"	P&L's EO, P&M
Unamb NP	"Look a red bottle! I want it/one."	P&M

4.1 Information in the data

There is a variety of information in referential data points. Figure 3 represents the information dependencies in any data point where a pronoun is used anaphorically and there is a potential antecedent that has been mentioned recently.[8]

Under SYNTACTIC USAGE, a learner can observe which pronoun is used (e.g., it, one, etc.). The syntactic category depends on which pronoun is used (e.g., NP, N', or N^0 for one). The learner can also observe the syntactic environment in which the pronoun is used, which depends on the latent syntactic category (e.g., "another one" indicates a syntactic environment of <NP, which

[7] Our learner also differs from the F&al learner by leveraging both syntactic and semantic information, instead of just syntactic information.

[8] Note that this represents a generative model for a referential data point, rather than a decision tree a learner would use to make inferences. That is, inferences flow both directions along the information dependencies.

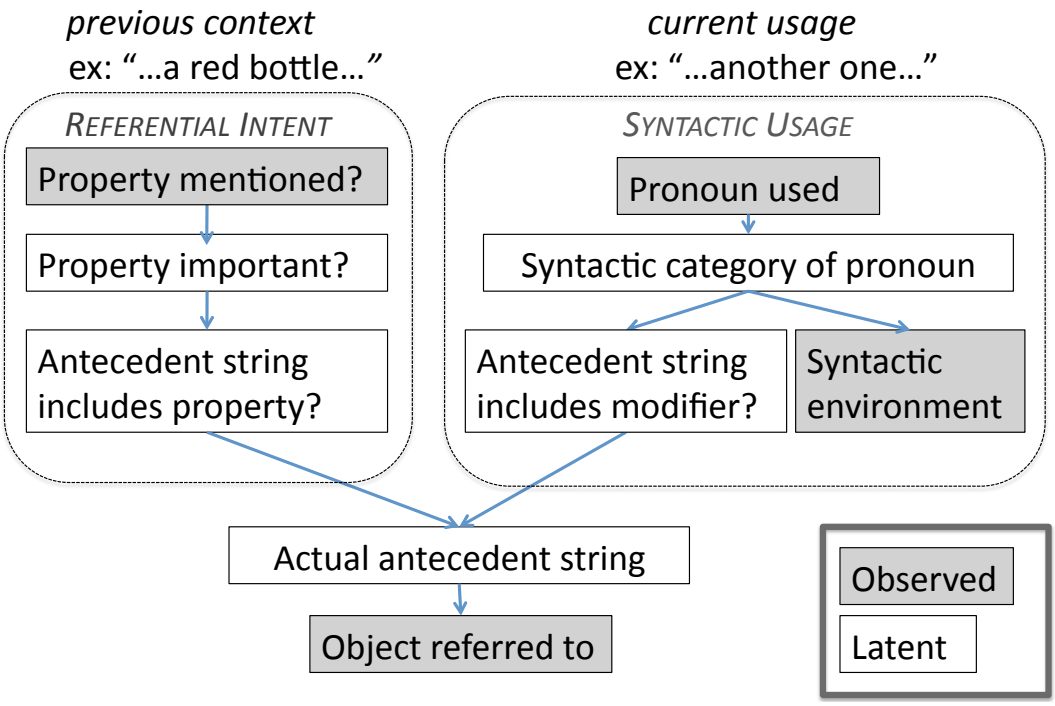

Figure 3: Information dependencies in referential data points.

means the category is N' or N^0). The syntactic category also determines whether the antecedent string can contain a modifier (e.g., category N^0 cannot, since it only allows noun-only strings like "bottle").

Under REFERENTIAL INTENT, a learner can observe whether the potential antecedent in the previous context mentioned a property or not (e.g., "a red bottle" vs. "a bottle"). If a property was mentioned, it is a latent variable whether the mentioned property was important for the referent of the pronoun to have. This then determines whether the antecedent string must include that property (e.g., it must if the property is important, and it must not if the property is not important).

Both the antecedent string variables determine the content of the actual antecedent string (e.g., if both a modifier and a property must be included, the antecedent would be "red bottle" rather than simply "bottle"). Finally, the antecedent string determines what object is being referred to, and whether that object has the mentioned property (e.g., whether it's a RED BOTTLE when the

previous context was “a red bottle”). This is observable (e.g., the learner can ascertain if the bottle that one refers to is in fact red).

These variables can take on the values shown in table 3.[9] The data types used by the different learning proposals have the observable and latent values in Table 4.

Table 3: Variable values in informative referential data points.

REFERENTIAL INTENT	property mentioned? $\in$ {Yes, No} property important $\in$ {Yes, No, N/A} antecedent string includes property? $\in$ {Yes, No, N/A}
SYNTACTIC USAGE	pronoun used $\in$ {one, it, him, her, etc.} syntactic category of pronoun $\in$ {NP, N', N^0} syntactic environment $\in$ {NP, <NP} antecedent string includes modifier? $\in$ {Yes, No, N/A}
COMBINED	actual antecedent string $\in$ {“red bottle”, “bottle”, etc.} object referred to $\in$ { has property, does not have property, N/A}

Table 4: Data types and variable values.

	Variable	**Unamb <NP**	**Sem-Syn Ambig**	**Syn Ambig**	**Unamb NP**
Observable	Prop Mentioned	Yes	Yes	No	Yes
	Pronoun	one	one	one	it, one, etc.
	Syn Env	<NP	<NP	<NP	NP
	Object	has property	has property	N/A	has property
Latent	Prop Important	Yes	Yes, No	N/A	Yes
	Antec Has Prop	Yes	Yes, No	N/A	Yes
	Syn Category	N'	N', N^0	N', N^0	NP
	Antec Has Mod	Yes	Yes, No	N/A	Yes
	Antec String	ex: “red bottle”	ex: “red bottle”, “bottle”	ex: “bottle”	ex: “a red bottle”

Unambiguous < NP data have a property mentioned in the potential antecedent (e.g., “Look - a red bottle!”), use the pronoun one (e.g., “There doesn’t seem to be another one here, though.”),

[9] Note that if no property was mentioned, the decision as to whether the mentioned property was important (property important?) is moot, and hence has the value N/A. This same logic applies to the decision about whether the antecedent string includes the modifier (antecedent string includes modifier?), whether the antecedent string includes the property (antecedent string includes property?), and whether the observed object has the property (object referred to).

have a syntactic environment that indicates the pronoun is smaller than NP (e.g., “another one”), and refer to an object that has the property mentioned (e.g., RED BOTTLE). Because these data are unambiguous, the learner can infer that the antecedent string includes the property (e.g., “red bottle”), which means the antecedent has a modifier (from the syntactic perspective) and also has a property (from the referential perspective). This indicates that the mentioned property is important and the syntactic category of the antecedent (and so of one) is N’.

Sem-Syn ambiguous data have a property mentioned in the potential antecedent (e.g., “Look - a red bottle!”), use the pronoun one (e.g., “Look - another one!”), have a syntactic environment that indicates the pronoun is smaller than NP (e.g., “another one”), and refer to an object that has the property mentioned (e.g., RED BOTTLE). Because these data are ambiguous both semantically and syntactically, the antecedent is unclear (e.g., “red bottle” or “bottle”). This means it is also unclear whether the antecedent includes a modifier and a property, whether the mentioned property is important, and what the syntactic category is (N’ or N^0).

Syn ambiguous data do not have a property mentioned in the potential antecedent (e.g., “Look - a bottle!”), use the pronoun one (e.g., “Look - another one!”), have a syntactic environment that indicates the pronoun is smaller than NP (e.g., “another one”), and refer to the object that is mentioned without indicating a property of that object. Because these data do not mention a property in the potential antecedent, they are uninformative about whether the antecedent should have a modifier that indicates the property, and whether a mentioned property is important. In addition, while the antecedent is unambiguous (e.g., “bottle”), the syntactic category is not (it could be N’ or N^0).

Unambiguous NP data have a property mentioned in the potential antecedent (e.g., “Look - a red bottle!”), use a number of different referential pronouns (e.g., “I want it/one”), have a syntactic environment that indicates the pronoun is category NP (e.g., “want one”), and refer to an object that has the property mentioned (e.g., RED BOTTLE). Because these data are unambiguous, the learner can infer the antecedent string is the entire NP (e.g., “a red bottle”), and note that the

antecedent string includes a modifier indicating the property (e.g., "red"). This in turn indicates that the property is important.

5 The online probabilistic learning framework

We now present an online probabilistic learning framework that uses the different kinds of information available in referential data points.

5.1 Important quantities

The two components of the correct representation for anaphoric one are (a) that a property mentioned in the potential antecedent is important for the referent of one to have (more specifically, p(*property important=yes* | *property mentioned=yes*)), and (b) that one is category N' when it is not an NP (more specifically, p(*category=N'* | *syntactic environment=<NP*)). These correspond to "property important?" and "syntactic category of pronoun" in Figure 3. We represent the probability of the former as p_I and the probability of the latter as $p_{N'}$. Note that p_I can only take the values Yes and No and $p_{N'}$ can only take the values N' or N^0.

We follow the update methods in P&L, and use equation (16) adapted from Chew (1971), which assumes p comes from a binomial distribution and the beta distribution is used to estimate the prior:

$$p_x = \frac{\alpha + data_x}{\alpha + \beta + totaldata_x}, \alpha = \beta = 1 \tag{16}$$

Parameters α and β represent a very weak prior when set to 1. The variable $data_x$ represents how many informative data points indicative of x have been observed, while $totaldata_x$ represents the total number of potential x data points observed. After every informative data point, $data_x$ and

$totaldata_x$ are updated as in (17), and then p_x is updated using equation (16). The variable ϕ_x indicates the probability that the current data point is an example of an x data point. For unambiguous data, ϕ_x = 1; for ambiguous data $\phi_X < 1$.

$$data_x = data_x + \phi_x \tag{17a}$$

$$totaldata_x = totaldata_x + 1 \tag{17b}$$

Probability p_I is updated for Unambiguous <NP data, Sem-Syn Ambiguous data, and Unambiguous NP data only - Syn Ambiguous data do not mention a property, and so are uninformative for p_I. Probability $p_{N'}$ is updated for Unambiguous <NP data, Sem-Syn Ambiguous data, and Syn Ambiguous data only - Unamb NP data indicate the category is not <NP, and so are uninformative for $p_{N'}$.

The value of ϕ_x depends on data type. We can derive the value of ϕ_I by using the information dependencies in Figure 3, and the basic Bayes equation. ϕ_I uses equation (18), which includes π (what pronoun was mentioned), σ (what the syntactic environment is), μ (whether the previous context mentioned a property), ω (whether the object has the mentioned property), and I (property important=yes). Note that p_I is predicated on a property being mentioned, which is why μ = yes.

$$\phi_I = p(I|\pi,\sigma,\mu = yes,\omega) = \frac{p(\pi,\sigma,\omega|I,\mu = yes) * p_I}{p(\pi,\sigma,\omega|\mu = yes)} \tag{18}$$

Unambiguous <NP and Unambiguous NP data end up having ϕ_I=1, which is intuitively satisfying since they unambiguously indicate that the property is important for the referent to have. Sem-Syn ambiguous data end up having ϕ_I calculated as in (19):

$$\phi_I = \frac{\rho_1}{\rho_1 + \rho_2 + \rho_3} \tag{19}$$

where

$$\rho_1 = \quad p_{N'} * \tfrac{m}{n+m} * p_I \tag{20a}$$

$$\rho_2 = \quad p_{N'} * \tfrac{n}{n+m} * (1 - p_I) * \tfrac{1}{t} \tag{20b}$$

$$\rho_3 = \quad (1 - p_{N'}) * (1 - p_I) * \tfrac{1}{t} \tag{20c}$$

In (20), *m* and *n* refer to how often N' strings are observed to contain modifiers (*m*) (e.g., "red bottle"), as opposed to containing only nouns (*n*) (e.g., "bottle"). These help determine the probability of observing an N' string with a modifier (20a), as compared to an N' string that contains only a noun (20b). Parameter *t* indicates how many property types there are in the learner's hypothesis space, which determines how suspicious a coincidence it is that the object just happens to have the mentioned property when there are *t* properties (types of objects) the learner is aware of.

The quantities in (20) correlate with anaphoric one representations. For ρ_1 (which is the adult representation), the syntactic category is N' ($p_{N'}$), a modifier is used ($\frac{m}{n+m}$), and the property is important (p_I). For ρ_2, the syntactic category is N' ($p_{N'}$), a modifier is not used ($\frac{n}{n+m}$), the property is not important (1- p_I), and the object has the mentioned property by chance ($\frac{1}{t}$). For ρ_3, the syntactic category is N^0 (1-$p_{N'}$), the property is not important (1- p_I), and the object has the mentioned property by chance ($\frac{1}{t}$). The numerator of (19) contains the only representation that has the property as important, while the denominator contains all three representations.

The value of $\phi_{N'}$ also depends on data type. We can derive the value of $\phi_{N'}$ similarly to ϕ_I, except that μ is not set to *yes* since $p_{N'}$ is not predicated on a property being mentioned. Instead,

σ is set to <NP since $p_{N'}$ is predicated on the syntactic environment indicating the category is smaller than NP. In addition, N' (syntactic category=N') is the variable of interest.

$$\phi_{N'} = p(N'|\pi,\sigma =< NP,\mu,\omega) = \frac{p(\pi,\mu,\omega|N',\sigma =< NP) * p_{N'}}{p(\pi,\mu,\omega|\sigma =< NP)} \tag{21}$$

Unambiguous <NP data end up having ϕ_I=1, which is again intuitively satisfying since they unambiguously indicate that the category is N' when the syntactic environment is <NP. Sem-Syn ambiguous data end up having $\phi_{N'}$ as in (22):

$$\phi_{N'Sem-Syn} = \frac{\rho_1 + \rho_2}{\rho_1 + \rho_2 + \rho_3} \tag{22}$$

where ρ_1, ρ_2, and ρ_3 are the same as in (20). Equation (22) is intuitively satisfying as only ρ_1 and ρ_2 are representations with syntactic category N'.

Syn Ambiguous data end up having $\phi_{N'}$ as the following:

$$\phi_{N'Syn} = \frac{\rho_4}{\rho_4 + \rho_5} \tag{23}$$

where

$$\rho_4 = \quad p_{N'} * \frac{n}{n+m} \tag{24a}$$

$$\rho_5 = \quad 1 - p_{N'} \tag{24b}$$

The quantities in (24) intuitively correspond to representations for anaphoric one when no property is mentioned in the previous context. For ρ_4, the syntactic category is N' ($p_{N'}$) and the N' string uses only a noun ($\frac{n}{n+m}$). For ρ_5, the syntactic category is N^0 (1-$p_{N'}$), and so the string is

noun-only by definition. The numerator of equation (23) contains the representation that has the category as N', while the denominator contains both possible representations.

Table 5 shows the different model parameters updated for each data type, as well as sample updates for p_I and $p_{N'}$, showing the value of each probability after one data point is seen at the beginning of learning when $p_I = p_{N'} = 0.50$. Other parameters take the following values for the sample updates, based on estimates from P&L: $m = 1$, $n = 3$, and $t = 5$. The values for m (number of modifier strings that are N') and n (number of noun-only strings that are N') are based on empirical estimates from corpus data, while t is a low estimate of the number of properties present in the learner's environment at the time the data point is encountered. When t is low, the beneficial impact of ambiguous data points on p_I is less, since each data point is less of a suspicious coincidence. For example, if there are five properties in the learner's environment (e.g., SILLY, STRIPED, NEXT TO THE DOLLY, BOUNCY, BEHIND MOMMY'S BACK), then it is less of a suspicious coincidence that the item in question happens to be STRIPED (1/5) than if there were twenty properties (1/20). A learner using this low t value thus boosts the value of p_I less for each informative ambiguous data point. Thus, by using low t values, we are biasing our learner away from a higher p_I (and so the learner is less likely to think the mentioned property is important and thus less likely to learn the correct representation of anaphoric one).

Table 5: Values for model parameters for each data type, and sample updates for p_I and $p_{N'}$, showing the value of each probability after one data point is seen at the beginning of learning when $p_I = p_{N'} = 0.50$, $\alpha = \beta = 1$, $m = 1$, $n = 3$, and $t = 5$.

	$data_x = data_x + \phi_x$		$p_x = \frac{\alpha + data_x}{\alpha + \beta + totaldata_x}$, $\alpha = \beta = 1$	
Data type	ϕ_I	$\phi_{N'}$	p_I	$p_{N'}$
Unamb <NP	1	1	0.67	0.67
Sem-Syn Amb	$\frac{\rho_1}{\rho_1+\rho_2+\rho_3}$	$\frac{\rho_1+\rho_2}{\rho_1+\rho_2+\rho_3}$	0.47	0.56
Syn Amb	N/A	$\frac{\rho_4}{\rho_4+\rho_5}$	0.50	0.48
Unamb NP	1	N/A	0.67	0.50

For Unamb <NP data, both ϕ_I and $\phi_{N'}$'s *phi* values are 1, and so $data_x$ is increased by 1. This leads to p_I and $p_{N'}$ both being increased. This is intuitively satisfying since unambiguous <NP data by definition are informative about both p_I (the mentioned property is indeed important) and $p_{N'}$ (the syntactic category is N').

For Sem-Syn Amb data, both p_I and $p_{N'}$ are altered, based on their respective ϕ values, which are less than 1 but greater than 0. The exact ϕ value depends on current values of p_I and $p_{N'}$. After one Sem-Syn Amb data point, p_I is lowered slightly (to .47), since the coincidence of the referent having the mentioned property is not suspicious enough. This is due to t being low.[10] However, $p_{N'}$ is increased slightly (to .56) since the current probabilities of the two representations that have the syntactic category as N' (ρ_1 and ρ_2) outweigh the current probability of the representation that has the syntactic category as N^0 (ρ_3).

Syn Amb data are only informative with respect to syntactic category, so only $p_{N'}$ is updated and only $\phi_{N'}$ has a value. Here, we see the misleading nature of the Syn Amb data that P&L discovered - the value of $p_{N'}$ is lowered because the representation using syntactic category N^0 (ρ_5) currently has a higher probability than the representation using category N' (ρ_4). This is because the N' representation in ρ_4 must include the probability of choosing a noun-only string (like "bottle") from all the N' strings available in order to account for the observed data point ($\frac{n}{n+m}$), while the N^0 category by definition only includes noun-only strings.

Unamb NP data are only informative with respect to whether the mentioned property is important, so only p_I is updated and only ϕ_I has a value. Since these data are unambiguous, ϕ_I=1, which is intuitively satisfying. This leads to an increase in p_I.

5.2 Learner input sets & parameter values

Table 6 indicates the availability of different data types in the learner's input, based on a corpus analysis on the Brown-Eve corpus (Brown, 1973) from the CHILDES database (MacWhinney,

[10] With t=20, for example, p_I = 0.58 and $p_{N'}$ = 0.62 after one Sem-Syn Amb data point.

2000). We chose the Eve corpus since it included naturalistic speech directed to a child starting at the age of 18 months and continuing through 27 months, containing 17,521 child-directed speech utterances.[11]

Table 6: Data type frequencies

Data type	**Brown-Eve**
Unamb <NP	0.00%
Syn-Sem Amb	0.66%
Syn Amb	7.52%
Unamb NP	8.42%
Uninformative	83.4%

We note that we did not find any Unamb <NP data, which accords with Baker's original intuition that such data are very scarce. We note also that uninformative data includes ungrammatical uses of anaphoric one, uses of one where no potential antecedent was mentioned in the previous linguistic context (e.g., "Do you want one?" with no previous linguistic context), and uses of pronouns as NPs where the antecedent did not contain a modifier (e.g., "Mmm - a cookie. Do you want it?"). This last kind of data is viewed as uninformative because NP data points can only help indicate whether a mentioned property is important. If no property is mentioned in the antecedent, then the data point is uninformative as to whether a referent must have the mentioned property.

Following P&L, we posit that the anaphoric one learning period begins at 14 months, based on experimental data supporting infant recognition of the category Noun and the ability to distinguish it from other categories such as Adjective at this age (Booth & Waxman, 2003). If children hear approximately 1,000,000 sentences from birth until 18 months (Akhtar et al., 2004), then we can use the data frequencies in table 6 to estimate the expected distribution of anaphoric one data during the learning period that spans from 14 to 18 months. Based on our analysis, we estimate that the child hears approximately 36,500 referential pronoun data points during the learning period. Table

[11] See Appendix A for a more thorough breakdown of the corpus analysis we have conducted here. See Appendix B for a comparison of the LWF corpus analysis to our corpus analysis.

7 below shows the input sets we will use to test the different learning proposals for anaphoric one.

Table 7: Input sets for different anaphoric one proposals

Data type	Baker	R&G, P&L	P&L's EO	P&M
Unamb <NP	0	0	0	0
Sem-Syn Ambig	0	242	242	242
Syn Ambig	0	0	2743	2743
Unamb NP	0	0	0	3073
Uninformative	36500	36258	33515	30442

For the free parameters in the model, we will follow the corpus-based estimate P&L used for m and n, which is approximately equivalent to $m = 1$ and $n = 3$.[12] These parameters matter when the learner is trying to decide whether the syntactic category should be N' or N^0, given that it is smaller than NP (i.e., $p_{N'}$). The smaller m is compared to n, the less that Syn ambiguous data cause a Bayesian learner to (incorrectly) favor the N^0 category over the N' category. P&L discuss why Syn ambiguous data have this effect in more detail, but for our purposes it suffices that if a learner using Syn Amb data cannot succeed with these values of m and n, the learner will not fare any better with other estimates that make m larger and/or n smaller.

We will also follow an estimate P&L used for t: $t = 5$. This is a lower estimate of t, which minimizes the benefit to any learners who heed suspicious coincidences (in particular, the suspicious coincidence of the referent just happening to have the mentioned property) for the reason discussed in 5.1. Heeding suspicious coincidences specifically aids the learner in deciding that the mentioned property is important (i.e., p_I is near 1). By making t low, we are biasing the learning environment against learners deciding the mentioned property is important. Thus, any learners who end up with a probability p_I near 1 with this low t value should end up with a p_I near 1 with higher t values.

[12]The actual numbers P&L found from their corpus analysis of N' strings were 119 noun+modifier N' strings to 346 noun-only N' strings, which is a ratio of 1 to 2.9.

5.3 Measures of success

One way to assess acquisition success is to measure p_I and $p_{N'}$ at the end of the learning period, since we would want these values to be near 1 for an adult representation.[13] In addition, we can also assess how likely a learner would be to reproduce the observed infant behavior from the LWF experiment. In particular, when presented with a scenario with utterances like "Look - a red bottle! Do you see another one?", how often will the learner look to the bottle with the mentioned property (RED)?

We can calculate the probability (p_{beh}) of the learner looking at the referent that has the mentioned property when given a choice between two referents. As before, π refers to what pronoun was mentioned, σ refers to what the syntactic environment is, μ refers to whether the previous context mentioned a property, and ω refers to whether the object has the mentioned property. Thus, the probability of reproducing the infant behavior in the LWF experiment is the probability of looking to the object that has the mentioned property ($\omega = hasproperty$), given that the observed pronoun is one ($\pi = \mathsf{one}$), the syntactic environment indicates the pronoun is smaller than NP ($\sigma =< NP$), and a property has been mentioned ($\mu = yes$).

$$p_{beh} = p(\omega = hasproperty | \pi = \mathsf{one}, \sigma =< NP, \mu = yes) \tag{25}$$

Using the information dependencies in figure 3, this works out to

$$p_{beh} = \frac{\rho_1 + \rho_2 + \rho_3}{\rho_1 + 2 * \rho_2 + 2 * \rho_3} \tag{26}$$

where ρ_1, ρ_2, and ρ_3 are defined as in (20), $m = 1$, $n = 3$, and $t = 2$ (since there are only two objects present in the experimental setup). As before, these quantities intuitively correspond to the differ-

[13] We note that this is the default adult representation, though there may be other pragmatic factors that impact the final adult representation. This is discussed further in section 6.2.

ent outcomes. For the correct representation where the property is important and the category is N' (ρ_1), the learner must look to the object with the property. For any of the incorrect representations (ρ_2 and ρ_3) where the antecedent string is effectively just the noun (e.g., "bottle"), the learner has a 1 in 2 chance of looking at the correct object by accident. The numerator represents all the outcomes where the learner looks to the correct object, while the denominator also includes the two additional outcomes where the learner looks to the incorrect object (ρ_2 and ρ_3 with incorrect behavior).

In addition, we can also assess the assumption LWF made about their experiment - in particular, if infants look at the object adults look at when adults have the correct representation of anaphoric one, it means that the children also have the correct representation. While this does not seem like an unreasonable assumption, it is worth asking if this is true. It is possible, for example, that children have an incorrect representation, but look at the correct object by chance (represented in the numerator of (26) as ρ_2 and ρ_3). Given this, there are two related questions that we can ask.

First, is it possible to get the correct behavior in the LWF experiment without having the correct representation for one *in general* (as represented by p_I and $p_{N'}$)? To answer this question, we can simply look at p_{beh} compared to p_I and $p_{N'}$. If p_{beh} is high when either p_I or $p_{N'}$ is low, this suggests that the correct behavior may not necessarily implicate the correct representation in general.

Second, is it possible to get the correct behavior in the LWF experiment without having the correct representation for one at the time the behavior is being generated? To answer this question, we can calculate calculate the probability ($p_{rep|beh}$) that the learner has the correct representation, given that the learner has produced the correct behavior (e.g., looking at the RED BOTTLE) in the experiment. This is, in effect, the contextually-constrained representation the learner is using, where the context is defined as the experimental setup.

$$p_{rep|beh} = p(N', I|\pi = \textsf{one}, \sigma =< NP, \mu = yes, \omega = hasproperty) \tag{27}$$

As before, π refers to what pronoun was mentioned, σ refers to what the syntactic environment is, μ refers to whether the previous context mentioned a property, and ω refers to whether the object has the mentioned property. In addition, *N'* refers to the syntactic category being N' (syntactic category = N', given that it is smaller than NP) and *I* refers to the property being important (property important = yes, given that a property has been mentioned). Thus, the probability of the learner having the correct representation, given that the learner has produced the correct behavior, is equivalent to the probability that the learner believes the syntactic category is N' (*N'*) and the mentioned property is important (*I*), given that the pronoun used was one ($\pi = \textsf{one}$), the syntactic environment indicates the category is smaller than NP ($\sigma =< NP$), a property was mentioned ($\mu = yes$), and the selected object has that property ($\omega = hasproperty$).

Using the information dependencies in figure 3, this works out to

$$p_{rep|beh} = \frac{\rho_1}{\rho_1 + \rho_2 + \rho_3} \tag{28}$$

where ρ_1, ρ_2, and ρ_3 are calculated as in (20), but with t = 2 (again, because there are only two objects to choose from in the LWF experimental setup). More specifically, given that the correct object has been looked at (whether on purpose (ρ_1) or by accident (ρ_2 and ρ_3)), we calculate the probability that the look is due to the correct representation (ρ_1).[14]

6 Results

Table 8 shows the results of the learning simulations over the different input sets, with averages over 1000 runs reported and standard deviations in parentheses.

[14]Note that this is the same equation as (19) (the only difference is the value of t). This has some intuitive appeal since ρ1 in (20) corresponds to the correct representation which has the mentioned property as important, while the

Table 8: Probabilities after learning

Prob	Baker	R&G, P&L	P&L's EO	P&M
$p_{N'}$	0.50 (<0.01)	0.97 (<0.01)	0.17 (0.02)	0.37 (0.04)
p_I	0.50 (<0.01)	0.95 (<0.01)	0.02 (0.01)	>0.99 (<0.01)
p_{beh}	0.56 (<0.01)	0.93 (<0.01)	0.50 (<0.01)	>0.99 (<0.01)
$p_{rep\|beh}$	0.22 (<0.01)	0.92 (<0.01)	<0.01 (<0.01)	>0.99 (<0.01)

Focusing first on $p_{N'}$ and p_I, we can see that our online learning model is producing results similar to what previous studies found when using the data sets proposed by those previous studies. Learning from unambiguous data alone does not work, as Baker supposed ($p_{N'} = 0.50$, $p_I = 0.50$). Including Sem-Syn ambiguous data will lead to the correct representation, as R&G and P&L found ($p_{N'} = 0.97$, $p_I = 0.95$). Additionally including Syn ambiguous data, as P&L's EO learner did, leads to the incorrect representation ($p_{N'} = 0.17$, $p_I = 0.02$).

The new result we have found is that expanding to unambiguous NP data (P&M) does not lead to the correct representation, since the learner's belief that the syntactic category is N' is low in general ($p_{N'} = 0.37$). However, perhaps surprisingly, this turns out not to matter for producing the correct behavior in the LWF experiment ($p_{beh} > 0.99$). That is, the learner could have the incorrect representation *in general* but still produce the correct behavior in that experimental setup with very high probability. How could this be? It turns out this is due to the high value of p_I, i.e., the learner's strong belief that a mentioned property is important. If the learner believes a mentioned property is important, then the object must have that property (e.g., be a RED BOTTLE when "red" was mentioned in the potential antecedent). So, the learner looks to the referent that has the property and this produces the correct behavior. Thus, it seems that LWF's assumption does not hold - producing adult-like behavior does not necessarily indicate that the learner has the correct representation in general.

However, a relaxed version of the LWF assumption does appear to hold. In particular, when

other two representations do not.

the child produces the correct behavior, the probability that the child has the correct representation *at the time the interpretation is being made* is very high ($p_{rep|beh} > 0.99$). This is again due to the learner's strong belief that the mentioned property is important. If the property is important, then the object must have that property (e.g., be a RED BOTTLE), which means the antecedent of one must include the mentioned modifier (e.g.,"red bottle" instead of just "bottle"). Since only category N' can contain modifiers, then one must be category N' *in this context*.

Thus, even though the learner has a incorrect representation in general, in the context where a modifier is present, the learner will end up with the correct interpretation and the correct representation. LWF were not wrong to assume correct behavior was due to a correct representation - it's simply that the correct representation may not apply generally. In particular, the P&M learner will have the incorrect representation when given Syn ambiguous data like "Look, a bottle! Do you see another one?" Since no property is mentioned, the high p_I value cannot help. Instead, the learner falls back on the $p_{N'}$ value alone, which is low ($p_{N'} = 0.37$), and so the learner will end up with one as N^0 for that data point.[15]

We note that this result is due to the input set the P&M learner is using - the learners using restricted input sets behave exactly as LWF would expect. When they have the correct representation in general (R&G, P&L), they produce the correct behavior and have the correct representation when producing that behavior. When they have the incorrect representation in general (Baker, P&L's EO), they produce chance behavior and likely have the incorrect representation if they happen to produce the correct behavior.

[15]Note however that the P&M learner would have the correct *behavior* when no property was mentioned, even with the incorrect representation. This is because the antecedent is clear (e.g., "bottle") and so the incorrect syntactic representation ([$_{N^0}$ bottle]) has no effect on identifying the correct referent.

6.1 Discussion

6.1.1 General discussion of results

Through our modeling study, we have learned several things about the acquisition of anaphoric one. First, indirect positive evidence can indeed aid a learner. Children may be able to learn the correct interpretation for one in certain situations (like the LWF experiment) by broadening the set of data they consider relevant, such as the Unambiguous NP data the P&M learner considered here. Second, we have discovered that the link between observed behavior, interpretation, and representation may not be so clear cut. Just because children demonstrate they have the correct interpretation some of the time (by displaying correct behavior) does not necessarily mean they have the correct representation all of the time. We have provided an example learner that would have the correct interpretation in the the context of the LWF experiment, but would not have the correct representation for other utterances, like those in Syn ambiguous data.

This discovery then tells us something about the acquisition process for anaphoric one. In particular, it suggests that while children must eventually learn that one is N', they do not need to do so by 18 months. Infants that learn as the P&M learners do here could produce the correct behavior even when they believe one is N^0 in general. This means that children just need to learn that one is N' sometime before they become adults, so that they find "side of the building" grammatical while finding "one of the road" ungrammatical.[16]

Since children do not need to have this syntactic category knowledge by 18 months, this may allow them time to develop the knowledge they need to follow the strategy proposed by F&al. In particular, recall that F&al's learner relied on subtle conceptual distinctions to leverage the syntactic distribution of one and learn that one is N': complements (indicating category N^0) conceptually evoke the head noun while modifiers (indicating category N') do not. While it is difficult to imagine 18-month-olds capable of making this subtle distinction, it is easier to imagine older children

[16] When this distinction is acquired by children is left to future experimental work.

doing so.

This would then lead to a more complex acquisition trajectory. Initially, children could use a broader input set (like the P&M learner) and learn the correct interpretation for one in most contexts, even if they believe one is N^0 by default. Later, children could be sophisticated enough to leverage the information in the syntactic distribution and identify one as definitively N'. The overall acquisition trajectory would look something like the one in Table 9.

Before 18 months, a learner using indirect positive evidence like the P&M learner would need to recognize that one is similar to other referential pronouns. This is domain-specific knowledge (since it refers to referential elements of the language), but it can likely be derived from the input by leveraging the distribution of referential elements. Though one does not have an identical distribution to other referential elements like *it* (e.g., "another one", but *"another it"), the distribution overlaps significantly (e.g., "I see one", "I see it", etc.). A learner can likely use innate, domain-general statistical learning abilities to leverage this distribution and learn this domain-specific knowledge.

Before 18 months, a P&M learner would then track how often a property mentioned in the potential antecedent is important for a referent to have (e.g., when hearing, "Look, a red bottle! Oh look, another one." or "Look, a red bottle! I want it.", how often is the bottle RED?). A Bayesian learner would track these suspicious coincidences using innate, domain-general statistical learning abilities.

At 18 months, a P&M learner can then produce the observed behavior in the LWF experimental context because the learner believes a mentioned property is important (even if the learner believes one is more likely to be N^0 in general). After 18 months, the learner could follow the F&al strategy, and leverage the syntactic distribution of one. Specifically, the learner keys into the subtle semantic distinction between complements and modifiers and knows the syntactic category implications for complements and modifiers. Leveraging the syntactic distribution likely involves innate, domain-general, statistical learning abilities, but recognizing the syntactic implications of complements

and modifiers may involve innate, domain-specific knowledge about language. While there may be a way to derive this domain-specific knowledge, we could not think of any obvious ways to do so (though of course it may be possible). However, to the extent that this innate domain-specific knowledge is required, acquisition of anaphoric one would then seem to require what is traditionally described as Universal Grammar.

Table 9: Learning trajectory for anaphoric one and learning biases required.

When	Bias	Innate	Derived	Dom-Spec	Dom-Gen	What To Learn	How To Learn
Before 18 months	one is like other referential elements	*	*	*	*	*	*
Before 18 months	recognize suspicious coincidence of antecedent property	*			*		*
After 18 months	know what syntactic complements vs. modifiers imply	?		*		*	
After 18 months	leverage complement vs. modifier distribution		*		*		*

With respect to the process of acquisition, we have shown that there may be a two-stage acquisition trajectory for anaphoric one. The first stage involves learning the correct representation in certain contexts, while the second stage involves learning the correct representation for all contexts. Though a variety of different learning biases are required, only the second stage may need a bias that is innate and domain-specific.

6.1.2 Broader implications

The results here also offer answers to some of the larger questions we're more generally interested in with respect to language acquisition. First, when induction problems exist, what does it take to solve them? We have provided a case study suggesting that broader data sets may be additional sources of information, providing indirect positive evidence. Thus, relaxing the direct evidence

assumption can be useful for understanding how children solve acquisition problems.

Second, when there is an induction problem, what learning biases are needed to solve it and are any of them part of Universal Grammar? Here, we have looked at a number of different learning biases a learner would need to match the behavior observed in children and (eventually) adults for anaphoric one. Only one bias seemed to be a candidate for an innate, domain-specific learning bias, and so something that would be part of Universal Grammar.

Third, can we learn anything about the acquisition trajectory by exploring the learning biases needed to solve induction problems? Through this case study, we have provided an example that does this. We identified learning biases that a learner might use to produce the behavior observed in 18-month-olds, and implemented them in a learning model. This allowed us to identify the knowledge state a learner using those biases would have when producing that observed behavior. Because this knowledge state did not match the adult knowledge state, this suggested a two-stage learning process.

6.2 Future directions

There are a number of ways to extend the research here, looking at the information sources available, the overall problem to be solved, and alternate learning strategies a learner might use.

6.2.1 Additional Sources of Information

Our learning model here was a Bayesian learning model that was able to track suspicious coincidences. Specifically, our learning model looked at the referent and the properties that referent had, comparing them to the property that was mentioned. The magnitude of the suspicious coincidence was determined only by how many other properties there were in the learner's consideration (i.e., the impact was inversely proportional to the chance that the referent had the mentioned property out of all the properties it could have had, implemented with parameter t).

However, there may be more nuanced ways to interpret how suspicious a coincidence is.[17] For example, consider Sem-Syn ambiguous data (e.g., "Look - a red bottle! Oh look - another one!", when the referent is a red bottle). These data may present a stronger suspicious coincidence if another object is present that does not have the mentioned property (e.g., a purple bottle), but the speaker specifically indicates (say, by gesture or gaze) that the object with the mentioned property is intended (e.g., a red bottle). This could be an additional cue that the mentioned property is relevant ("red"), because there was another object present that didn't have that property and the speaker specifically didn't pick that other object. Given this, data points like this might have update values closer to that of unambiguous data (which has $\phi_I = \phi_{N'} = 1$), since it is more likely that the mentioned property is important (p_I) and so more likely that the category is N' ($p_{N'}$). Without a corpus analysis that includes this kind of situational information, it is unclear how frequent these "more influential" Sem-Syn ambiguous data are. However, see Appendix C for one way to estimate the impact these kind of data could have on learning anaphoric one.

Another source of information involves more sophisticated contextual cues. Some examples are shown below in (29):

(29a) "I hate that red bottle - do you have another one?"

(29b) "I want this *red* bottle, and you want *that* one." (*italics* indicate emphasis)

Most adults would interpret the referent of one in both cases as a BOTTLE that is not red. For (29a), this is perhaps based on the verb "hate", and the inference that someone would not ask for another of something they hate . For (29b), this is perhaps based on the contrastive focus that occurs between "red" and "that". In both cases, this involves an inference that draws from information beyond the default syntactic and semantic representation. In (29a), this is an inference about when a speaker would use "hate" in this way; in (29b), this is an inference about when speakers use contrastive focus. The default interpretation of one seems to include the modifier (see 30). In (30a), it seems the speaker is requesting another red bottle. In (30b), while there is contrastive

[17] Thanks to the UChicago audiences for pointing the ideas in this section out.

focus with "that", it doesn't interfere with the interpretation of one's antecedent as "red bottle".

(30a) "I love that red bottle - do you have another one?"

(30b) "I want *this* red bottle, and you want *that* one." (*italics* indicate emphasis)

We note that we did not find any occurrences of data like (29) in our corpus analysis, which suggests that young children probably do not encounter these data very often. In addition, it is unclear how sensitive very young children (younger than 18 months, for example) would be to this additional contextual information, and how well they would be able to make the pragmatic inferences that adults would make. Incorporating this additional contextual information when forming an interpretation is clearly something children must eventually learn to do since adults do it, but we speculate that the initial target state for learning is the default interpretation where the mentioned property is important. It would be useful to assess when children have the adult interpretations for non-default anaphoric one examples like those in (29), as this would allow us to further fine-tune the acquisition trajectory.

6.2.2 Alternate forms of the learning problem

In the present study, we have examined learning the representation of anaphoric one assuming one standard syntactic structure. In particular, we assumed the following: (i) noun phrases are category NP, (ii) modifiers are sister to N', and (iii) complements are sister to N^0. This would give the structure for the noun phrase "a delicious bottle of wine" represented in the left side of figure 4, and shown in bracket notation in (31). However, an alternate representation of noun phrases is available[18], shown in (32) and the right side of figure 4. It assumes the following: (i) noun phrases are category DP (Determiner Phrase), (ii) modifiers are sisters to N' and children of NP, and (iii) complements are sisters of N'.

[18]Thanks to Greg Kobele for noting this.

(31) [$_{NP}$ a [$_{N'}$ delicious [$_{N'}$ [$_{N^0}$ bottle] [$_{PP}$ of wine]]]]

(32) [$_{DP}$ a [$_{NP}$ delicious [$_{N'}$ [$_{N'}$ [$_{N^0}$ bottle]] [$_{PP}$ of wine]]]]

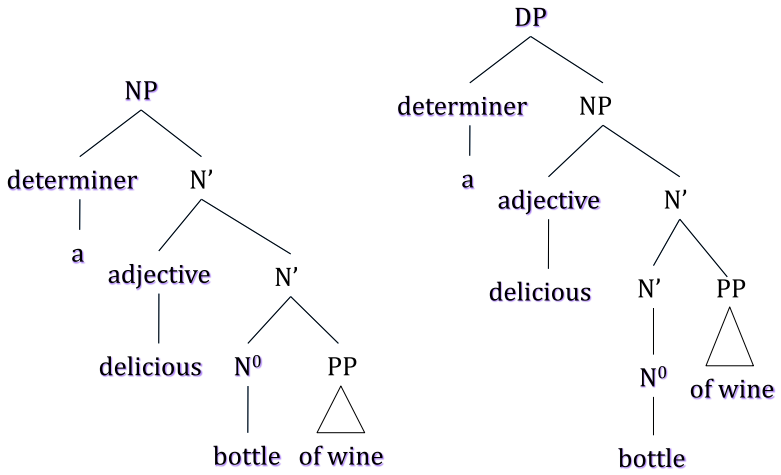

Figure 4: Phrase structure trees corresponding to the bracket notation in examples (31) and (32) for "a delicious bottle of wine".

Practically speaking, this means that the learner must learn that the antecedent of anaphoric one can be category NP (e.g., "delicious bottle of wine") or category N' (e.g., "bottle of wine") but never category N^0 (e.g., "bottle" in (33)). This means there are three syntactic categories smaller than an entire noun phrase (DP), and a child must learn that only two of them are valid antecedents for one. Moreover, in the LWF experiment, a child should have the preference that one's antecedent is category NP, so that it can include the modifier (i.e., "red bottle" is an NP in this representation).

(33) "I have a delicious bottle of wine...

(a) ...and you have one, too." [one = "delicious bottle of wine", category NP]

(b) ...and you have a flavorful one, too." [one = "bottle of wine", category N']

(c) ...*and you have a flavorful one of beer. [one $\neq$ "bottle", category N^0]

While we have not implemented a model that uses this syntactic representation, we can speculate on the results we might find. First, when faced with Syn ambiguous data (e.g., "Look - a

bottle! Oh, look - another one!"), there is a three-way ambiguity (NP vs. N' vs. N^0) instead of a two-way ambiguity. Since a Bayesian learner will prefer the smallest syntactic category consistent with the data point[19], the learner will still prefer N^0 as our learner did here. Thus, Syn ambiguous data remain misleading about the syntactic category of one (i.e., category = N^0).

Second, both Sem-Syn ambiguous data and Unamb NP data would lead a learner to assume the category is NP when a modifier is present (e.g., "red bottle"). This is because both these data types increase the probability that the mentioned property is important for one's referent to have (p_I). In this syntactic representation, only category NP can include modifiers. Therefore, the learner will likely perform well in the LWF experiment, as long as p_I is high.

Because no data favor N', we would expect that the learner disprefers one as N' at the end of learning. Instead, the learner assumes one is NP (e.g., antecedent = "red bottle") in contexts like the LWF experiment that have a property mentioned and assumes one is N^0 in general when no property is mentioned. This is qualitatively the same result that we have found here, and would still predict a two-stage acquisition trajectory. Learning in the second stage might again be able to make use of the complement vs. modifier distinction, though not quite as directly. In particular, in this representation, both modifiers and complements are sisters to N', as shown in the right side of Figure 4. However, complements are sisters to an N' whose only child is N^0. The learner would thus need to connect the subtle semantic distinction between complements and modifiers to the syntactic structure shown on the right side of Figure 4, which involves the syntactic knowledge that complements are sisters to particular kinds of N'. For the same reasons discussed in section 6.1.1, this knowledge may be a good candidate for Universal Grammar. So, a learner using this syntactic representation would likely still need to rely on innate, domain-specific knowledge, as we found with the learner implemented in the present study.

[19]This is due to the Size Principle (Tenenbaum & Griffiths, 2001). In particular, the set of strings covered by category N^0 is smaller than the set of strings covered by category N', which is smaller than the set of strings covered by category NP. A noun-only string like "bottle" is consistent with all three categories, and so the category covering the smallest set of strings is favored.

6.2.3 Alternate learning strategies

We have explored learners that use a particular probabilistic learning strategy (Bayesian learning) that implicitly favors the smallest set compatible with the observable data (Tenenbaum & Griffiths, 2001). However, there are alternate strategies learners might use. For example, a learner might have an explicit bias to prefer the largest set compatible with the observable data.[20] For instance, given a noun-only string like "bottle" that is compatible with category N' and category N^0, this learner would prefer to choose the larger syntactic category (N').

We speculate that this kind of bias could lead to the correct representation at 18 months. To briefly sketch how this would work, consider that the misleading Syn ambiguous data cause the current learner to prefer category N^0 over category N'. However, a learner who prefers the larger structure will not be led astray the same way - that learner would prefer category N' in this situation, which is the correct representation. In fact, a learner with that bias would not even need to use indirect positive evidence as the P&M learner does here - using only the Unambiguous <NP, Sem-Syn ambiguous, and Syn ambiguous data should lead this learner to the correct representation.

So how could a learner come to have this kind of learning strategy? It must be explicit because it does not implicitly fall out from the mechanics of Bayesian inference. For the dimension of what to learn vs. how to learn, it clearly is a bias about how to learn (choose the largest set/structure compatible). For the dimension of domain-specific vs. domain-general, it could be domain-general if it applies to other data besides language data, but domain-specific if it only applies to learning language knowledge. For the dimension of innate vs. derived, it could certainly be an innate preference (though it would go against the implicit preference to choose the smallest compatible set that comes from Bayesian inference). On the other hand, it may be possible to derive this preference if other data demonstrate that choosing the larger set/structure is the correct answer. Something in the linguistic domain that does this is verb phrase ellipsis, such as "I promised to help him and you *did*, too." Most adults interpret this as "I promised to help him and you *promised*

[20] Thanks to Ming Xiang for suggesting this.

to help him, too", rather than "I promised to help him and you *helped him*, too." This suggests that *did* is replacing the larger verb phrase, rather than the smaller one. However, it is unclear how frequently verb phrase ellipsis occurs in child directed speech - if it occurs less frequently than anaphoric one, it may not be a good way for children to derive that useful learning strategy in time to learn anaphoric one.

7 Conclusion

We have demonstrated that indirect positive evidence can be leveraged effectively by an online probabilistic learner in order to produce behavior consistent with young children's anaphoric one behavior, even if the learner does not achieve the adult representation. This suggested that the acquisition process may require more than one stage. Though the first stage would not require innate domain-specific knowledge, a subsequent acquisition stage might.

Indirect evidence does not necessarily negate the need for learning biases - it may, however, alter the nature of the necessary learning biases. Considering indirect evidence and its impact on acquisition can help define concrete proposals of the contents of Universal Grammar. We believe this general approach of looking at broader input sets for learning linguistic phenomena may be fruitful for identifying what is and is not necessarily part of Universal Grammar. Knowing the impact of the necessary learning biases on acquisition may also inform us about the acquisition trajectory, and provide guidance for additional experimental investigation.

8 Acknowledgements

We are very grateful to Vance Chung and Erika Webb for their assistance with the corpus analysis. In addition, we have benefited from some very enlightening discussion with Max Bane, Morgan Sonderegger, Greg Kobele, Ming Xiang, the Computation of Language laboratory at UC Irvine, the

2010 Computational Models of Language Learning seminar at UC Irvine, and the audiences at the UChicago 2011 workshops on Language, Cognition, and Computation and Language, Variation, and Change. In addition, this research was supported by NSF grant BCS-0843896 to LP.

References

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

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

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

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

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

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

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

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

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

Chomsky, N. (1988). *Language and problems of knowledge: The managua lectures*. Cambridge, MA: MIT Pressl.

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

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

Dresher, E. (2003). Meno's paradox and the acquisition of grammar. In S. Ploch (Ed.), *Living on*

the edge: 28 papers in honour of jonathan kaye (studies in generative grammar 62 (pp. 7–27). Berlin: Mouton de Gruyter.

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

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

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*, 287–300.

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

Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein (Ed.), *Explanation in linguistics: The logical problem of language acquisitions* (pp. 9–31). London: Longman.

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

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

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

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65–B73.

Lightfoot, D. (1982). Review of geoffrey sampson, making sense. *Journal of Linguistics*, *18*, 426–431.

Lightfoot, D. (1989). The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences*, *12*, 321–334.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

McMurray, B., & Hollich, G. (2009). Core computational principals of language acquisition: can statistical learning do the job? introduction to special section. *Developmental Science*, *12*(3),

365–368.

Pearl, L. (2007). *Necessary bias in natural language learning*. Unpublished doctoral dissertation, University of Maryland, College Park, College Park, MD.

Pearl, L., & Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. *Language Learning and Development*, *5*(4), 235–265.

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.

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

Reali, F., & Christiansen, M. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028.

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

Scholz, B., & Pullum, G. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review*, *19*, 185–223.

Soderstrom, M., Conwell, E., Feldman, N., & Morgan, J. (2009). The learner as statistician: three principles of computational success in language acquisition. *Developmental Science*, *12*(3), 409–411.

Spelke, E. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626–636.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103–114.

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

Tomasello, M. (2004). Syntax or semantics? response to lidz et al. *Cognition*, *93*, 139–140.

A Frequency of different pronouns in the input

Since the P&M learner uses all informative referential pronoun data, we included all available referential pronouns in our corpus analysis instead of focusing only on anaphoric one. Table 10 shows the breakdown of the pronouns observed in the Eve corpus (Brown, 1973). We note that not all these pronouns belonged to informative data points (where informative is defined as in section 5.2).

Table 10: Pronoun frequencies in Eve corpus

Pronoun	**Frequency**	**%**
it	1538	53.7%
he	321	11.2%
one<NP	302	10.5%
them	182	6.4%
she	165	5.8%
they	142	5.0%
her	80	2.8%
him	76	2.7%
one=NP	52	1.8%
itself	3	0.1%
himself	1	<0.1%
total	2862	100%

From this distribution, we can see that it is the most frequent pronoun, which makes up the bulk of the Unamb NP examples in the P&M input set.

B Corpus analysis comparison

LWF conducted a corpus analysis on the Suppes (Suppes, 1974) and Brown-Adam (Brown, 1973) corpora from CHILDES (MacWhinney, 2000), which contained approximately 54,800 child-directed utterances total, but they did not include the Unamb NP data points that the P&M learner uses. Given this, we also conducted an analysis on the Brown-Eve corpus (Brown, 1973), which in-

cluded all four data types. Table 11 compares the availability of different data types in the learner's input, based on the two corpus analyses. Note that because we included Unamb NP data points, LWF's uninformative data points proportion was much lower than ours - specifically, only ungrammatical <NP data points where uninformative for their analysis while ungrammatical data points, data points that didn't have a mentioned antecedent (e.g. "Do you want one?" with no previous linguistic context), and NP data points where the antecedent did not contain a modifier (e.g., "Mmm - a cookie. Do you want it?") were uninformative for our analysis.

Table 11: Data type frequencies

Data type	**LWF: Suppes & Brown-Adam**	**P&M: Brown-Eve**
Unamb <NP	0.25%	0.00%
Syn-Sem Amb	4.56%	0.66%
Syn Amb	94.72%	7.52%
Unamb NP	N/A	8.42%
Uninformative	0.47%	83.4%

Comparing the two corpus analyses, one striking observation is that we were unable to find any Unamb <NP data in our analysis (P&M). This is perhaps not so surprising, given that such data require a specific conjunction of utterance and situation (and this lack of Unamb <NP data correlates with Baker's original intuition that these data are very rare). In the original LWF analysis, only 0.25% of the data were of this type.

If we look at the other data types both analyses looked at, i.e., the Sem-Syn Amb and Syn Amb data, we find that the Syn Amb data points outnumber the Sem-Syn Amb data points in both corpus analyses. The main difference is that LWF found a higher ratio (about 21 Syn Amb to 1 Sem-Syn Amb) than we did (about 11 Syn Amb to 1 Sem-Syn Amb).

For the Unamb NP data in our analysis, we find that such data are fairly similar in quantity to the Syn Amb data in our analysis (about 11 Unamb NP data points for every 10 Syn Amb data

points).

C More influential data

A certain subset of Sem-Syn ambiguous data may be more influential than how we've implemented them here. Recall that Sem-Syn ambiguous data involve utterances like "Look - a red bottle! Oh, look - another one!" when a red bottle is present. If another non-red bottle is also present, but the speaker indicates the red bottle (say, by gesture or gaze), this seems like an additional source of information that the property is important - namely, given the choice between a referent with the property and a referent without the property, the speaker chose the referent with the property. This additional information should increase the learner's belief that the property is important, above and beyond the increase that comes just from the suspicious coincidence of picking a a referent that has the property.

Without a corpus analysis (presumably including video files that show the child's learning environment when referential data examples are uttered), it is unclear how frequently data like these appear. However, one way to explore the effect of these kind of data would be to treat some proportion of the Sem-Syn ambiguous data as if they were as influential as Unambiguous <NP data. Treating these special Sem-Syn ambiguous data as Unambiguous <NP data allows them to have the maximal effect they could have - in reality, they would likely not be as influential as Unambiguous <NP data. Table 12 shows the effect of treating *all* (100% of) Sem-Syn ambiguous data as if they were as influential as Unamb <NP data - this is the maximal amount of Sem-Syn ambiguous data that could have this additional influence. In reality, it is more likely that only a subset of the Sem-Syn ambiguous data are of this kind. Thus, we provide an estimate of the best learning performance scenario. Results are the average of 1000 simulations per learner, with standard deviations shown in parentheses. Note that results for the Baker learner remain the same as in Table 8 because that learner does not heed Sem-Syn ambiguous data and so cannot treat them

as if they were Unambiguous <NP data.

Table 12: Probabilities after learning, assuming all Sem-Syn ambiguous data are as effective as Unambiguous <NP data.

Prob	**Baker**	**R&G, P&L**	**P&L's EO**	**P&M**
$p_{N'}$	0.50 (<0.01)	>0.99 (<0.01)	0.38 (0.05)	0.38 (0.05)
p_I	0.50 (<0.01)	>0.99 (<0.01)	>0.99 (<0.01)	1.00 (<0.01)
p_{beh}	0.56 (<0.01)	>0.99 (<0.01)	0.98 (<0.01)	>0.99 (<0.01)
$p_{rep\|beh}$	0.22 (<0.01)	>0.99 (<0.01)	0.98 (<0.01)	>0.99 (<0.01)

We can observe that the results do not change qualitatively for three of the learners: the Baker learner still fails, the R&G (equivalent to the filtered P&L learner) still succeeds, and the P&M learner succeeds in the LWF experimental context ($p_{beh} = p_{rep|beh} > 0.99$) but has the incorrect representation in general ($p_{N'} = 0.38$). The main change we see is that P&L's EO learner now appears to have the same performance as the P&M learner, where before P&L's EO learner failed. In particular, if we look at Table 8 for the P&M results with no highly influential Sem-Syn ambiguous data, we see they are nearly identical to the results from P&L's EO learner here. This tells us that having just a few "unambiguous" data points (here, P&L's EO learner's influential Sem-Syn ambiguous data) has the equivalent effect of learning from Unambiguous NP data (which is what the P&M learner does).

Of course, this is the best possible learning scenario; in reality, less of the Sem-Syn ambiguous data will be highly influential and the subset that is more influential will likely not be as influential as true Unambiguous <NP data. However, this tells us that even in that best case scenario, we would still expect a two-stage acquisition trajectory: Learners who do not implement a filter to ignore Syn ambiguous data (P&L's EO, P&M) do not learn the correct representation by 18 months. Being sensitive to this additional influence of some Sem-Syn ambiguous data does not negate the impact of the Syn ambiguous data.