Induction problems, indirect positive evidence, and Universal Grammar: Anaphoric one revisited

Lisa S. Pearl and Benjamin Mis

Department of Cognitive Sciences

3151 Social Science Plaza

University of California, Irvine

Irvine, CA 92697

lpearl@uci.edu, bmis@uci.edu

May 18, 2012

Abstract

One motivation for Universal Grammar (UG) comes from the existence of induction problems in language acquisition, and their solutions. Previous induction problem characterizations have typically assumed that only direct evidence is relevant; however, *indirect positive evidence* may also be useful. We examine the case study of English anaphoric one and investigate whether a probabilistic learner using indirect positive evidence can solve this induction problem. We find that this learner, given realistic input, can reproduce child anaphoric one behavior – surprisingly, even when a non-adult representation underlies this behavior. This suggests that the previous characterization of the one induction problem may need to be updated, as the link between observable behavior and underlying knowledge is not straightforward. We also discuss the nature of the learning biases leading to this result, and how this impacts the larger debate about the motivation for UG and its contents.

1 Universal Grammar: Making an argument from acquisition

One explicit motivation for Universal Grammar (**UG**) comes from an *argument from acquisition*: UG allows children to acquire language knowledge as effectively and rapidly as they do (Chomsky, 1980a; Crain, 1991; Hornstein & Lightfoot, 1981; Legate & Yang, 2002; Lightfoot, 1982b). In particular, UG is meant to be knowledge that is part of our biological endowment (*innate*) and is only used for learning language (*domain-specific*). This knowledge allows children to solve induction problems, where the available data appear to be compatible with multiple hypotheses about the generalizations for the language.[1]

This motivation for UG thus comes directly from the existence of induction problems, and the solutions to those problems. But what exactly is *in* UG that allows children to solve the induction problems they encounter? Traditionally, proposals for the contents of UG have come from characterizing a specific induction problem pertaining to a particular linguistic phenomenon (e.g., structure-dependent rules to relate the declarative and interrogative forms of utterances (Chomsky, 1980a), the structure of English anaphoric one in certain utterances (Baker, 1978), and constraints on long-distance dependencies (Chomsky, 1973)). A specific characterization of an induction problem not only makes it possible to precisely describe a potential solution, but also to explicitly test that solution and compare it to other potential solutions.

This is where computational modeling studies have recently attempted to make progress on the debate about UG's existence and its contents (Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Pearl & Lidz, 2009; Pearl & Sprouse, in press; Perfors, Tenenbaum, & Regier, 2011; Regier & Gahl, 2004). In each case where a specific induction problem has been characterized and used to propose a learning strategy involving at least one UG learning bias, researchers have examined whether that strategy is the only solution. When other learning strategies are also capable of

[1]The induction problem in language acquisition is often referred to as 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).

solving the induction problem of interest, the nature of the learning biases that comprise those learning strategies can then be discussed. To the extent that they are UG learning biases, this provides motivation for UG and a proposal for its contents, even if those contents are not the ones originally proposed. In contrast, if the necessary learning biases are unlikely to be in UG – perhaps because they are unlikely to be both innate and domain-specific – this takes away the support for UG that comes from that characterization of the induction problem.

1.1 Characterizing induction problems

Since the characterization of an induction problem is crucial for providing support to UG and making concrete proposals about its contents, how are induction problems characterized? We believe an induction problem involves at least the following parts: the initial state, the data intake, the learning period, and the target state.

The *initial state* includes both the initial knowledge state and the existing learning capabilities of the learner at that time. The initial knowledge can be defined by specifying what children already know by the time they are trying to learn the specific linguistic knowledge in question. This can be stipulated – for example, we might assume that children already know there are different grammatical categories before they learn the syntactic representation of some item in the language. However, this may also be assessed by experimental methods that can tell us what knowledge children seem to have at a particular point in development – for example, do they behave as if they have grammatical categories? Similarly, experimental methods can also be used to assess what learning capabilities children have, such as whether they can track distributional information in the input.

The *data intake* for an induction problem refers to the available input children use to learn from (Fodor, 1998). This is defined by the assumptions and biases the learner has in the initial state. For example, if children assume only syntactic information is relevant, they may ignore semantic cues that might otherwise be useful. Once the type of information children use is defined, corpus

analysis methods can provide realistic estimates of the input children get.

The *learning period* defines how long children have to reach the target state. Experimental methods can provide this information, usually by assessing the knowledge children have at a particular age, as demonstrated by their behavior. Often in computational studies, the learning period is implemented as children receiving a specific amount of data, which is the amount they would encounter during the learning period. After that quantity of data, they should then reach the target knowledge state.

The *target state* defines what knowledge children are trying to attain. Theoretical methods will specify this knowledge, and the particular representation it has. Notably, there may be different specifications, depending on the theoretical framework assumed. Sometimes, these different specifications are equivalent for the purposes of the induction problem. For example, determining which of two syntactic categories is the correct one for a particular item may be common to two frameworks, even if the two frameworks involve different labels for the syntactic category options.

An induction problem can then be characterized using these four components: Given a specific initial state, data intake, and learning period, the specified target state is not the only knowledge state that could be reached. Clearly, there can be different characterizations of an induction problem pertaining to the same linguistic phenomenon, because there may be differences in any one of these components. Thus, it is important to investigate the specific characterization that has been used to motivate a given UG learning bias. Notably, the initial state can be affected by the learning strategy used, since the learning strategy may involve particular biases about the data relevant for learning, for example. Since these aspects of the initial state are often stipulated, what looks like an induction problem with one characterization of the initial state may not be an induction problem with a different characterization. Relatedly, it is also useful to ask whether a particular learning strategy will work for other induction problem characterizations; to the extent that it does, this is stronger support for that learning strategy and the learning biases that comprise it.

1.2 The direct evidence assumption

Previous characterizations of induction problems motivating UG have tended to include a particular assumption in the initial state of the learner: the *direct evidence* assumption. The basic intuition of the direct evidence assumption is that in order to learn some linguistic knowledge L, children learn from examples of L in the linguistic input (*direct positive evidence*). It is 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, children would pay attention to examples of complex yes/no questions like (1a) and potentially notice 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?[2]

When learning the representation of anaphoric one in English, children would pay attention to examples of one being used anaphorically (2a) and potentially notice 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.

When learning to form complex wh-questions in English, children would pay attention to examples of complex wh-questions in English (3a-c) and potentially notice the absence of ungrammatical examples like (3d).

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

(3) Complex wh-question examples

(a) Who _*who* thinks the necklace is expensive?

(b) What does Jack think _*what* is expensive?

(c) Who _*who* thinks the necklace for Lily is expensive?

(d) *Who does Jack think the necklace for _*who* is expensive?

However, another kind of data that could be informative to children is *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 children (for example, due to their learning biases in the initial state). In particular, if the initial state includes knowledge of what counts as indirect positive evidence, the induction problem can now be characterized differently and may be solvable using different learning strategies than the ones previously proposed.[3] Recently, some computational modeling studies have been exploring the utility of indirect positive evidence for different induction problems (Foraker et al., 2009; Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2008; Pearl & Sprouse, in press; Perfors et al., 2011; Reali & Christiansen, 2005). We follow this promising approach here.

1.3 Case study: English anaphoric one

A specific characterization of an induction problem concerning 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, Waxman, and Freedman (2003); Lidz and Waxman (2004); Pearl (2007); Pearl and Lidz (2009); Pullum and Scholz (2002); Regier and Gahl (2004); Tomasello (2004); among others). Computational modeling studies have examined this character-

[3]Interestingly, indirect positive evidence is similar to what linguistic parameters are meant to do in generative linguistic theory – if multiple linguistic phenomena are controlled by the same parameter, data for any of these phenomena can be treated as an equivalence class, where learning about some linguistic phenomena yields information about others (Chomsky, 1981; Pearl & Lidz, in press; Viau & Lidz, 2011). The knowledge of the linguistic parameter is part of the initial state, and allows a broader set of data to be utilized.

ization and investigated learning strategies that alter the initial state of the learner in various ways affecting the data intake, while keeping the learning period and target state the same (Pearl & Lidz, 2009; Regier & Gahl, 2004). More specifically, each study has broadened the set of direct positive evidence a learner could use. In the current study, we investigate a learning strategy that broadens it further to include indirect positive evidence.

In the rest of this paper, we first briefly review the characterization of the learning problem under consideration, including the adult knowledge indicative of the target state and the child behavior thought to specify the learning period. We then highlight why anaphoric one has been considered an induction problem, given the available direct evidence, which specifies the data intake. Following this, we review previous proposals for learning strategies that solve this induction problem, and we describe a new learning strategy that additionally uses indirect positive evidence. We review the different kinds of information that are available in informative data points, and then present an online Bayesian learner adapted from Pearl and Lidz (2009) that uses this learning strategy. We show that a learner using this strategy reproduces the child behavior associated with correct knowledge of one. Surprisingly, this learning strategy leads to a different knowledge state than the target state, even though it produces the behavior thought to implicate the target state. This suggests that the link between observed behavior, interpretation, and knowledge representation may not be as transparent as once thought. In particular, very young children may not have the adult representation for one, and so the learning period characterizing this induction problem should actually be longer.

In addition, we compare our learner's indirect positive evidence strategy to the direct evidence learning strategies previously proposed, using the same online Bayesian learning framework. We replicate results previously found with the more restricted learning strategies, which suggests that it is the learner's view of the data intake that causes the new results we find, rather than something about the specific probabilistic learning framework chosen. We then discuss the nature of the learning biases comprising this learner's strategy, and more generally what children require in order

to solve this induction problem for anaphoric one. We also discuss alternate learning strategies as well as alternate characterizations of the induction problem. We conclude with how this impacts the larger debate about the contents of UG.

2 Characterizing the anaphoric one induction problem

While knowledge of one clearly goes beyond being able to correctly interpret examples like (2a) and recognize the ungrammaticality of (2b), the specific issue of representation for one in those cases has often been cited as an example of an induction problem for language acquisition (Baker, 1978; Crain, 1991; Hornstein & Lightfoot, 1981; Lightfoot, 1982a, 1989; Ramsey & Stich, 1991). More specifically, adult knowledge has been characterized as involving both a syntactic and semantic component. An example is shown in (4).

(4) Situation: Two red bottles are present.

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

Default interpretation of one:

syntactic antecedent of one = "red bottle"

semantic referent of one = RED BOTTLE

In order to interpret an utterance like (4), the listener must first identify the linguistic antecedent of one, i.e., what previously mentioned string one is standing in for. This is the syntactic component. In (4), adults generally interpret one's antecedent as "red bottle", so the utterance is equivalent to "Look - a red bottle! Oh, look - another *red bottle*!".[4] Then, the listener uses this antecedent to identify the referent of one, e.g., what object in the world one is referring to, and what properties that object has. This is the semantic component. Given the antecedent "red bottle", adults interpret

[4]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 the non-default interpretations more in section 8.3.

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.

An influential theoretical framework posited that the string "red bottle" has the structure in (5), while "a red bottle" has the structure in (6) (Chomsky, 1970; Jackendoff, 1977). 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]]]

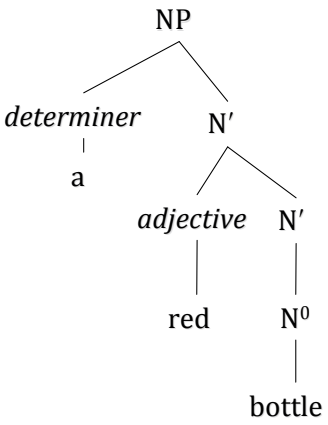

Figure 1: Phrase structure tree corresponding to the bracket notation in examples (5) and (6).

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" projects to both N^0 and N').[5]

[5]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'

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

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

This theoretical framework also posited that an anaphoric element (like one) can only have a linguistic antecedent of the same syntactic category as the element itself. Since one's antecedent can be "red bottle", then one should be category N′. 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, adults should be perfectly happy to have one's referent be a purple bottle. Since adults do not have this as the default 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 of the default interpretation of one for data like (4) 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 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 [$_{N'}$ [$_{N^0}$ bottle]]] rather than [$_{N'}$ [$_{N^0}$ bottle]].

Behavioral evidence from Lidz et al. (2003) (henceforth **LWF**) suggests that 18-month-olds also have this same interpretation for utterances like (4).[6] Using an intermodal preferential looking

here), while "bottle" can also be labeled with a smaller category label (N^0 here). However, see discussion in section 8.3 for what happens with alternate theoretical representations that additionally differentiate "red bottle" from "bottle".

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

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! Now look - 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.[7] Thus, LWF interpreted this to mean that by 18 months, children have acquired the same representation for anaphoric one that adults have. This then specifies the learning period.

In terms of the learner's initial state, the original proposal (Baker, 1978) (henceforth, **Baker**) assumed that only direct evidence was relevant, and that only unambiguous data were informative. LWF's corpus analysis of child-directed speech samples verified that these data were indeed too sparse to reach that target state given such a short learning period. In particular, they found that a mere 0.25% of child-directed anaphoric one utterances were unambiguous data, which is far below what theory-neutral estimates would suggest is necessary for acquisition by 18 months (Yang, 2004, 2011).

The induction problem for anaphoric one[8] can then be characterized as follows, and appears very real indeed.

(i) **Initial state**:

Knowledge: Syntactic categories exist, in particular N^0, N', and NP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

Bias: Only direct evidence of one is useful.

Bias: Only unambiguous evidence of one is useful.

[7]Moreover, LWF confirmed that infants responded similarly when the utterance was "Look, a red bottle! Now look - do you see another red bottle?", suggesting that they had correctly inferred that the antecedent of one in the original utterance was "red bottle". In addition, infants did not have this looking preference with control utterances such as "Look, a red bottle! Now look - what do you see now?", which suggests that they were using the language in the original utterance to determine which object to look at (in that case, the object indicated by the linguistic antecedent "red bottle").

[8]For ease of exposition, when we refer to knowledge of "anaphoric one" henceforth, we will mean knowledge of anaphoric one in examples such as (4).

(ii) **Data intake** (following biases in initial state):

All unambiguous one evidence in the input.

(iii) **Learning period**:

Completed by 18 months.

(iv) **Target state**:

Knowledge: In utterances like the example in (4), one is category N′ and its linguistic antecedent includes the modifier.

3 The direct evidence

Unambiguous data using anaphoric one are very rare because they require a very specific conjunction of situation and utterance.

(9) Unambiguous (**Unamb**) data example

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 (9), if the child mistakenly believes the referent is just a BOTTLE, then the antecedent of one is "bottle" and it's surprising that the speaker would claim there isn't "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 for the referent to have, so the referent must be a RED BOTTLE. Since there isn't another red bottle present, 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′.

There are other one data available, but they are ambiguous in various ways. Many one data are ambiguous with respect the syntactic category of one (10), even if children already know that the choice is between N′ and N^0.

(10) Syntactic (**Syn**) ambiguity example

Situation: There are two bottles present.

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

Syn ambiguous data like (10) do not clearly indicate the category of one, even though the property the referent must have is clear. In (10), the referent must be a 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].

As we saw in Figure 1, sometimes there is also more than one N' antecedent to choose from (e.g., “red bottle”: [$_{N'}$ red [$_{N'}$ [$_{N^0}$ bottle]]] vs. “bottle”: [$_{N'}$ [$_{N^0}$ bottle]]). In these cases, there is also ambiguity with respect to the referent (e.g., a RED BOTTLE vs. any BOTTLE), as shown in (11).

(11) Semantic and Syntactic (**Sem-Syn**) ambiguity example

Situation: There are two red bottles present.

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

Sem-Syn ambiguous data like (11) are unclear about both the properties of the referent and the category of one. In (11), if the child held the mistaken hypothesis that the referent must simply be a BOTTLE (unlike the adult interpretation of a 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 would be viewed as merely a coincidence. The alternative hypothesis is that the referent is a 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 referent are available, this data point is ambiguous semantically. This data point is ambiguous syntactically for the same reason Syn data like (10) are: If the referent is a BOTTLE, then the antecedent is “bottle”, which is either N^0 or N'.

4 Previous solutions to the induction problem

4.1 Adding additional knowledge to the initial state

The solution proposed by Baker was that children must know that anaphoric elements (like one) cannot be syntactic category N^0. Instead, children automatically rule out that possibility from their hypothesis space.[9] Baker's solution thus updates the initial state as follows:

Baker's update of the initial state:

Knowledge: Syntactic categories exist, in particular N^0, N', and NP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

Bias: Only direct evidence of one is useful.

Bias: Only unambiguous evidence of one is useful.

Knowledge: One is not category N^0.

Because this knowledge is domain-specific and was assumed to be innate, this solution is a UG learning bias, and in fact specified a proposal for one piece of UG. Of course, as is apparent from the original characterization of the induction problem, domain-specific knowledge was already assumed in the initial state of the learner. Whether that other knowledge must be innate or could instead be derived from prior experience with language is unclear – importantly, that was not relevant to the debate concerning the solution to this characterization. In particular, even if that other initial state knowledge was necessarily innate (which is not at all clear), the induction problem *still* exists, and one solution is this UG knowledge that Baker proposed.

[9] 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].

4.2 Updating the initial state in other ways

4.2.1 Regier & Gahl 2004

Regier and Gahl (2004) (henceforth **R&G**) investigated a learning strategy that assumed children had the ability to do Bayesian inference and were not restricted to learning from unambiguous data. Specifically, a Bayesian learner could learn something from Sem-Syn data like (11) by tracking 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. Their solution involved updating the initial state as follows:

R&G's update of the initial state:

Knowledge: Syntactic categories exist, in particular N^0, N′, and NP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

Bias: Only direct evidence of one is useful.

~~**Bias: Only unambiguous evidence of one is useful.**~~

Bias: Use Bayesian inference.

R&G reasoned that removing the restriction to unambiguous evidence and using Bayesian inference were unlikely to be part of UG. Thus, their solution to the induction problem did not require additional UG components.

4.2.2 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**) Bayesian learner able to extract information from ambiguous data (like R&G's learner) would view Syn ambiguous data like (10) as informative, as well. 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). Moreover, Syn ambiguous data far outnumber the Sem-Syn ambiguous and unambiguous data combined (about 20 to 1 in P&L's corpus analysis). Thus, a Bayesian learner like R&G proposed would need to explicitly filter out the Syn ambiguous data. This learning strategy updates the initial state as follows:

P&L's update of the initial state:

Knowledge: Syntactic categories exist, in particular N^0, N', and NP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

Bias: Only direct evidence of one is useful.

~~**Bias: Only unambiguous evidence of one is useful.**~~

Bias: Use Bayesian inference.

Bias: Only unambiguous and Sem-Syn data are useful.

P&L suggested that this kind of data intake filter is domain-specific, since it involves ignoring a specific kind of linguistic data. While this could be innate (and so part of UG), they speculate how this restriction could be derived from innate domain-general learning biases.[10] To the extent that is true, P&L's solution to the induction problem also did not require a UG component, though it did add a restriction to the data intake.

[10] In particular, they suggest that a learner who learns only when the current utterance's referent is ambiguous 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, and how it derives from domain-general learning principles).

5 Another solution: Removing the direct evidence bias

5.1 The learning strategy

Instead of restricting the input set, we consider a learning strategy that expands it beyond unambiguous (9), Sem-Syn ambiguous (11), and Syn ambiguous (10) 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 a specific string is not unique to one. These other pronouns would be category NP in the current induction problem characterization, since they replace an entire noun phrase (NP) when they are used, as in (12):

(12) "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":

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

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

(14) "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)).[11] However, since one is similar to other pronouns referentially (by being anaphoric and having linguistic antecedents) and shares some syntactic distribution properties with them (since

[11] This shows that one clearly has some categorical flexibility, since it can be both NP and smaller than NP. However, it appears to be conditional on the linguistic context, rather than being a probabilistic choice for any given context. For example, it is not the case that in examples like (14) one can alternate between NP and N′. Instead, in (14) it is always NP, while in unambiguous utterances like (9), it is always N′. We will assume (along with previous studies) that children prefer referential elements to have as few categories as possible (ideally, just a single category), which is why they must choose between N′ and N^0 when one is smaller than NP for ambiguous examples like (4), (10), and (11).

it can appear as an NP), a learner could decide that information gleaned from other pronouns is relevant for interpreting one.

This bias to use other pronoun data can be combined with a bias to use Bayesian inference, similar to R&G's and P&L's learners. In particular, a learner could track how often a property mentioned in the potential antecedent (e.g., "red" in "a red bottle" in (14)) is important for the referent to have (and so also important for the antecedent to contain). 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 ((12) and (14)). 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 mentioned property (e.g., "a red bottle"). Data points like (12) and (14) 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. The initial state for the P&M learning strategy is thus updated as follows:

P&M's update of the initial state:

Knowledge: Syntactic categories exist, in particular N^0, N', and NP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

~~**Bias: Only direct evidence of one is useful.**~~

~~**Bias: Only unambiguous evidence of one is useful.**~~

Bias: Use Bayesian inference.

Bias: Learn from other pronoun data.

Like the R&G and P&L learning strategies, our learning strategy differs from the Baker strategy by learning from data besides the unambiguous <NP data. However, our strategy differs from the strategies in R&G and P&L by learning from data containing anaphoric elements besides one, since this is viewed as indirect positive evidence. Table 1 shows which learning strategies use which data.

Table 1: Data sets used by different learning strategies.

Data type	Example	Learning strategies
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

We will save detailed discussion of the nature of the biases involved in the P&M learning strategy for section 8.2, specifically the bias to learn from other pronoun data. If this is a UG bias, then this is a specific proposal about the contents of UG that differs from the Baker proposal. Conversely, if this bias is unlikely to be a UG bias, this is another solution to the induction problem that does not appear to require a UG learning bias.

5.2 Information in the data

There is a variety of information in anaphoric data points. Figure 2 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.[12] These data include both referential and syntactic information (the variables grouped under REFERENTIAL INTENT and SYNTACTIC USAGE), and both types of information are relevant for determining the antecedent, which determines the properties the referent must have.

Under REFERENTIAL INTENT, there are three variables. First, 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") (**Property mentioned?**). If a property was mentioned, it is a latent variable whether the mentioned property is important for the referent of one to have (**Property important?**), which

[12]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, it encodes the dependencies between the different variables, and inferences flow both directions along the information dependencies.

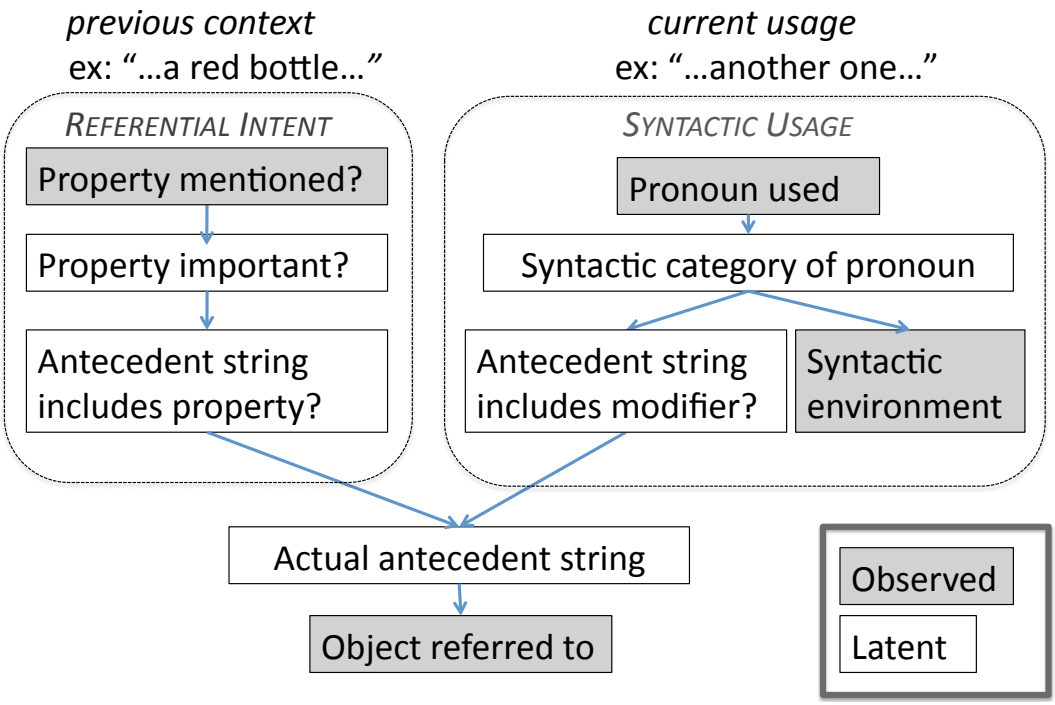

Figure 2: Information dependencies in referential data points.

determines whether the antecedent string must contain the property (e.g., it must if the property is important, and it must not if the property is not important) (**Antecedent string includes property?**).

Under SYNTACTIC USAGE, there are four variables. A learner can observe which pronoun is used (e.g., it, one, etc.) (**Pronoun used**). The syntactic category depends on which pronoun is used (e.g., NP, N′, or N^0 for one) (**Syntactic category of pronoun**). The learner can also observe the syntactic environment in which the pronoun is used (**Syntactic environment**), which depends on the latent syntactic category (e.g., "another one" indicates a syntactic environment of $<$NP, which 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"). If a modifier is possible, it is a latent variable whether the antecedent actually does include the modifier (**Antecedent string contains modifier?**).

Both the antecedent string variables determine the content of the actual antecedent string, which is a latent variable (**Actual antecedent string**). The referential component determines whether the antecedent must contain the property, while the syntactic component determines whether the antecedent must contain a modifier. Only two combinations are viable:

(i) If the antecedent must contain the property mentioned and must contain the modifier mentioned (e.g., the antecedent would be "red bottle")

(ii) If the antecedent must not contain the property mentioned and must not contain the modifier mentioned (e.g., the antecedent would be "bottle")[13]

The antecedent string determines what object is being referred to, and whether that object must have the mentioned property (e.g., whether it's a RED BOTTLE when the previous context was "a red bottle"). The learner can observe whether the object has the mentioned property (e.g., the learner can ascertain if the bottle that one refers to is in fact red) (**Object referred to**), even though the learner cannot directly observe the antecedent.

These variables can take on the values shown in table 2.[14] The data types used by the different learning proposals have the observable and latent values in Table 3.

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."), 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., a 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

[13]The remaining combinations involve an incompatibility:

(iii) If the antecedent must not contain the property mentioned but must contain the modifier mentioned, there is no property that can fill the modifier position of the antecedent.

(iv) If the antecedent must contain the property mentioned but must not contain the modifier mentioned, there is no place syntactically for the property to be instantiated in the antecedent.

[14]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).

Table 2: 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 3: Data types and variable values.

	Variable	**Unamb <NP**	**Sem-Syn Ambig**	**Syn Ambig**	**Unamb NP**
Observable	Property mentioned	Yes	Yes	No	Yes
	Pronoun	one	one	one	it, one, etc.
	Syntactic environment	<NP	<NP	<NP	NP
	Object	has property	has property	*N/A*	has property
Latent	Property important	Yes	Yes, No	*N/A*	Yes
	Antec has property	Yes	Yes, No	*N/A*	Yes
	Syntactic category	N′	N′, N^0	N′, N^0	NP
	Antec has modifier	Yes	Yes, No	*N/A*	Yes
	Antec string	ex: "red bottle"	ex: "red bottle", "bottle"	ex: "bottle"	ex: "a red bottle"

property (from the referential perspective). This indicates that the mentioned property is important for the referent to have 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., a 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 for the referent to have, 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 an additional property of that object (e.g., a BOTTLE). 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 for the referent to have. 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., a 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 for the referent to have.

6 The online probabilistic learning framework

We now present an online probabilistic learning framework that uses the different kinds of information available in the anaphoric data types described above.

6.1 Important quantities

The two components of the correct representation for anaphoric one in the default context are

(a) that a property mentioned in the potential antecedent is important for the referent of one to have (and so is part of one's antecedent), and

(b) that one is category N' when it is not an NP.

Recall that these variables can take only take two values: Yes or No for the former and N′ or N^0 for the latter. These correspond to **property important?** (given that it has been mentioned) and **syntactic category of pronoun** (given that it is smaller than NP) in Figure 2. Our modeled learner will determine the probability associated with a particular value for both of these variables, specifically *p*(*property important=yes* | *property mentioned=yes*) and *p*(*category=N′* | *syntactic environment=<NP*). We represent the probability of the former as p_I and the probability of the latter as $p_{N'}$. If the correct representation of one has been learned, both probabilities should be near 1.

We follow the update methods in P&L, and use equation (15) adapted from Chew (1971), which assumes *p* comes from a binomial distribution and the beta distribution is used to estimate the prior. It is reasonable to think of both p_I and $p_{N'}$ as parameters in binomial distributions, given that each variable takes on only two values, as noted above.

$$p_x = \frac{\alpha + data_x}{\alpha + \beta + totaldata_x}, \alpha = \beta = 1 \tag{15}$$

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 (16)[15], and then p_x is updated using equation (15). The variable ϕ_x indicates the probability that the current data point is an example of an x data point. For unambiguous data, $\phi_x = 1$; for ambiguous data $\phi_x < 1$.

[15]Note that the := symbol should be read as "becomes equal to".

$$data_x := \quad data_x + \phi_x \tag{16a}$$

$$totaldata_x := \quad totaldata_x + 1 \tag{16b}$$

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 2, and the basic Bayes equation. ϕ_I uses equation (17), 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 (i.e., $p_I = p(I|\mu = yes)$).

$$\phi_I = p(I|\pi,\sigma,\mu = yes,\omega) = \frac{p(\pi,\sigma,\omega|I,\mu = yes) * p_I}{p(\pi,\sigma,\omega|\mu = yes)} \tag{17}$$

When ϕ_I is calculated for Unambiguous <NP and Unambiguous NP data using (17), it can be shown that ϕ_I = 1, which is intuitively satisfying since these data unambiguously indicate that the property is important for the referent to have. When ϕ_I is calculated for Sem-Syn ambiguous data using (17), it can be shown that ϕ_I is equal to (18):

$$\phi_I = \frac{\rho_1}{\rho_1 + \rho_2 + \rho_3} \tag{18}$$

where

$$\rho_1 = \quad p_{N'} * \tfrac{m}{n+m} * p_I \tag{19a}$$

$$\rho_2 = \quad p_{N'} * \tfrac{n}{n+m} * (1 - p_I) * \tfrac{1}{t} \tag{19b}$$

$$\rho_3 = \quad (1 - p_{N'}) * (1 - p_I) * \tfrac{1}{t} \tag{19c}$$

In (19), *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 (19a), as compared to an N′ string that contains only a noun (19b). 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. Parameters *m*, *n*, and *t* are implicitly estimated by the learner, and will be estimated from child-directed speech corpus frequencies when possible when we implement our learner.

The quantities in (19) can be intuitively correlated 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) - this corresponds to the antecedent being "red bottle" = [$_{N'}$ red [$_{N'}$ [$_{N^0}$ bottle]]]. 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 ($\frac{1}{t}$) - this corresponds to the antecedent being "bottle" = [$_{N'}$ [$_{N^0}$ bottle]]. 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 ($\frac{1}{t}$) - this corresponds to the antecedent being "bottle" = [$_{N^0}$ bottle]. The numerator of (18) 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. Thus, $p_{N'}$ = $p(N'|\sigma =< NP)$.

$$\phi_{N'} = p(N'|\pi,\sigma =< NP,\mu,\omega) = \frac{p(\pi,\mu,\omega|N',\sigma =< NP) * p_{N'}}{p(\pi,\mu,\omega|\sigma =< NP)} \tag{20}$$

When $\phi_{N'}$ is calculated for Unambiguous <NP data using equation (20), it can be shown that $\phi_{N'}$=1, which is again intuitively satisfying since these data unambiguously indicate that the category is N$'$ when the syntactic environment is <NP. When $\phi_{N'}$ is calculated for Sem-Syn ambiguous data using (20), it can be shown that $\phi_{N'}$ is equal to (21):

$$\phi_{N'Sem-Syn} = \frac{\rho_1 + \rho_2}{\rho_1 + \rho_2 + \rho_3} \tag{21}$$

where ρ_1, ρ_2, and ρ_3 are the same as in (19). Equation (21) is intuitively satisfying as only ρ_1 and ρ_2 are correlated with representations with syntactic category N$'$.

When $\phi_{N'}$ is calculated for Syn Ambiguous data using equation (20), it can be shown that $\phi_{N'}$ is equal to (22):

$$\phi_{N'Syn} = \frac{\rho_4}{\rho_4 + \rho_5} \tag{22}$$

where

$$\rho_4 = \quad p_{N'} * \frac{n}{n+m} \tag{23a}$$

$$\rho_5 = \quad 1 - p_{N'} \tag{23b}$$

The quantities in (23) 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 ($\frac{n}{n+m}$) - this corresponds to the antecedent being "bottle" = [$_{N'}$ [$_{N^0}$ bottle]]. For ρ_5, the syntactic category is N^0 (1-$p_{N'}$), and so the string is noun-only by definition - this corresponds to the antecedent being "bottle" = [$_{N^0}$ bottle]. The numerator of equation (22) contains the representation that has the category as N′, while the denominator contains both possible representations.

Table 4 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 conservative 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). This means that if our learners converge on a high p_I given this estimate of t, they should certainly converge on a high p_I with the higher t values that are likely to exist in realistic learning scenarios.

Table 4: 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 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 for the referent to have) and $p_{N'}$ (the syntactic category is N′).

For Sem-Syn ambiguous 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.[16] 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 ambiguous 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 ambiguous

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

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.

6.2 Learner input sets & parameter values

Table 5 indicates the availability of different data types in the learner's input, based on a corpus analysis of the Brown-Eve corpus (Brown, 1973) from the CHILDES database (MacWhinney, 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.[17]

Table 5: 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 include ungrammatical uses of anaphoric one, uses of one where no potential antecedent was mentioned in the previous

[17]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.

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 5 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.[18] Table 6 shows the input sets we will use to test the different learning proposals for anaphoric one, based on the data each learning strategy considers relevant for learning.

Table 6: 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 $n = 3$.[19] These parameters matter when

[18] Specifically, 2,874 of the 17,521 utterances from the Eve corpus were referential pronoun data points, which is approximately 16.4%. The number of utterances children would hear between 14 and 18 months is approximately 1,000,000*4/18, which is 222,222. We multiply 222,222 by 16.4% to get the number of referential pronoun data points heard during this period, which is 36,452, and we round that to 36,500.

[19] 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.

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 ambiguous 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 conservative 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 6.1. Heeding suspicious coincidences specifically aids the learner in deciding that the mentioned property is important for the referent to have (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.

6.3 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 the default adult representation. 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! Now look - 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 ($\mu = yes$).

$$p_{beh} = p(\omega = hasproperty | \pi = \mathsf{one}, \sigma =< NP, \mu = yes) \tag{24}$$

Using the information dependencies in Figure 2 and Bayes equation, this works out to

$$p_{beh} = \frac{\rho_1 + \rho_2 + \rho_3}{\rho_1 + 2 * \rho_2 + 2 * \rho_3} \tag{25}$$

where ρ_1, ρ_2, and ρ_3 are defined as in (19), 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 different 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 interpreting their experiment - in particular, if infants look at the object adults look at when adults have the target representation of anaphoric one, it means that the children also have the target representation. While this does not seem like an unreasonable assumption, it is worth verifying that this is true. It is possible, for example, that children have a different representation, but look at the correct object by chance (represented in the numerator of (25) as ρ_2 and ρ_3). Given this, there are two related questions that

we can ask.

First, is it possible to get adult-like behavior in the LWF experiment without having the adult 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 adult-like behavior may not necessarily implicate the target representation in general.

Second, is it possible to get adult-like behavior in the LWF experiment without having the target representation for one at the time the behavior is being generated? To answer this question, we can calculate the probability ($p_{rep|beh}$) that the learner has the target representation, given that the learner has produced the adult 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.

$$p_{rep|beh} = p(N', I|\pi = \mathsf{one}, \sigma =< NP, \mu = yes, \omega = hasproperty) \tag{26}$$

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 target 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 for the referent to have (I), given that the pronoun used was one ($\pi = \mathsf{one}$), 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 2 and Bayes equation, this works out to

$$p_{rep|beh} = \frac{\rho_1}{\rho_1 + \rho_2 + \rho_3} \tag{27}$$

where ρ_1, ρ_2, and ρ_3 are calculated as in (19), 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 target representation (ρ_1).[20]

7 Results

Table 7 shows the results of the learning simulations over the different input sets, with averages over 1000 runs reported and standard deviations in parentheses.[21]

Table 7: Probabilities after learning

Prob	**Baker**	**R&G, P&L filtered**	**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}	0.53 (<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 replicating the results that previous studies found when using the data sets proposed by those learning strategies. 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 target representation, as R&G and P&L's filtered learners did ($p_{N'} = 0.97$, $p_I = 0.95$). Additionally including Syn ambiguous data, as P&L's

[20] Note that this is the same equation as (18) (the only difference is the value of t). This has some intuitive appeal since $\rho 1$ in (19) corresponds to the correct representation which has the mentioned property as important, while the other two representations do not.

[21] Note that averaging over 1000 runs means that the learner's input distribution was drawn from the distribution in Table 6 for each run, but the order of data types encountered may differ from run to run.

EO learner did, leads to a different representation where one's category is N^0 and the mentioned property is not important for the referent to have ($p_{N'} = 0.17$, $p_I = 0.02$).

The new result we have found is that expanding to include unambiguous NP data (P&M) does not lead to the target 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 adult-like behavior in the LWF experiment ($p_{beh} > 0.99$). That is, the learner could have a different 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 for the referent to have, then one's antecedent must have that property in it (e.g., "red bottle", when "red" was mentioned in the potential antecedent), and so the object referred to by one must have that property (e.g., be a RED BOTTLE). So, the learner infers the correct antecedent, looks to the referent that has the property, and so produces adult-like behavior. Thus, it seems that LWF's assumption does not hold - producing adult-like behavior does not necessarily indicate that the learner has the target representation in general.

However, a relaxed version of the LWF assumption does appear to hold. In particular, when the child produces adult-like behavior, the probability that the child has the target 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 for the referent to have. If the property is important, then 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 linguistic context*.

Thus, even though the learner has a non-adult representation in general, in the context where a modifier is present, the learner will end up with the adult-like interpretation and the adult representation. LWF were not wrong to assume adult-like behavior was due to an adult representation - it's simply that the adult representation may not apply generally. In particular, the P&M learner

will have a different 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.[22] The only way to tell that the P&M learner has a non-adult representation in general would be to test its acceptance of ungrammatical utterances where one is used as syntactic category N^0, such as "I'll sit by the side of the river and you sit by the one of the building." Given that the P&M learner believes one can be category N^0 with some probability (p_{N^0}=1-$p_{N'}$ = 0.63), it would find utterances like these grammatical some of the time, which clearly differs from adult behavior.

We note that this result is due to the input set the P&M learning strategy is using - the learning strategies that use restricted input sets behave exactly as LWF would expect. When they have the target representation in general (R&G, P&L filtered), they produce the correct behavior and have the target 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 a different representation if they happen to produce the correct behavior.

We additionally note that this result is not due to the particular duration of the learning period we chose. As Figure 3 shows, the P&M learner converges on these probabilities fairly quickly, with very little change to the probabilities occurring after the first few hundred data points. Thus, we would not predict the behavior of the P&M learner to alter appreciably if it was exposed to more data, unless those data were very different from the data it had been learning from already or it was able to use those data in a very different way.

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

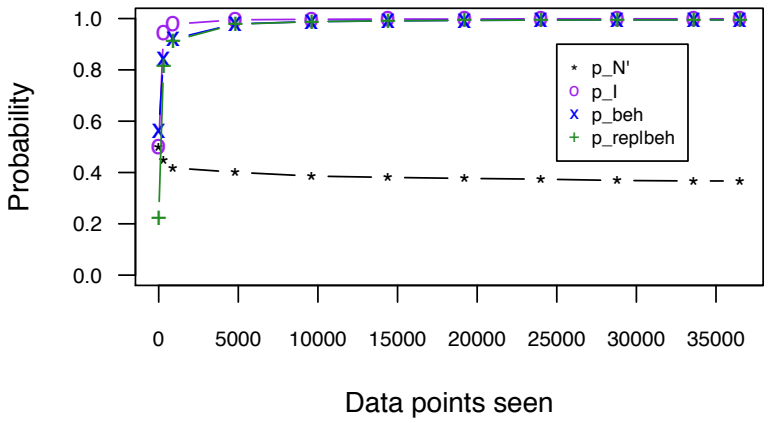

Figure 3: P&M probabilities over the learning period.

8 Discussion

8.1 General discussion of results

Through this modeling study, we have provided new information about the acquisition of knowledge concerning English anaphoric one. First, using a learning strategy that draws on indirect positive evidence, a child would be able to produce the behavior at 18 months that was thought to indicate the target knowledge state, presumably solving the induction problem. However, surprisingly, this behavior can be produced without reaching the target state - instead, a child with an immature context-dependent representation of one could produce the observed behavior. This suggests that the link between observed behavior, interpretation, and representation may not be as clear cut as once thought. Even though children demonstrate they have the adult interpretation some of the time (by displaying adult-like behavior), this does not necessarily mean they have the adult representation all of the time. We have provided an example learning strategy that would lead to the adult-like interpretation in the the context of the LWF experiment, but would not lead to the adult representation for other utterances, like those in Syn ambiguous data.

This suggests an update of the learning problem characterization. If we want the target state to remain unchanged, then the learning period may not be restricted to 18 months. Instead, it could be

that children achieve the target knowledge state later on. If so, this means they may have access to additional data, knowledge, and learning capabilities to solve the induction problem that we did not allow the learners modeled here. We briefly discuss one example of this kind of solution in section 8.4.1. More generally, it would suggest a two-stage acquisition trajectory for anaphoric one, with the first stage completed by 18 months and the second stage completed sometime afterwards.

More broadly, the results here also demonstrate how using indirect positive evidence may be useful for investigating solutions to induction problems. In particular, by relaxing the direct evidence assumption, we may find that the behavior we observe in children can be explained, given the data in children's input.

With respect to testing proposals for the contents of UG, we have also described how specific characterizations of induction problems motivate specific proposals. In particular, when a learning strategy succeeds, we can examine the learning biases that comprise it and discuss whether they are likely to be in UG. We thus examine the biases that are part of the indirect positive evidence learning strategy used here below in section 8.2. In addition, we discuss alternate learning strategies that might be useful for the induction problem characterization explored here, as well as alternate characterizations of the induction problem, and how this impacts the debate about the contents of UG.

8.2 The learning biases of the P&M learning strategy

The P&M learning strategy includes two biases that enrich the initial state of the learner:

(a) Use Bayesian inference.

(b) Learn from other pronoun data.

The bias to use Bayesian inference to leverage information in the data has been part of proposed learning strategies before, specifically the strategy of R&G and P&L that restricted the data intake. Since Bayesian inference can be used for other kinds of data besides language data, it is unlikely to be a domain-specific strategy (though it is likely innate). This means it would not be a UG learning

bias.

The bias to learn from other pronoun data clearly concerns language data, and so would be domain-specific. But is it innate or derived? It is possible that this bias results from innate knowledge that referential pronoun data can be treated as an equivalence class. If this were true, this would be a UG learning bias. Conversely, it could be possible to derive this bias from prior linguistic experience with the pronouns of English. In particular, while one does not have an identical distribution to other referential elements like it (e.g., "another one", but not "another it"), the distribution overlaps significantly (e.g., "I see one", "I see it", etc.). If a child was sensitive to this distributional data, it may be possible to derive the knowledge that these data are relevant for learning about anaphoric one, and so can serve as indirect positive evidence.

While we have no evidence that discerns between these two options, the study here can be seen as either providing a different characterization of the contents of UG or providing another non-UG way to generate the behavior we see in 18-month-olds. In particular, if the second bias is innate, this is then a specific proposal about the contents of UG that differs from the original Baker proposal: Instead of explicitly limiting the hypothesis space, the desired behavior can be produced by broadening the data intake. If this second bias is instead derived, this is an alternate non-UG learning strategy that will produce the desired behavior, in addition to the R&G and P&L filtering strategy that restricts the data intake.

8.3 Other learning strategies

8.3.1 Using data more effectively

The Bayesian learning model we used 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.[23] 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.

8.3.2 Using sophisticated contextual cues

Another source of information involves more sophisticated contextual cues. Some examples are shown below in (28):

(28a) "I hate that red bottle - do you have another one?"

(28b) "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 (28a), this is perhaps based on the verb "hate", and the inference that someone would not ask for another of something they hate . For (28b), this is perhaps based on the contrastive focus that occurs

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

between “red” and “that”. In both cases, this involves an inference that draws from information beyond the default syntactic and semantic representation. In (28a), this is an inference about when a speaker would use “hate” in this way; in (28b), this is an inference about when speakers use contrastive focus. The default interpretation of one seems to include the modifier (see 29). In (29a), it seems the speaker is requesting another red bottle. In (29b), while there is contrastive focus with “that”, it doesn’t interfere with the interpretation of one’s antecedent as “red bottle”.

(29a) “I love that red bottle - do you have another one?”

(29b) “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 (28a) in our corpus analysis, which suggests that young children probably do not encounter these data very often.[24] 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 (28), as this would allow us to further fine-tune the acquisition trajectory.

8.3.3 Favoring the larger category

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, an alternate strategy is to prefer the largest set compatible with the observable data.[25] For instance, given a noun-only string like “bottle” that is compatible with category N′ and

[24] Our corpus was not marked for constrastive focus, so it is unclear how often data like (28b) appear.

[25] Thanks to Ming Xiang for suggesting this.

category N^0, this learner would prefer to choose the hypothesis that covers a larger set of strings (N').

This kind of bias would 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.

Where does this learning bias come from? It must be explicit because it does not implicitly fall out from the mechanics of Bayesian inference. For the dimension of domain-specific vs. domain-general, it could be domain-general if it applies to other data besides language data; conversely, it could be 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 correct.

8.4 Alternate induction problem characterizations

There are different ways to characterize the learning problem concerning anaphoric one, only one of which we have explored here. Below we briefly discuss two additional ways which are similar, but crucially differ on the target state, or both the initial state and the target state. We highlight when and how these characterizations lead to different proposals about the contents of UG.

8.4.1 A different target state

Another characterization of this learning problem focuses on the syntactic representation alone, where one is N′ when it is smaller than NP. The target state of this characterization can be described as follows, updated from the characterization explored in the current study:

(iv) **Target state**:

Knowledge: In utterances like the example in (4), one is category N′. ~~and its linguistic antecedent includes the modifier~~.

This was actually the target state in Baker's original formulation of the induction problem. Recently, Foraker et al. (2009) (henceforth **F&al**) have investigated a learning strategy that could be used to solve this characterization of the learning problem. Unlike the other strategies explored here, this learner only learned from syntactic data, rather than also using the semantic information available. Similar to the indirect positive evidence strategy explored in this study, F&al removed the bias to learn only from direct evidence. Similar to all the strategies investigated here, Bayesian inference was used. In addition, F&al's learning strategy employed subtle conceptual knowledge in order to identify that category of one. Specifically, their learner was able to distinguish syntactic *complements* from syntactic *modifiers*, where a syntactic 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 4 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.

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 (30a), while "one with dots" is grammatical (30b).

(30a) *Lily waited by the side of the building while Jack sat by the one of the road.

(30b) Lily was fond of the ball with stripes while Jack preferred the one with dots.

Thus, the initial state for F&al's learning strategy would be updated as follows:

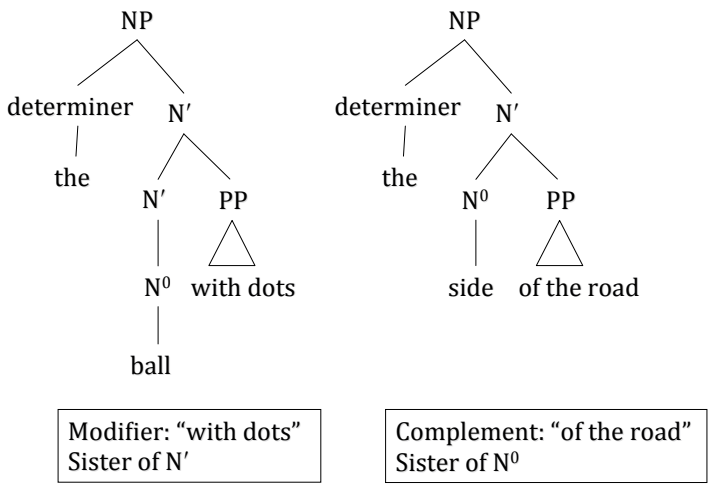

Figure 4: Phrase structure trees corresponding to a modifier and a complement.

(i) **Initial state**:

Knowledge: Syntactic categories exist, in particular N^0, N', and NP.

~~Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.~~

~~Bias: Only direct evidence of one is useful.~~

Bias: Only unambiguous evidence of one is useful.

Bias: Only syntactic data are useful.

Bias: Use Bayesian inference.

Bias: Learn from all linguistic elements that take complements or modifiers.

Knowledge: Complements conceptually evoke their head noun while modifiers do not.

Knowledge: Syntactic category N^0 is sister to a complement, not a modifier.

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 learning strategy can track the complement-modifier distribution of linguistic elements such as one and compare it to other elements that are syntactic category N^0. In particular, a Bayesian learner can note the absence of one being used with complements. This then indicates that one is not N^0, but rather N'. While there were not

many informative one data points in their data, F&al's ideal Bayesian learner was able to learn the correct syntactic category for one.

But what of the additional biases and knowledge in the initial state required to achieve this solution? We consider each in turn. The bias to use only syntactic data is clearly domain-specific, but could perhaps be derived from the target state concerning only the syntactic representation - syntactic data could be the natural choice for informative data in this case. The bias to use Bayesian inference is likely innate, but also likely domain-general since Bayesian inference can be used in many cognitive domains. The bias to learn from all linguistic elements taking complements or modifiers is the indirect positive evidence bias. Similar to the indirect positive evidence bias the P&M learning strategy used, it could be specified innately that these elements should be heeded, and so be a UG bias. Conversely, it could be derived somehow, perhaps from noticing salient properties of nominal categories. The semantic knowledge that complements conceptually evoke their head nouns seems to be clearly domain-specific, as does the syntactic knowledge relating N^0 to complements. While it is possible that this knowledge is derived somehow, we could not think of an obvious way to do so - thus, these knowledge components would likely be part of UG.

From this, we see that considering this version of the induction problem leads to a different proposal for the contents of UG. At the very least, detailed semantic and syntactic knowledge is required concerning complements and modifiers, and it is also possible that the bias to pay attention to the indirect positive evidence offered by other linguistic elements taking complements and modifiers is part of UG. Still, the target state is reachable, given this enriched initial state. Since this learning strategy does not consider the semantic component of anaphoric one, it is unclear how well it would match the behavior of 18-month-olds observed in the LWF experiment, however.

8.4.2 A different initial and target state

Another characterization of the induction problem assumes different syntactic categories than the ones in the characterization we examined here. 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 5, and shown in bracket notation in (31a). However, an alternate representation of noun phrases is available (Bernstein, 2003; Longobardi, 2003)[26], shown in (31b) and the right side of Figure 5. 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′.

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

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

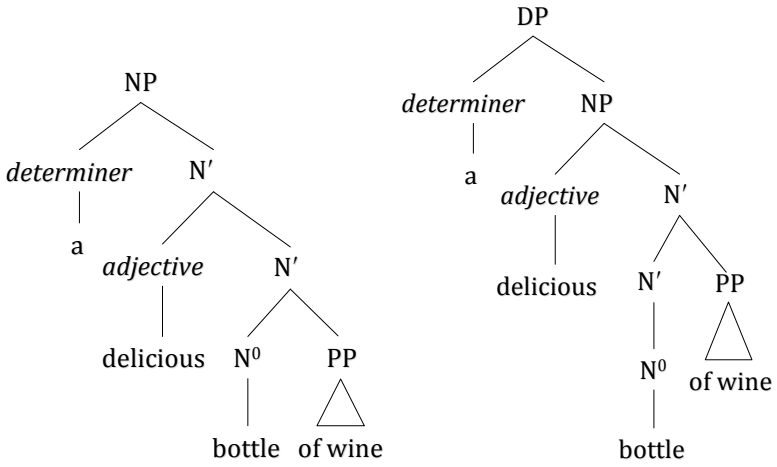

Figure 5: Phrase structure trees corresponding to the bracket notation in examples (31) and (32) for "a delicious bottle of wine".

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 (32)), when it is smaller than DP. 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

[26]Thanks to Greg Kobele for noting this.

bottle" is an NP in this representation).

(32) "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 ≠ "bottle", category N^0]

The initial and target states for the induction problem can then be updated as follows:

(i) **Initial state**:

Knowledge: Syntactic categories exist, in particular N^0, N′, NP, and DP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

Bias: Only direct evidence of one is useful.

Bias: Only unambiguous evidence of one is useful.

(iv) **Target state**:

Knowledge: In utterances like the example in (4), one is category NP and its linguistic antecedent includes the modifier.

While we have not implemented a learning strategy that uses this syntactic representation, we can easily speculate on the results we might find with an indirect positive evidence strategy like the P&M strategy proposed here, as there are still many similarities in the learning problem. As before this strategy would update the initial state as follows:

(i) **Initial state**:

Knowledge: Syntactic categories exist, in particular N^0, N′, NP, and DP.

Knowledge: Anaphoric elements like one take linguistic antecedents of the same category.

~~Bias: Only direct evidence of one is useful.~~

~~Bias: Only unambiguous evidence of one is useful.~~

Bias: Use Bayesian inference.

Bias: Learn from other pronoun data.

When faced with Syn ambiguous data (e.g., "Look - a bottle! Oh, look - another one!"), there is still a two-way ambiguity (N' vs. N^0), since "bottle" projects to both N' and N^0. When given data compatible with two hypotheses, a Bayesian learner will prefer the hypothesis that covers a smaller set of items (due to the Size Principle (Tenenbaum & Griffiths, 2001)). This is the N^0 category hypothesis, since all noun strings (like "bottle") are included in both hypotheses, but noun+complement strings (like "bottle of wine") are additionally included in the N' hypothesis. This means that the Syn ambiguous data will cause the learner to 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).

In addition, 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 when one is smaller than DP. Therefore, the learner will likely perform well in the LWF experiment, as long as p_I is high. This is again similar to the behavior the P&M learning strategy produced.

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. Moreover, the learning biases involved are the same as before the P&M strategy, and so the implications for UG remain the same as discussed above in section 8.2. This is an example where the same learning strategy will work over multiple characterizations of an induction problem. Thus, the distinction between these characterizations does not affect the proposal for the contents of UG.

9 Conclusion

In this paper, we have explicitly characterized an induction problem concerning English anaphoric one that has been used to motivate specific proposals for the contents of UG. In particular, we noted how theoretical assumptions about the knowledge representation and experimental data concerning the acquisition trajectory have been used to specify different components of this induction problem. We then demonstrated that a probabilistic learning strategy using indirect positive evidence can produce the behavior observed experimentally in young children – even when the target knowledge state had not been reached. This suggests that immature representations may persist longer than realized, with children producing adult-like behavior even though their representations are not adult-like. This in turn motivates an alternate form of the learning problem where acquisition of anaphoric one knowledge proceeds in stages, and the learning period for anaphoric one is longer than previously thought. In addition, we described how explicit computational models implementing different strategies can be used to offer concrete proposals for the contents of UG. In particular, indirect positive evidence does not necessarily negate the need for innate, domain-specific learning biases - it may, however, alter the exact form those biases take. We believe this general approach of broadening the data intake for language acquisition may be fruitful for identifying what is and is not necessarily part of UG.

10 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 suggestions from Max Bane, Morgan Sonderegger, Greg Kobele, Ming Xiang, Sue Braunwald, five anonymous reviewers, 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, as well as the audiences at

CogSci 2011. All errors are, of course, are own and not at all their fault. 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.

Bernstein, J. (2003). The DP Hypothesis: Identifying Clausal Properties in the Nominal Domain. In M. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory.* Oxford, UK: Blackwell.

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. (1970). Remarks on monimalization. In R. Jacobs & P. Rosenbaum (Eds.), *Reading in English Transformational Grammar* (pp. 184–221). Waltham: Ginn.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 237–286). New York: Holt, Rinehart, and Winston.

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

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

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

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

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. (1998). 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 Acquisition* (pp. 9–31). London: Longman.

Jackendoff, R. (1977). *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.

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. *Cognition*, *93*, 157–165.

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

Lightfoot, D. (1982a). *The Language Lottery: Toward a Biology of Grammars*. Cambridge: MIT Press.

Lightfoot, D. (1982b). 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.

Longobardi, G. (2003). The Structure of DPs: Some Principles, Parameters, and Problems. In M. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory.* Oxford, UK: Blackwell.

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

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.

Pearl, L., & Lidz, J. (in press). Parameters in Language Acquisition. In K. Grohmann & C. Boeckx (Eds.), *The Cambridge Handbook of Biolinguistics.* Cambridge: Cambridge University Press.

Pearl, L., & Sprouse, J. (in press). Computational Models of Acquisition for Islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Islands Effects.* Cambridge: Cambridge University Press.

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.

Ramsey, W., & Stich, S. (1991). Connectionism and three levels of nativism. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and Connectionist Theory.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

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.

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.

Viau, J., & Lidz, J. (2011). Selective learning in the acquisition of Kannada ditransitives. *Language*.

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

Yang, C. (2011). Computational models of syntactic acquisition. *WIREs Cognitive Science*.

A Frequency of different pronouns in the input

Since the P&M learner uses all informative referential pronoun data, we included all available referential personal pronouns in our corpus analysis instead of focusing only on anaphoric one. Table 8 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 6.2).

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 data intake.

Table 8: Pronoun frequencies in the Brown-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%

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 included all four data types. Table 9 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 NP and <NP 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.

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 cor-

Table 9: 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%

relates 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 ambiguous and Syn ambiguous data, we find that the Syn ambiguous data points outnumber the Sem-Syn ambiguous data points in both corpus analyses. The main difference is that LWF found a higher ratio (about 21 Syn to 1 Sem-Syn) than we did (about 11 Syn to 1 Sem-Syn).

For the Unamb NP data in our analysis, we find that such data are fairly similar in quantity to the Syn ambiguous data in our analysis (about 11 Unamb NP data points for every 10 Syn ambiguous 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 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 10 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 7 because that learner does not heed Sem-Syn ambiguous data and so cannot treat them as if they were Unambiguous <NP data.

Table 10: Probabilities after learning, assuming all Sem-Syn ambiguous data are as effective as Unambiguous <NP data.

Prob	**Baker**	**R&G, P&L's filtered**	**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}	0.53 (<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 7 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 target 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.