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

April 28, 2013

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 not recognized that INDIRECT POSITIVE evidence may be both available and useful, in addition to other types of evidence. As a case study, we investigate whether a probabilistic learner utilizing indirect positive evidence can solve the induction problem traditionally associated with English anaphoric *one*. Though full adult knowledge of the representation and interpretation of anaphoric *one* is extensive, recent experimental work with young children has provided an empirical basis to focus on that accords with the original characterization of this induction problem. In particular, given a specific anaphoric *one* context, an adult interpretation appears to be generated by toddlers, based on their preferential looking behavior. We find that our modeled learner, given realistic input, can reproduce this toddler anaphoric *one* behavior, which was thought to indicate the induction problem had been solved by this age. However, we find this behavior can be generated even when a non-adult representation underlies it. This suggests that the previous characterization of this anaphoric *one* induction problem may need revision, as the link between observable behavior and underlying knowledge is not straightforward. Nonetheless, at least some aspects of the adult representation are in place, and the question remains as to how even that level of knowledge is attained so early in development. We discuss the nature of the learning biases that can generate child behavior, and how this impacts the larger debate about the motivation for UG and its contents.

Keywords: anaphoric *one*; acquisition; computational modeling; indirect positive evidence; online probabilistic learning; poverty of the stimulus; Universal Grammar

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; Lightfoot, 1982b; Anderson & Lightfoot, 2000, 2002; Crain & Pietroski, 2002; Fodor & Crowther, 2002; Legate & Yang, 2002; Lidz, Waxman, & Freedman, 2003; Lidz & Waxman, 2004; Sugisaki, 2005; Gualmini, 2007; Yang, 2004; Gualmini, 2007; Berwick, Pietroski, Yankama, & Chomsky, 2011; Yang, 2011; Crain & Thornton, 2012; Anderson, in press). In particular, UG is meant to be one or more learning biases that are part of our biological endowment (INNATE) and are only used for learning language (DOMAIN-SPECIFIC). These learning biases allow children to solve induction problems, a specific kind of learning problem where the available data are compatible with multiple hypotheses about the generalizations for the language.[1] Thus, this motivation for the *existence* of UG comes directly from the existence of induction problems.

Proposals for the *contents* of UG have traditionally come from characterizing a specific learning problem pertaining to a particular linguistic phenomenon, identifying it as an induction problem, and describing the (UG) solution to that specific characterization. Some examples of this approach include investigating linguistic phenomena such as these:

(i) structure-dependent rules to relate the declarative and interrogative forms of utterances (Chomsky, 1980a; Fodor & Crowther, 2002; Legate & Yang, 2002; Berwick et al., 2011; Anderson, in press)

(ii) structure-dependent rules for interpreting pronouns (Anderson & Lightfoot, 2000)

(iii) the structure and interpretation of English anaphoric *one* (Baker, 1978; Ramsey & Stich, 1991; Lidz et al., 2003; Lidz & Waxman, 2004; Regier & Gahl, 2004; Sugisaki, 2005; Gualmini, 2007; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Pearl & Lidz, 2009; Pearl & Mis, 2011)

(iv) constraints on long-distance dependencies (Chomsky, 1973; Yang, 2004, 2011; Crain & Thornton, 2012; Pearl & Sprouse, 2013, in press)

(v) rules for interpreting scope relationships when two or more logical operators are present (Crain & Pietroski, 2002; Crain & Thornton, 2012)

A specific characterization of a learning problem makes it possible to identify it as an induction problem and precisely describe a potential solution to it. More importantly, a specific characterization allows us to explicitly test that solution and compare it to other potential solutions. When the solutions all involve UG biases, this both supports the existence of UG and provides specific proposals for its contents. If it instead turns out that some solutions do not involve UG biases, this takes away the support for UG that comes from that characterization of the learning problem.

Our goals in this article are (i) to characterize the learning problem traditionally associated with English anaphoric *one*, as it has been used to motivate both the existence and contents of UG, and (ii) describe and test a potential solution that draws on a type of evidence not previously used

[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).

for this induction problem (INDIRECT POSITIVE EVIDENCE, discussed below in section 1.2). Our methodology is straightforward. We first characterize the learning problem by drawing on theoretical and experimental work that describes the knowledge to be attained and the observed behavior of young children demonstrating their knowledge. Then, we describe the evidence available in the input that children can use to learn the required knowledge. At this point, we classify this learning problem as an induction problem, as has been traditionally proposed.

We subsequently discuss several proposed solutions to this induction problem, including a new proposal that uses indirect positive evidence. We test their effectiveness by embedding each solution in a probabilistic learning model that incorporates both syntactic and semantic information, and investigate their ability to reproduce the observed child behavior. The results suggest a somewhat nuanced picture, as it is possible to produce the observed child behavior without having the adult representation of anaphoric *one*. This suggests the learning problem associated with anaphoric *one* likely needs to be redefined to allow for a longer learning period, which may cause it to no longer be classified as an induction problem. This also serves as a general cautionary note about the non-trivial relationship between underlying knowledge and observed behavior. Nonetheless, our modeling results suggest that at least some aspects of the adult representation are in place, and the question remains as to how even that level of knowledge is attained so early in development. Computational modeling results based on realistic input (e.g., Sakas & Fodor, 2001; Sakas & Nishimoto, 2002; Yang, 2002; Sakas, 2003; Regier & Gahl, 2004; Yang, 2004; Legate & Yang, 2007; Foraker et al., 2009; Pearl & Lidz, 2009; Perfors, Tenenbaum, & Regier, 2011; Yang, 2011; Sakas & Fodor, 2012; Legate & Yang, 2013; Pearl & Sprouse, 2013, in press) allow us to make progress on the debate surrounding UG by providing a formal mechanism for exploring whether learning strategies can solve language learning problems – or at least, generate the behavior observed in children. We are then able to compare the types of biases required by each successful strategy, consider whether any are UG, and compare precisely what kinds of UG biases each successful strategy motivates if UG biases are indeed involved.

1.1 Characterizing learning problems

Since the characterization of a learning problem is crucial for providing support to UG and making concrete proposals about its contents, how do we characterize learning problems? We believe a learning 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 and biases 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 and biases children have, such as whether they can track distributional information in the input and what information they are sensitive to. We allow a broad definition of "learning bias", where "bias" simply represents a preference of some kind. Under this view, a learning bias can be about the hypothesis space or about the learning mechanism.[2] An example bias about the hypothesis space might involve viewing the learning problem as a decision between two grammatical categories vs. three. An example bias about the learning mechanism might involve what decision method to use, such as probabilistic inference (e.g., Pearl & Lidz, 2009; Yang, 2011) vs. a random step algorithm (e.g., Gibson & Wexler, 1994; Niyogi & Berwick, 1996; Sakas, 2003).

The DATA INTAKE for a learning problem refers to the data children use for learning (Fodor, 1998), and may be a subset of the available input. This is often 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 information children use is defined, corpus analysis methods can provide realistic estimates of the input children encounter.

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 up to that age. 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 identified using these four components: Given a specific initial state, data intake, and learning period, an induction problem occurs when 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, when describing learning strategies for solving an induction problem, these four components can still be used - thus, it is a useful framework for investigating both learning problems and their solutions. For example, the learning strategy may involve particular biases about the data relevant for learning, which are then characterized as part of the learner's initial state and impact the set of data in the data intake. Importantly, once the pieces of a given learning strategy are included in the induction problem characterization, we may find that what previously looked like an induction problem no longer seems to be (e.g., once the learner has additional learning biases, the problem is simply a solvable learning problem instead of an induction problem). Relatedly, it

[2] We note that many proposals about the contents of UG have typically involved biases that define the learner's hypothesis space in some helpful way (e.g., Chomsky, 1973, 1980a; Baker, 1978; Berwick et al., 2011; Crain & Thornton, 2012), rather than biasing the learning mechanism in some helpful way.

is also useful to ask whether a particular learning strategy will be successful for different induction problem characterizations; to the extent that it is, this is stronger support for that learning strategy and the learning biases that comprise it.

1.2 Relaxing 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 (Chomsky, 1980a; Crain, 1991; Hornstein & Lightfoot, 1981; Lightfoot, 1982b; Anderson & Lightfoot, 2000, 2002; Crain & Pietroski, 2002; Legate & Yang, 2002; Lidz et al., 2003; Gualmini, 2007; Crain & Thornton, 2012; Anderson, in press). The basic intuition of the direct evidence assumption is that in order to learn some linguistic knowledge L, children learn from examples of L appearing in the linguistic input (DIRECT POSITIVE EVIDENCE). More recently, several computational induction problem investigations have considered statistical learners that are sensitive to INDIRECT NEGATIVE EVIDENCE related to the directly informative data, and so notice what direct evidence examples are missing from the input (e.g. Regier & Gahl, 2004; Foraker et al., 2009; Perfors, Tenenbaum, & Wonnacott, 2010; Perfors et al., 2011).[3] Thus, prior investigations have typically assumed that the learner's data intake consists of direct positive evidence and (potentially) indirect negative evidence for L.

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 $\_\_{is}$ happy?
(b) *Is the boy who $\_\_{is}$ in the corner is happy?[4]

As another example, when learning the representation and interpretation 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.

As a third example, 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 *wh*-question examples like (3d).

(3) Complex *wh*-question examples
(a) Who $\_\_{who}$ thinks the necklace is expensive?

[3] Note that this is *indirect* negative evidence as the ungrammatical examples are not explicitly presented as ungrammatical, but rather are simply missing from the input.

[4] The * will be used to indicate ungrammaticality henceforth.

(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 – for example, such data might be about linguistic knowledge L2 when the learner is trying to learn about knowledge L1. Nonetheless, these L2 data can be informative about L1 if viewed as relevant by children (for example, due to children's learning biases in the initial state). Thus, these data are examples explicitly observed in the input – making them POSITIVE evidence – but not directly about the specific linguistic knowledge of interest – making them INDIRECT evidence. If the learner's initial state includes knowledge of what counts as indirect positive evidence, the learning problem characterization changes, and may be solvable using different learning strategies than the ones previously proposed.[5] Recently, some computational modeling studies have been exploring the utility of indirect positive evidence for different induction problems, either using it implicitly (Reali & Christiansen, 2005; Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2008; Foraker et al., 2009; Perfors et al., 2011) or explicitly (Pearl & Mis, 2011; Pearl & Sprouse, 2013, in press), with varying results. We follow this approach to investigate the acquisition of anaphoric *one* knowledge.

1.3 Case study: English anaphoric *one*

A specific characterization of an induction problem concerning English anaphoric *one* (from example (2) above) has been vigorously debated recently (e.g., Pullum and Scholz (2002); Lidz et al. (2003); Akhtar, Callanan, Pullum, and Scholz (2004); Lidz and Waxman (2004); Regier and Gahl (2004); Tomasello (2004); Sugisaki (2005); Gualmini (2007); Pearl (2007); Foraker et al. (2009); Pearl and Lidz (2009), among others). Computational modeling studies have examined this characterization 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 (Regier & Gahl, 2004; Pearl & Lidz, 2009). More specifically, each study has investigated the impact of broadening the set of direct positive and indirect negative evidence a learner could use. In the current study, we investigate a learning strategy that broadens the data intake further to include indirect positive evidence.

The rest of this paper is organized as follows. Section 2 briefly reviews the characterization of the induction problem under consideration, including the adult knowledge indicative of the target state and the child behavior thought to specify the learning period. Section 3 highlights why anaphoric *one* has been considered an induction problem, given the available direct positive evidence used for the data intake. Section 4 reviews previous proposals for learning strategies

[5] Notably, 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 (e.g. Chomsky, 1981; Viau & Lidz, 2011; Pearl & Lidz, 2013). The knowledge of the linguistic parameter is part of the initial state, and allows a broader set of data to be utilized for the data intake.

that solve this induction problem, and describes a new learning strategy that additionally uses indirect positive evidence. Section 5 reviews the different kinds of information that are available when understanding referential pronoun data points and presents an online probabilistic learning framework adapted from Pearl and Lidz (2009) that incorporates these different information types, which we then use to compare the different learning strategies.

We show in section 6 that a learner using the indirect positive evidence strategy reproduces the child behavior associated with adult knowledge of *one*. However, 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 have a non-adult representation for *one*, and so the learning period characterizing this induction problem should actually be longer. Because this in turn may change the learner's knowledge state (as older children have acquired more linguistic knowledge) and the data intake the learner uses, this may affect this learning problem's status as an induction problem.

In section 6, we also replicate results found with the previously proposed learning strategies, which suggests that it is the learner's view of the data intake that yields the new results we find, rather than simply something about the specific probabilistic learning framework chosen. Section 7 discusses the nature of the learning biases comprising this learner's strategy for generating toddler anaphoric *one* behavior, and more generally what children would require in order to solve the anaphoric *one* learning problem. Section 7 also discusses alternate characterizations of the learning problem that change the target state and/or the initial state. Section 8 concludes with how these results impact the larger debate about the motivation for UG and its contents.

2 Characterizing the anaphoric one learning 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; Hornstein & Lightfoot, 1981; Lightfoot, 1982a, 1989; Crain, 1991; Ramsey & Stich, 1991; Lidz et al., 2003; Lidz & Waxman, 2004; Sugisaki, 2005; Gualmini, 2007; Foraker et al., 2009; Pearl & Lidz, 2009; Pearl & Mis, 2011). 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* = the second RED BOTTLE present

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 effectively 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!***[6] Then, the listener uses this antecedent to identify the referent of *one*, e.g., what thing in the world *one* is referring to, and what properties that referent has. This is the semantic component. Given the antecedent *red bottle*, adults interpret the referent of *one* as a bottle that is red (RED BOTTLE), as opposed to just any bottle (BOTTLE). That is, the *one* the speaker is referring to is a bottle that specifically has the property RED and this utterance would sound somewhat strange if the speaker actually was referring to a purple bottle.

An influential theoretical framework has 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*]]]]

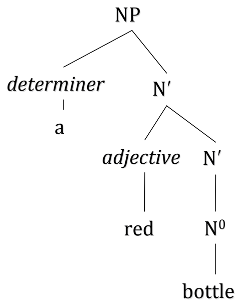

Figure 1: Phrase structure tree corresponding to the bracket notation in examples (5) and (6).

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 modifier+noun 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′).[7]

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

[6] 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 that *one*'s antecedent is *red bottle*. We discuss non-default interpretations in Appendix F.

[7] We note that while we use the labels N′ and N^0, other theoretical implementations may use different labels to distinguish these hierarchical levels. The actual labels themselves are immaterial – it is only relevant for our purposes that these levels are distinguished the way we have done here, i.e., that *red bottle* and *bottle* are the same label (N′ here), while *bottle* can also be labeled with a smaller category label (N^0 here). However, see discussion in section 7.3 for what happens with alternate theoretical representations that additionally differentiate *red bottle* from *bottle*.

This theoretical framework also posits 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).[8] Using an intermodal preferential looking paradigm (Spelke, 1979; Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987), 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.[9] LWF interpreted this to mean that by 18 months, children have acquired the same representation for anaphoric *one* that adults have. This specifies the learning period for anaphoric *one*.

In terms of the learner's initial state, the original proposal by Baker (1978) assumed that only direct evidence was relevant, and that only unambiguous data were informative. LWF's corpus analysis of child-directed speech samples (as well as our own corpus analysis, discussed in section 5.2.2) verified that these data were indeed too sparse to reach that target state within the specified learning period. In particular, LWF 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 (Legate & Yang, 2002; Yang, 2004, 2011). More strikingly,

[8] Though see Tomasello (2004) and Gualmini (2007) for critiques of LWF's interpretation of their experiment and Lidz and Waxman (2004) for a rebuttal to Tomasello (2004).

[9] 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?* - in the control case, they looked at the non-red bottle, showing a default novelty preference. This 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*).

our own corpus analysis found NO examples of these kind of unambiguous data.

The induction problem for anaphoric *one*[10] 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.

(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* like (9) are very rare because they require a 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 isn't 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 referent is 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 to 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 referent 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

[10] For ease of exposition, when we refer to knowledge of "anaphoric *one*" henceforth, we will mean knowledge of anaphoric *one* in examples such as (2), (4), and (8).

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 so 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 Learning strategies

4.1 Previous solutions to the induction problem

4.1.1 Adding additional knowledge to the initial state

The solution proposed by Baker (1978) 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.[11] 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.

Notably, while this does alter the initial state, it does not alter the learner's data intake, which is still restricted to direct, unambiguous data (it is a DIRECTUNAMB learner). 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

[11] 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*$]$.

state of the learner (e.g., that anaphoric elements take linguistic antecedents of the same category). Whether that other knowledge must be innate or could instead be derived from prior experience with language is unclear. More importantly, that was not relevant to the debate concerning the solution to this characterization – even if that other initial state knowledge was necessarily innate (which may or may not be true), the induction problem STILL exists, and one solution is the UG knowledge that Baker proposed.

4.1.2 Updating the initial state in other ways

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 adult representation for anaphoric *one* quite quickly. 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 probabilistic inference.

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

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. Thus, the EO Bayesian learner characterized by R&G's learning strategy would learn from unambiguous data, Sem-Syn ambiguous data, and Syn ambiguous data (this is a DIRECTEO learner). 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 is a DIRECTFILTERED learner). 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 probabilistic inference.
+Bias: Only unambiguous and Sem-Syn ambiguous 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.[12] 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.

4.2 Another solution: Removing the direct evidence bias

Instead of restricting the data intake, 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. – i.e., 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)).[13] However, since *one* is similar to other pronouns referentially (by being anaphoric and

[12] 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).

[13] 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).

having linguistic antecedents) and shares some syntactic distribution properties with them (since it can appear as an NP), a learner could decide that information gleaned from other pronouns is relevant for interpreting *one*. These "other pronoun" data would then become indirect positive evidence for the learner trying to acquire the representation for anaphoric *one* (a +OTHERPRO learner), since the learner is leveraging information from the presence of these data.

This bias to use other pronoun data can be combined with a bias to use probabilistic inference, similar to R&G's and P&L's learners. In particular, a learner could track how often the intended referent has a property mentioned in the potential antecedent (e.g., *red* in *a red bottle* in (14)), which relates to whether this property should be included in the antecedent. 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 (as in (12) and (14)), 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 the +OtherPro learner will learn from. The initial state for the +OtherPro learning strategy is thus updated as follows:

+OTHERPRO'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 probabilistic inference.
+Bias: Learn from other pronoun data.

Like the R&G and P&L learning strategies, the +OtherPro learning strategy differs from Baker's (DirectUnamb) strategy by learning from data besides the unambiguous <NP data. However, the +OtherPro strategy also differs from the strategies in R&G and P&L (DirectFiltered, DirectEO) by learning from data containing anaphoric elements besides *one*, since these are viewed as indirect positive evidence. Table 1 shows which learning strategies use which data.

We will save detailed discussion of the nature of the biases involved in the +OtherPro learning strategy for section 7.2, specifically the bias to learn from other pronoun data. If this turns out to be 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 a(nother) strategy for solving this induction problem that does not require a UG learning bias.

Table 1: Data intake for different learning strategies.

Data type	Example	Learning strategies
Unamb <NP	*Look– a red bottle! There isn't another one here, though.*	DirectUnamb, DirectFiltered, DirectEO, +OtherPro
Sem-Syn amb	*Look– a red bottle!* *Oh, look– another one!*	DirectFiltered, DirectEO, +OtherPro
Syn amb	*Look– a bottle!* *Oh, look– another one!*	DirectEO, +OtherPro
Unamb NP	*Look a red bottle!* *I want it/one.*	+OtherPro

5 Learning anaphoric one

5.1 Information in the data

There is a variety of information that a learner uses to understand a referential expression in an anaphoric data point, some of which is observable and some of which is latent, as shown in Figure 2 (observed variables are shaded). This figure represents the information used by the learner when understanding a referential expression containing a pronoun. Notably, both syntactic and referential information can be used by the learner, as will become clear when we step through the different variables involved. All variables in this model are discrete, with variables that are binary[14] in lowercase.

Beginning at the top lefthand side of Figure 2, **R** is the referential phrase itself, i.e., the words used in the referential expression, such as *another one* or *it*. This is observable from the data point, and from this, the learner can observe the pronoun used in the referential expression (**Pro**), e.g., *one* or *it*. In addition, from **R**, the learner can observe the syntactic environment (**env**) of the referential pronoun. In particular, the learner can observe whether the pronoun is used in an environment that indicates it is smaller than a noun phrase (**env=<NP**), such as *another one* or instead is in an environment that indicates it is a noun phrase (**env=NP**), such as *it*. The values of **Pro** and **env** are used to infer the syntactic category (**C**) of the referential pronoun, which can be N^0, N', or NP. The learner assumes the syntactic category of the pronoun is the same as the syntactic category of the linguistic antecedent, and so uses the syntactic category information from **C** to determine two properties of the linguistic antecedent: (1) if the antecedent includes a determiner (**det=yes**) or not (**det=no**), and (2) if the antecedent includes a modifier (**mod=yes**) or not (**mod=no**). If **C=NP**, both a determiner and modifier must be included if present (**det=yes**, **mod=yes**); if **C=N′**, a determiner is not possible (**det=no**) though a modifier is and so may either be included (**mod=yes**) or not (**mod=no**); if **C=N^0**, neither a determiner nor a modifier is possible (**det=no**, **mod=no**). All of these variables depend on the syntactic information available from the data point.

Moving to the top righthand side of Figure 2, **m** concerns whether a property was mentioned

[14] Some of these variables take on an additional value of "not applicable" in certain cases, but otherwise take on one of only two values. See Table 2 for details.

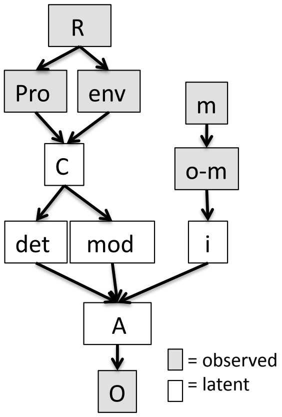

Figure 2: Model of understanding a referential expression containing a pronoun. The variables correspond to (i) **syntactic** information (R = the referential phrase, Pro = the pronoun in the referential phrase, env = the syntactic environment of the pronoun, C = the syntactic category of the pronoun, det = whether a determiner is included in the antecedent, mod = whether a modifier is included in the antecedent), (ii) **referential** information (m = whether a property was mentioned in the potential linguistic antecedent, o-m = whether an object in the present context has the mentioned property, i = whether the mentioned property is included in the antecedent), (iii) the **linguistic antecedent** (A = the linguistic antecedent), and the (iv) **intended referent** (O = the intended referent, which is an object in our examples).

in the potential linguistic antecedent or not, e.g., *Look – a red bottle* (**m=yes**) vs. *Look – a bottle* (**m=no**). If a property is mentioned, **o-m** concerns whether a referent (object) in the present context has the mentioned property (**o-m=yes**) or not (**o-m=no**). Both these variables' values can be observed from the previous linguistic context (**m**) and the current environment (**o-m**). If an object in the present context has the mentioned property (**o-m=yes**), the learner will infer whether the property should be included in the linguistic antecedent (**i=yes**) or not (**i=no**), which concerns the speaker's intentions (specifically, did the speaker intend to refer to that property when identifying the referent?) All these variables can be thought of as concerning the referential intentions of the speaker.

Both syntactic information (**det**, **mod**) and referential information (**i**) are used to infer the linguistic antecedent (**A**) of the referential pronoun, e.g., *red bottle* vs. *bottle*. Only certain combinations of variable values are licit when a property is mentioned (**m=yes**), due to the constraints placed on the antecedent by **mod** and **i**:[15]

[15] In particular, **mod** and **i** must agree. If **mod=no** and **i=yes**, the referential intent is to include the mentioned property in the antecedent (**i=yes**), but there is no place syntactically for the property to go, as no modifier is possible (**mod=no**) – as is the case for category N^0. If **mod=yes** and **i=no**, the referential intent is not to include the property (**i=no**), but the syntax requires a modifier to be present (**mod=yes**) – and this is impossible as no property can fill the modifier slot.

In addition, if **mod** (and so **i**) = **no**, **det** $\neq$ **yes** since including a determiner (**det=yes**) necessarily includes any modifier present (requiring **mod=yes**) due to the structure of NPs (see Figure 1).

(a) **det=yes**, **mod=yes**, **i=yes** yielding e.g., **A=** ***the red bottle***
(b) **det=no**, **mod=yes**, **i=yes** yielding e.g., **A=** ***red bottle***
(c) **det=no**, **mod=no**, **i=no** yielding e.g., **A=** ***bottle***

The antecedent is used to infer the intended object (**O**). Notably, despite this depending on the linguistic antecedent **A**, the actual intended referent is often observable from context (as is the case for our various data types discussed above). That is, the learner can infer what object was the intended referent, even if the linguistic antecedent is ambiguous. For example, consider Sem-Syn ambiguous data, such as *Look – a red bottle! Oh, look – another one!* The issue with such data is that it is unclear whether the antecedent is *red bottle* or *bottle* since both are compatible with the object present (e.g., a RED BOTTLE). Thus, though the intended referent depends on the latent variable **A**, the learner can often observe what properties the intended object **O** has, e.g., whether it is a RED BOTTLE or a PURPLE BOTTLE, etc.

These variables can take on the values shown in table 2.[16] The data types used by the different learning proposals have the observable and latent values in Table 3. We now walk through the variable values for each of the four data types.

Table 2: Variable values in informative referential data points. Observable variables are in **bold**.

R ∈ {*another one*, *it*, etc.}	
Pro ∈ {*one*, *it*, etc.}	**m** ∈ {yes, no}
env ∈ {<NP, NP}	**o-m** ∈ {yes, no, *N/A*}
C ∈ {NP, N′, N^0}	i ∈ {yes, no, *N/A*}
det ∈ {yes, no}	
mod ∈ {yes, no}	
A ∈ {*a red bottle*, *red bottle*, *bottle*, etc.}	
O ∈ {RED BOTTLE, PURPLE BOTTLE, etc.}	

Unambiguous <NP data have a referential expression **R** such as *another one*, which uses the pronoun *one* (**Pro=*one***) and indicates the pronoun is smaller than an NP (**env=<NP**). In addition, a property is mentioned in the potential linguistic antecedent (**m=yes**) and an object in the present context has the mentioned property (**o-m=yes**) – in particular, the intended referent has the mentioned property (e.g, **O=RED BOTTLE**). Because these data are unambiguous, the learner can infer the antecedent **A** (e.g., *red bottle*), which indicates that the property is included in the antecedent (**i=yes**) on the referential side, while a modifier is included in the antecedent (**mod=yes**) and a determiner is not included (**det=no**) on the syntactic side. Given that a modifier is included, the category **C** must be N′.

Similar to Unambiguous <NP data, Sem-Syn ambiguous data have a referential expression **R** such as *another one*, which uses the pronoun *one* (**Pro=*one***) and indicates the pronoun is smaller than an NP (**env=<NP**). In addition, a property is mentioned in the potential linguistic antecedent (**m=yes**) and an object in the present context has the mentioned property (**o-m=yes**) – in particular,

[16] Note that if no property was mentioned (**m=no**), the decision as to whether an object present has the mentioned property is moot (**o-m=N/A**), as is the decision to include the mentioned property in the antecedent (**i=N/A**).

Table 3: Data types and variable values. Observable variables are in **bold**.

Variable	**Unamb <NP**	**Sem-Syn Amb**	**Syn Amb**	**Unamb NP**
R	ex: *another one*	ex: *another one*	ex: *another one*	ex: *it*
Pro	*one*	*one*	*one*	ex: *it*
env	<NP	<NP	<NP	NP
C	N'	N', N^0	N', N^0	NP
det	no	no	no	yes
mod	yes	yes, no	no	yes
m	yes	yes	no	yes
o-m	yes	yes	*N/A*	yes
i	yes	yes, no	*N/A*	yes
A	ex: *red bottle*	ex: *red bottle*, *bottle*	ex: *bottle*	ex: *a red bottle*
O	ex: RED BOTTLE	ex: RED BOTTLE	ex: BOTTLE	ex: RED BOTTLE

the intended referent has the mentioned property (e.g, **O=RED BOTTLE**). However, because these data are ambiguous, it is unclear whether the antecedent **A** includes the mentioned property as a modifier or not (e.g., *red bottle* vs. *bottle*). Thus, while it is clear the determiner is not included (**det=no**), it is unclear whether the mentioned property is included in the modifier position (**i=yes, no**, **mod=yes, no**). Because of this, it is also unclear whether the syntactic category **C** is N' or N^0.

Like Unambiguous <NP and Sem-Syn ambiguous data, Syn ambiguous data have a referential expression **R** such as *another one*, which uses the pronoun *one* (**Pro=*one***) and indicates the pronoun is smaller than an NP (**env=<NP**). However, a property is not mentioned in the potential linguistic antecedent (**m=no**) and so it is moot whether an object in the present context has the mentioned property (**o-m=*N/A***) – in particular, it does not matter what properties the intended referent has (e.g, **O=BOTTLE**). Nonetheless, given the nature of these data, the learner can infer the antecedent **A** (e.g., *bottle*), which indicates that no determiner or modifier is included in the antecedent (**det=no**, **mod=no**). Because no property was mentioned, it is moot whether the mentioned property is included in the antecedent (**i=*N/A***). Nonetheless, it is unclear from the antecedent whether the category **C** is N' or N^0.

Unambiguous NP data have a referential expression **R** such as *it*, which uses a pronoun such as *it* (**Pro=*it***) and indicates the pronoun is category NP (**env=NP, C=NP**). In addition, a property is mentioned in the potential linguistic antecedent (**m=yes**) and an object in the present context has the mentioned property (**o-m=yes**) – in particular, the intended referent has the mentioned property (e.g, **O=RED BOTTLE**). Because these data are unambiguous, the learner can infer the antecedent **A** (e.g., *a red bottle*), which indicates that the property is included in the antecedent (**i=yes**) on the referential side, while a modifier and determiner are included in the antecedent (**mod=yes**, **det=yes**) on the syntactic side.

5.2 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. We will use this framework to evaluate the different proposed learning strategies.

5.2.1 Important quantities

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

(a) when an object has the property mentioned in the potential antecedent (**o-m=yes**), that property is included in the antecedent of *one* (**i=yes**), and

(b) when the syntactic environment indicates *one* is smaller than an NP (**env=<NP**), it is category N′ (**C=N′**).

Importantly for the update equations we will use in the online probabilistic learning framework, the variables of interest (**i** and **C**) can only take on two values in these situations: **i** $\in$ {yes, no} when **o-m=yes** and **C** $\in$ {N′, N^0} when **env=<NP**. Our modeled learner will determine the probability associated with a particular value for both these variables in these situations, specifically *p*(*i=yes* | *o-m=yes*) and *p*(*C=N′* | *env=<NP*). We represent the probability of the former as p_{incl} and the probability of the latter as $p_{N'}$. If the target representation of *one* has been learned for the default context, both probabilities should be near 1.[17]

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_{incl} and $p_{N'}$ as parameters in binomial distributions, given that each variable takes on only two values, as noted above.

$$p_x = \frac{\alpha + d_x}{\alpha + \beta + D_x}, \alpha = \beta = 1 \qquad (15)$$

Parameters α and β represent a very weak prior when set to 1.[18] The variable d_x represents how many informative data points indicative of x have been observed, while D_x represents the total number of potential x data points observed. After every informative data point, d_x and D_x are updated as in (16), 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, ϕ_x = 1; for ambiguous data $\phi_x < 1$.

[17] Note that there may be different strategies for translating probabilities about underlying knowledge into actual behavior in any given anaphoric *one* situation. For example, one strategy is where p = the chance of selecting that underlying representation, so p = 0.60 represents a stronger chance of selecting that underlying representation, compared to p = 0.51. Another might be a "winner-take-all" strategy where any $p >$ 0.50 is equivalent to 1 and any $p <$ 0.50 is equivalent to 0. Since it is unknown how this process works, we assume (following Regier and Gahl (2004) and Pearl and Lidz (2009)) the most direct translation, where the higher the probability, the higher the chance of selecting this underlying representation, as this does not require a commitment to any additional filtering process on the part of the learner.

[18] Before seeing any data at all, the learner effectively imagines that one data point has been observed in favor of one value of the variable (α=1) and one data point has been observed in favor of the other value of the variable (β=1). These numbers are quickly overwhelmed by actual observations of data.

$$d_x = d_x + \phi_x \tag{16a}$$
$$D_x = D_x + 1 \tag{16b}$$

Probability p_{incl} 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_{incl} since **o-m=*N/A***). 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 (**env=NP**), and so are uninformative for $p_{N'}$).

The value of ϕ_x depends on data type. We can derive the values of ϕ_{incl} and $\phi_{N'}$ by doing probabilistic inference over the graphical model in Figure 2. The details of this inference are presented in Appendix A. Both ϕ_{incl} and $\phi_{N'}$ involve three free parameters: m, n, and s. Two of these, m and n, correspond to syntactic information: They 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*). We will follow the corpus-based estimates P&L used for m and n, which are m = 1 and n = 2.9.[19] The other parameter, s, corresponds to referential information: It indicates how many salient properties there are in the learner's hypothesis space at the time the data point is observed. This determines how suspicious a coincidence it is that the object just happens to have the mentioned property, given that there are s salient properties the learner is aware of. It is unclear how best to empirically ground our estimate as it concerns what is salient to the child, which is not easily observable from existing empirical data. It may be that a child is only aware of a few salient properties out of all the properties known (e.g., PURPLE and IN MOMMY'S HAND for a purple bottle in Mommy's hand). In contrast, it may be that the child considers all known properties, which we can conservatively estimate as the number of adjectives known by this age (e.g., P&L estimate 14- to 16-month-olds know approximately 49 adjectives, using the MacArthur CDI (Dale & Fenson, 1996)). We use s=10 in the simulations reported in section 6, but also explore a variety of values ranging from 2 to 49 in Appendix E.

Table 4 shows a sample update after a single data point of each type at the beginning of learning when p_{incl} = $p_{N'}$ = 0.50, using the values $m = 1$, $n = 2.9$, and $s = 10$.

For unambiguous <NP data, both ϕ_{incl} and $\phi_{N'}$ are 1, and so d_x is increased by 1. This leads to p_{incl} and $p_{N'}$ both being increased. This is intuitively satisfying since unambiguous <NP data by definition are informative about both p_{incl} (the mentioned property should indeed be included in the antecedent) and $p_{N'}$ (the syntactic category is indeed N′).

For Sem-Syn ambiguous data, both p_{incl} 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_{incl} and $p_{N'}$ (which are both 0.50 initially). After one Sem-Syn Amb data point, p_{incl} increases to 0.53, and $p_{N'}$ increases to 0.59. This is again intuitively satisfying since the learner capitalizes on the suspicious coincidence that the intended object has the mentioned property, but is not as confident in this data point as the learner would be about an unambiguous <NP data point.

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

Table 4: The value of p_{incl} and $p_{N'}$ after one data point is seen at the beginning of learning when p_{incl} = $p_{N'}$ = 0.50, $\alpha = \beta$ = 1, $m = 1$, $n = 2.9$, and $s = 10$.

	$p_x = \frac{\alpha+d_x}{\alpha+\beta+D_x},\ \alpha = \beta = 1$	
Data type	p_{incl}	$p_{N'}$
Unamb <NP	0.67	0.67
Sem-Syn Amb	0.53	0.59
Syn Amb	0.50	0.48
Unamb NP	0.67	0.50

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 data that P&L discovered, where these data cause the learner to believe less that *one* is category N' when it is smaller than an NP. The formal details of why this occurs are described in Appendix B.

Unambiguous NP data are only informative with respect to whether the mentioned property is included in the antecedent, so only p_{incl} is updated and only ϕ_{incl} has a value. Since these data are unambiguous, ϕ_{incl}=1, which is intuitively satisfying. This leads to an increase in p_{incl}.

5.2.2 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.[20]

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

[20] See Appendix C for a more thorough breakdown of the corpus analysis we have conducted here.

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 included in the antecedent. If no property is mentioned, then the data point is uninformative as to whether the antecedent must contain 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.[21] 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* learning strategies

Data type	**DirectUnamb**	**DirectFiltered**	**DirectEO**	**+OtherPro**
Unamb <NP	0	0	0	0
Sem-Syn Amb	0	242	242	242
Syn Amb	0	0	2743	2743
Unamb NP	0	0	0	3073
Uninformative	36500	36258	33515	30442

5.2.3 Measures of success

One way to assess acquisition success is to measure p_{incl} 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. As an additional checkpoint, we can also assess how likely a learner would be to reproduce the observed child 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), given a choice between that bottle and a bottle of a different color? Notably, this is a metric that the previous studies by R&G and P&L did not use, as they were only assessing the probability of the target representation where p_{incl}=$p_{N'}$=1. However, given that we have empirical data about children's behavior, it seems reasonable to also use it in assessing acquisition success if we can.

We can use almost the same graphical model shown in Figure 2 to calculate the probability of the learner looking at the referent that has the mentioned property when given a choice between

[21] Specifically, 2,874 of the 17,521 utterances from the Eve corpus were referential data points containing a pronoun, 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.

two referents (p_{beh}). The only difference is that the intended object **O** is no longer an observed variable – instead, the child infers the intended object from the information available and looks to one of the two objects present. Thus, s=2, though the other free parameter values remain the same (m=1, n=2.9). More formally, given a data point that has a referential expression **R=*another one***, a pronoun **Pro=*one***, a syntactic environment that indicates the pronoun is smaller than NP (**env=<NP**), a property mentioned (**m=yes**), and an object in the present context that has that property (**o-m=yes**), we can calculate how probable it is that a learner would look to the object that has the mentioned property (e.g., **O=RED BOTTLE**), which is what 18-month-olds in the LWF experiment did. We describe the formal details of the probabilistic inference involved in calculating p_{beh} in Appendix D.1.

In addition to assessing the probability of the observed 18-month-old behavior in the LWF experiment, we can also assess the assumption LWF made about interpreting their experiment: If children 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. 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_{incl} and $p_{N'}$)? To answer this question, we can simply look at p_{beh} compared to p_{incl} and $p_{N'}$. If p_{beh} is high when either p_{incl} or $p_{N'}$ is low, this suggests that adult-like behavior does 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 that the learner has the target representation, given that the learner has produced the adult behavior (e.g., looking at the object with the mentioned property) in the experiment ($p_{rep|beh}$). This is, in effect, the contextually-constrained representation the learner is using, where the context is defined as the experimental setup. More formally, given that the referential expression is *another one* (**R=*another one***), the pronoun is *one* (**Pro=*one***), the syntactic environment indicates the pronoun is smaller than an NP (**env=<NP**), a property was mentioned (**m=yes**), an object present has the mentioned property (**o-m=yes**), AND the child has looked at the object with the mentioned property (e.g., **O=RED BOTTLE**), what is the probability that the representation is the adult representation, where the antecedent = e.g., *red bottle* (**A=*red bottle***)? This would mean that the antecedent includes the property (**i=yes**), the antecedent does not include the determiner (**det=no**), the antecedent includes a modifier (**mod=yes**), and the antecedent category is N′ (**C=N′**). Probability $p_{rep|beh}$ can be calculated by using probabilistic inference over the graphical model in Figure 2. The formal details of the probabilistic inference involved in calculating $p_{rep|beh}$ are discussed in Appendix D.2.

6 Results

Table 7 shows the results of the learning simulations over the different input sets with s (the number of properties salient to the learner when interpreting the data point) set to 10.[22] Averages over 1000 runs reported, with standard deviations in parentheses.[23]

Table 7: Probabilities after learning, with s=10, which is the number of properties salient to the learner when interpreting a data point.

Probability	**DirectUnamb**	**DirectFiltered**	**DirectEO**	**+OtherPro**
$p_{N'}$	0.50 (<0.01)	0.99 (<0.01)	0.25 (0.06)	0.37 (0.04)
p_{incl}	0.50 (<0.01)	0.96 (<0.01)	0.38 (0.18)	>0.99 (<0.01)
p_{beh}	0.56 (<0.01)	0.95 (<0.01)	0.53 (0.04)	>0.99 (<0.01)
$p_{rep\|beh}$	0.23 (<0.01)	0.95 (<0.01)	0.11 (0.11)	>0.99 (<0.01)

A few observations can be made. First, since the DirectUnamb learner uses only unambiguous <NP data in its intake and these data were not found in our dataset, this learner effectively learns nothing. Thus, the DirectUnamb learner remains completely uncertain whether *one* is N′ when it is smaller than NP ($p_{N'}$=0.5) and whether the antecedent includes the mentioned property (p_{incl}=0.5). Given these general non-preferences, it is only slightly more likely to look at the correct bottle in the LWF experiment (p_{beh}=0.56) and is fairly unlikely to have the adult representation if it happens to do so ($p_{rep|beh}$=0.23). Specifically, if the DirectUnamb learner looks at the bottle with the mentioned property, it has only a 23% of doing so because it has the same antecedent as adults do. Thus, learning from unambiguous <NP data alone runs into an induction problem, as Baker supposed and we have affirmed.

Turning now to the +OtherPro learner, we see that this learner which includes the indirect positive evidence of unambiguous NP data decides that the antecedent should include the mentioned property (p_{incl}>0.99). This seems intuitively satisfying as this probability is exactly what unambiguous NP data boost. However, this learner also has a moderate dispreference for believing *one* is N′ when it is smaller than an NP ($p_{N'}$=0.37). That is, this learner is inclined to incorrectly believe that *one* is category N^0 in general, which is not the target state. This means that, given a Syn ambiguous data point like *Look, a bottle! Do you see another one?*, the +OtherPro learner would interpret *one*'s antecedent as [$_{N^0}$ bottle], rather than as [$_{N'}$ [$_{N^0}$ bottle]].[24] In addition, unlike adults, it would judge utterances like the following grammatical, since they use *one* as an N^0: **I sat by the side of the river, and you sat by the one of the road.*

Interestingly, this lack of the target state knowledge does NOT prevent the +OtherPro learner from producing the observed infant behavior in the LWF experiment (p_{beh}>0.99). How can this

[22] For the results with different values of s, ranging from 2 to 49, see Appendix E.

[23] 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.

[24] Note that the +OtherPro learner would have still have adult-like *behavior* (believing the antecedent string = *bottle*, and so the referent is a BOTTLE).

be? This is due to the linguistic context in the experiment, where a property is mentioned in the potential antecedent. Because the learner believes so strongly that a mentioned property must be included in the antecedent (e.g., the antecedent is *red bottle* rather than *bottle*), the only representation that allows this (e.g., [$_{N'}$ *red* [$_{N'}$ [$_{N^0}$ *bottle*]]]) overpowers the other potential representations' probabilities. Thus, the +OtherPro learner will conclude the antecedent includes the mentioned property, and so it and the pronoun referring to it (*one*) must be N′ IN THIS CONTEXT – even if the learner believes *one* is not N′ in general. 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. This is a somewhat surprising result (though see Gualmini (2007), who supposed this could occur for other reasons).

Nonetheless, 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 ($p_{rep|beh} > 0.99$). As described above, this is due to believing a mentioned property must be included in the antecedent. Thus, LWF's assumption is true in linguistic contexts where a property is mentioned in the potential antecedent, such as their experiment, and any unambiguous <NP, Sem-Syn ambiguous, and unambiguous NP data points. So, in essence, LWF were correct to believe 18-month-olds had the adult representation in their experimental context - it's simply that 18-month-olds may not have the adult representation in all contexts.

A reasonable question is whether this somewhat surprising behavior is due solely to the data intake of the +OtherPro learner, or is instead due to the probabilistic learning model that we have assumed underlies learning and generation of child behavior. One way to evaluate this is by examining the results of the other two learning strategies that have previously been investigated to see how they compare with previous results for those learning strategies.

For the DirectFiltered learner, previous studies (Regier & Gahl, 2004; Pearl & Lidz, 2009) found that this filtered learner has a very high probability of acquiring the adult representation. We replicate this qualitative result here ($p_{N'}$=0.99, p_{incl}=0.96). For the DirectEO learner, Pearl and Lidz (2009) found that this learner has a very low probability of learning the adult representation. We replicate this qualitative result here ($p_{N'}$=0.25, p_{incl}=0.38). Thus, the differences in the data intake used by the learners modeled here are the most likely cause of the different learning behavior we observe.

To summarize, we find that indirect positive evidence coming from Unambiguous NP data has a significant beneficial impact on learning and allows a probabilistic learner to reliably generate the anaphoric *one* behavior observed in 18-month-olds. More surprisingly, we also find that evaluating learners based on their ability to replicate experimental results LWF found with 18-month-olds leads to an unexpected result: Learners WITHOUT the target representation can still produce the target behavior. Specifically, even if a learner does not believe *one* is N′ in general, if that learner believes a mentioned property should be included in the linguistic antecedent, that learner can still generate the target behavior in the LWF experiment. This suggests a more complicated relationship between underlying knowledge and observed behavior. Specifically for anaphoric *one*, it is possible to correctly interpret *one* in certain linguistic contexts and have the target representation in those contexts, even if the target representation is not the preferred default representation.

7 Discussion

7.1 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 generate this situation: It leads 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 induction 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, where *one* is always category N′, 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 7.3.1. 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 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 addition, we have provided a framework for (i) making this characterization and (ii) representing the components of any proposed learning strategy for solving an induction problem and/or generating the observed child behavior. Then, when a learning strategy succeeds, we can examine the learning biases that comprise it and discuss whether they are likely to be in UG. With this idea in mind, we examine the biases that are part of the indirect positive evidence learning strategy used by the +OtherPro learner. In addition, we discuss alternate characterizations of the induction problem that alter the target and/or initial state, and how this impacts the debate about the contents of UG.[25]

[25] For discussion of other learning strategies for the induction problem characterization already explored, see Appendix F.

7.2 The learning biases of the +OtherPro learning strategy

The +OtherPro learning strategy includes two biases that enrich the initial state of the learner:

(a) Use probabilistic inference.
(b) Learn from other pronoun data.

The bias to use probabilistic 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 (the DirectFiltered learner). Since probabilistic 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 **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 a 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 a non-UG learning strategy that will produce the desired behavior, in addition to the potentially non-UG one proposed by R&G/P&L and implemented in the DirectFiltered learner.

7.3 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.

7.3.1 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, probabilistic inference was used. In addition, F&al's learning strategy employed subtle conceptual knowledge in order to identify the 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 3 shows the syntactic structure associated with modifiers and complements, where a modifier like *with dots* is sister to N′ and a complement like *of the road* is sister to N^0.

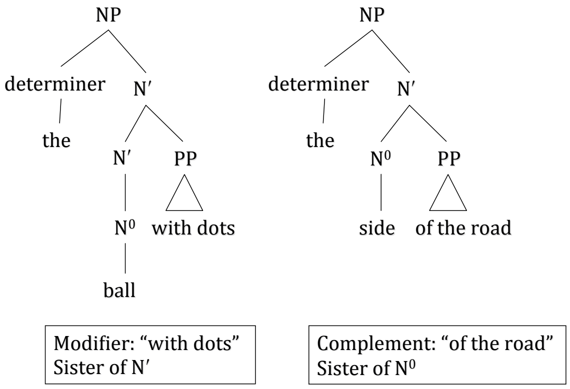

Figure 3: Phrase structure trees corresponding to a modifier and a complement.

Because of this, *one* (being N′) cannot appear with complements, since complements adjoin with N^0. This is why **one of the road* is ungrammatical (19a), while *one with dots* is grammatical (19b).

(19a) *Lily waited by the side of the building while Jack sat by the one of the road.

(19b) 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:

(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 probabilistic 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 (viewing other N^0 elements as informative constitutes indirect positive evidence for this learner). In particular, a Bayesian learner can note the absence of *one* being used with complements (which is indirect negative evidence for this learner). 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 probabilistic inference is likely innate, but also likely domain-general since probabilistic 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 +OtherPro 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. Moreover, the ability to make this rather sophisticated semantic distinction between complements and modifiers may not be available before 18 months, and so children would only be able to use this learning strategy if the learning period for anaphoric *one* is longer.

From this, we see that considering this characterization 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* nor calculate a preference for including a mentioned property in the antecedent, it is unclear how well it would match the behavior of 18-month-olds observed in the LWF experiment, however.

7.3.2 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 4, and shown in bracket notation in (20a). However, an alternate representation of noun phrases is available (Bernstein, 2003; Longobardi, 2003)[26], shown in (20b) and the right side of Figure 4. It assumes the following: (i) noun phrases are category DP (Determiner Phrase), (ii) modifiers are sisters to N′ and children of NP, and (iii) complements are sisters of N′ and children of N′.

(20a) [$_{NP}$ *a* [$_{N'}$ *delicious* [$_{N'}$ [$_{N^0}$ *bottle*] [$_{PP}$ *of wine*]]]]
(20b) [$_{DP}$ *a* [$_{NP}$ *delicious* [$_{N'}$ [$_{N'}$ [$_{N^0}$ *bottle*]] [$_{PP}$ *of wine*]]]]

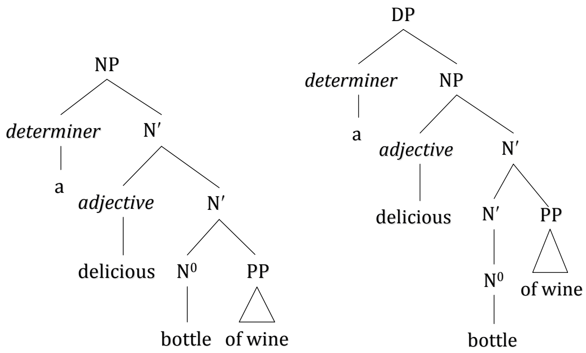

Figure 4: Phrase structure trees corresponding to the bracket notation in examples (20) and (21) for *a delicious bottle of wine*.

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

(21) *I have a delicious bottle of wine...*
(a) *...and you have another one.* [*one* = *delicious bottle of wine*, category NP]
(b) *...and you have a flavorful one.* [*one* = *bottle of wine*, category N′]
(c) *... *and you have a flavorful one of beer.* [*one* $\neq$ *bottle*, category N^0]

This means there are three syntactic categories smaller than an entire noun phrase (DP), and a child must learn that only two of them are valid antecedents for *one*. Moreover, in the LWF experiment, a child should have the preference that *one*'s antecedent is category NP, so that it can include the modifier (i.e., *red bottle* is an NP in this representation).

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

[26] Thanks to Greg Kobele and several anonymous reviewers for noting this.

(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 so 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 +OtherPro strategy proposed here, as there are still many similarities in the induction problem characterization. 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 probabilistic 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 (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 unambiguous 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 included in the antecedent (p_{incl}). 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_{incl} is high. This is again similar to the behavior the +OtherPro 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 for the +OtherPro strategy, and so the implications for UG remain the same as discussed above in section 7.2. 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.

8 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 can be used to specify different components of this induction problem. We then demonstrated that a probabilistic learning strategy using indirect positive evidence that comes from data containing other referential pronouns can produce the behavior observed experimentally in young children – even when the target knowledge state has 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 induction problem where acquisition of anaphoric *one* knowledge proceeds in stages, and the learning period for anaphoric *one* may be longer than previously thought. Nonetheless, to achieve even this immature representation that can generate adult interpretations in certain contexts requires certain learning biases. By characterizing the anaphoric *one* learning problem and the indirect positive evidence learning strategy using our framework, we explicitly identified what those learning biases would be and discussed whether they are likely to be UG biases. More generally, we described how explicit computational models implementing different strategies can be used to offer concrete proposals for the contents of UG. This case study demonstrates that indirect positive evidence does not automatically 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.

9 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 and discussion from LouAnn Gerken, Jeff Lidz, Max Bane, Morgan Sonderegger, Greg Kobele, Ming Xiang, Sue Braunwald, several 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 NYU 2012 Linguistics Colloquium, 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.

Anderson, S. R. (in press). What is Special About the Human Language Faculty, and How Did It Get That Way? In R. Botha & M. Everaert (Eds.), *The Evolutionary Emergence of Human Language.* Oxford, UK: Oxford University Press.

Anderson, S. R., & Lightfoot, D. W. (2000). The Human Language Faculty as an Organ. *Annual Review of Physiology*, *62*, 697–722.

Anderson, S. R., & Lightfoot, D. W. (2002). *The Language Organ: Linguistics as Cognitive Physiology*. Cambridge, UK: Cambridge University Press.

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.

Berwick, R., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*, 1207–1242.

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.

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

Crain, S., & Thornton, R. (2012). Syntax Acquisition. *Wiley Interdisciplinary Reviews Cognitive Science*, *3*, 185–203.

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

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.

Fodor, J. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The Linguistic Review*, *19*, 105–145.

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.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(4), 407–454.

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.

Gualmini, A. (2007). On that One Poverty of the Stimulus Argument. *Nordlyd*, *34*(3), 153–171.

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.

Legate, J., & Yang, C. (2007). Morphosyntactic Learning and the Development of Tense. *Linguistic Acquisition*, *14*(3), 315–344.

Legate, J., & Yang, C. (2013). Assessing Child and Adult Grammar. In R. Berwick & M. Piatelli-Palmarini (Eds.), *Rich Languages from Poor Inputs.* Oxford, UK: Oxford University Press.

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.

Niyogi, P., & Berwick, R. C. (1996). A language learning model or finite parameter spaces. *Cognition*, *61*, 161–193.

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. (2013). Parameters in Language Acquisition. In K. Grohmann & C. Boeckx (Eds.), *The Cambridge Handbook of Biolinguistics* (pp. 129–159). Cambridge, UK: Cambridge University Press.

Pearl, L., & Mis, B. (2011). How Far Can Indirect Evidence Take Us? Anaphoric One Revisited. In L. Carlson, C. Höschler, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 879–884). Austin, TX: Cognitive Science Society.

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*, 19–64.

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.

Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607 - 642.

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.

Sakas, W. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 415–422). Sapporo, Japan: Association for Computational Linguistics.

Sakas, W., & Fodor, J. (2012). Disambiguating Syntactic Triggers. *Language Acquisition*, *19*(2), 83–143.

Sakas, W., & Fodor, J. D. (2001). The Structural Triggers Learner. In S. Bertolo (Ed.), *Language Acquisition and Learnability* (pp. 172–233). Cambridge, UK: Cambridge University Press.

Sakas, W., & Nishimoto, E. (2002). *Search, Structure or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition.* City University of New York, NY. (Manuscript)

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

Sugisaki, K. (2005). *One* issue in acquisition. In *Proceedings of the Sixth Tokyo Conference on Psycholinguistics* (pp. 345–360). Tokyo, Japan: Hituzi Syobo.

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

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 Deriving ϕ_{incl} and $\phi_{N'}$

The values ϕ_{incl} and $\phi_{N'}$ are used for updating p_{incl} and $p_{N'}$, respectively, which are the probabilities associated with the target referential (p_{incl}) and syntactic ($p_{N'}$) representation for anaphoric *one*. We can derive the values of ϕ_{incl} and $\phi_{N'}$ by doing probabilistic inference over the graphical model in Figure 2.

A.1 ϕ_{incl}

ϕ_{incl} uses the expanded equation in (1), which calculates the probability that the antecedent includes the property (**i=yes**) given that an object present has the mentioned property (**o-m=yes**), summing over all values of intended object **O**, antecedent **A**, determiner in the antecedent **det**, modifier in the antecedent **mod**, syntactic category **C**, pronoun **Pro**, syntactic environment **env**, referential expression **R**, and property mentioned **m**.

$$\phi_{incl} = p(i = yes|o\text{-}m = yes) \tag{1a}$$

$$= \frac{p(i = yes, o\text{-}m = yes)}{p(o\text{-}m = yes)} \tag{1b}$$

$$= \frac{\sum_{O,A,det,mod,C,Pro,env,R,m} p(i = yes, o\text{-}m = yes)}{\sum_{O,A,det,mod,C,Pro,env,R,i,m} p(o\text{-}m = yes)} \tag{1c}$$

The value of ϕ_{incl} depends on data type. When ϕ_{incl} is calculated for Unambiguous <NP and Unambiguous NP data using (1), it can be shown that ϕ_{incl} = 1, which is intuitively satisfying since these data unambiguously indicate that the property should be included in the antecedent. When ϕ_{incl} is calculated for Sem-Syn ambiguous data using (1), it can be shown that ϕ_{incl} is equal to (2):

$$\phi_{incl} = \frac{rep_1}{rep_1 + rep_2 + rep_3} \tag{2}$$

where

$$rep_1 = p_{N'} * \frac{m}{m+n} * p_{incl} \tag{3a}$$

$$rep_2 = p_{N'} * \frac{n}{m+n} * (1 - p_{incl}) * \frac{1}{s} \tag{3b}$$

$$rep_3 = (1 - p_{N'}) * (1 - p_{incl}) * \frac{1}{s} \tag{3c}$$

In (3), 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 (3a), as compared to an N′ string that contains only a noun (3b). Parameter s indicates how many salient properties there are in the learner's hypothesis space at the time the data point is observed, which determines how suspicious a coincidence it is that the object just happens to have the mentioned property given that there are s salient properties the learner is aware of. Parameters m, n, and s are implicitly estimated by the learner based on prior experience, and are estimated from child-directed speech corpus frequencies when possible when implementing the modeled learners.

The quantities in (3) can be intuitively correlated with anaphoric *one* representations. For rep_1 (which is the adult representation), the syntactic category is N′ ($p_{N'}$), a modifier is used ($\frac{m}{m+n}$), and the property is included in the antecedent (p_{incl}) – this corresponds to the antecedent **A** being *red bottle* = $[_{N'}$ *red* $[_{N'}$ $[_{N^0}$ *bottle*]]]. For rep_2, the syntactic category is N′ ($p_{N'}$), a modifier is not used ($\frac{n}{m+n}$), the property is not included in the antecedent (1- p_{incl}), and the intended object **O** has the mentioned property by chance ($\frac{1}{s}$) – this corresponds to the antecedent **A** being *bottle* = $[_{N'}$ $[_{N^0}$ *bottle*]]. For rep_3, the syntactic category is N^0 (1-$p_{N'}$), the property is not included in the antecedent (1- p_{incl}), and the intended object **O** has the mentioned property by chance ($\frac{1}{s}$) – this corresponds to the antecedent **A** being *bottle* = $[_{N^0}$ *bottle*]. The numerator of (2) contains the only representation that has the property included in the antecedent (rep_1), while the denominator contains all three representations.

A.2 $\phi_{N'}$

We can derive the value of $\phi_{N'}$ similarly to ϕ_{incl}, by again doing probabilistic inference over the graphical model in Figure 2. $\phi_{N'}$ uses the expanded equation in (4), which calculates the probability that the syntactic category is N′ (**C=N′**) when the syntactic environment indicates the referential pronoun is a category smaller than NP (**env=<NP**), summing over all values of intended object **O**, antecedent **A**, determiner in the antecedent **det**, modifier in the antecedent **mod**, pronoun **Pro**, referential expression **R**, property included in the antecedent **i**, object in the present context with mentioned property **o-m**, and property mentioned **m**.

$$\phi_{N'} = p(C = N'|env =< NP) \tag{4a}$$
$$= \frac{p(C = N', env =< NP)}{p(env =< NP)} \tag{4b}$$
$$= \frac{\sum_{O,A,det,mod,Pro,R,i,o-m,m} p(C = N', env =< NP)}{\sum_{O,A,det,mod,C,Pro,R,i,o-m,m} p(env =< NP)} \tag{4c}$$

The value of $\phi_{N'}$ also depends on data type. When $\phi_{N'}$ is calculated for Unambiguous <NP data using equation (4), 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 (4), it can be shown that $\phi_{N'}$ is equal to (5):

$$\phi_{N'Sem-Syn} = \frac{rep_1 + rep_2}{rep_1 + rep_2 + rep_3} \tag{5}$$

where rep_1, rep_2, and rep_3 are the same as in (3). Equation (5) is intuitively satisfying as only rep_1 and rep_2 correspond to representations where *one* is syntactic category N′.

When $\phi_{N'}$ is calculated for Syn ambiguous data using equation (4), it can be shown that $\phi_{N'}$ is equal to (6):

$$\phi_{N'Syn} = \frac{rep_4}{rep_4 + rep_5} \tag{6}$$

where

$$rep_4 = p_{N'} * \frac{n}{m+n} \tag{7a}$$
$$rep_5 = 1 - p_{N'} \tag{7b}$$

The quantities in (7) intuitively correspond to representations for anaphoric *one* when no property is mentioned in the previous context. For rep_4, the syntactic category is N′ ($p_{N'}$) and the N' string uses only a noun ($\frac{n}{m+n}$) – this corresponds to the antecedent **A** being *bottle* = [$_{N'}$ [$_{N^0}$ *bottle*]]. For rep_5, the syntactic category is N^0 (1-$p_{N'}$), and so the string is noun-only by definition – this corresponds to the antecedent **A** being *bottle* = [$_{N^0}$ *bottle*]. The numerator of equation (6) contains the representation that has *one*'s category as N′, while the denominator contains both possible representations.

B Syn ambiguous data effects

Pearl and Lidz (2009) discovered that Syn ambiguous data can be misleading for a Bayesian learner. In the probabilistic learning model we describe, this effect is represented as the value of $p_{N'}$ lowering. This occurs even at the very beginning of learning (when p_{incl} = $p_{N'}$ = 0.50) because the representation using syntactic category N^0 (rep_5 above in section A.2) at that point has a higher probability than the representation using category N′ (rep_4 above in section A.2).

This occurs because the N′ representation in rep_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 ($\frac{n}{n+m}$); in contrast, the N^0 category by definition only includes noun-only strings. Because of this, the N′ representation is penalized, and the amount of the penalty depends on the values of m and n. More specifically, the learner we implement here considers the sets of strings covered by category N^0 and category N′, where the set of N^0 strings (size n), which contains noun-only strings, is included in the set of N′ strings (size $m + n$), which also includes modifier+noun strings. The higher the value of m is with respect to n, the more likely N′ strings are to have modifiers in the learner's experience. If m is high, it is a suspicious a coincidence to find a noun-only string as the antecedent, if the antecedent is actually category N′. For a Bayesian learner that capitalizes on suspicious coincidences, this means that when m is higher, a noun-only string causes the learner to favor the smaller of the two hypotheses, namely that *one* is category N^0. Thus, the larger that m is compared to n, the more that Syn ambiguous data cause a Bayesian learner to (incorrectly) favor the N^0 category over the N′ category.

C Frequency of different pronouns in the input

Since the +OtherPro 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 5.2.2).

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%

From this distribution, we can see that *it* is the most frequent pronoun, which makes up the bulk of the unambiguous NP examples in the +OtherPro data intake.

D p_{beh} and $p_{rep|beh}$

D.1 p_{beh}

Given a data point that has a referential expression **R=*another one***, a pronoun **Pro=*one***, a syntactic environment that indicates the pronoun is smaller than NP (**env=<NP**), a property mentioned (**m=yes**), and an object in the present context that has that property (**o-m=yes**), we can calculate how probable it is that a learner would look to the object that has the mentioned property (e.g., **O=RED BOTTLE**). For ease of exposition in the equations below, we will represent the situation where the object has the mentioned property as **O=O-M**. We can calculate p_{beh} by doing probabilistic inference over the graphical model in Figure 2, as shown in the equations in (8).

$$p_{beh} = p(O = \textsc{o-m}| \\ R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes) \quad \text{(8a)}$$

$$= \frac{p(O = \textsc{o-m}, R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes)}{p(R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes)} \quad \text{(8b)}$$

$$= \frac{\sum_{det,mod,C,i,A} p(O = \textsc{o-m}, R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes)}{\sum_{det,mod,C,i,A,O} p(R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes)} \quad \text{(8c)}$$

When p_{beh} is calculated, it can be shown that it is equivalent to the quantity in (9).

$$p_{beh} = \frac{rep_1 + rep_2 + rep_3}{rep_1 + 2 * rep_2 + 2 * rep_3} \quad \text{(9)}$$

where rep_1, rep_2, and rep_3 are defined as in (3), m = 1, n = 2.9, and s = 2 (since there are only two salient objects present in the LWF experimental setup). As before, these quantities intuitively correspond to the different outcomes. For the target representation where the property is included in the antecedent and the category is N′ (rep_1), the learner must look to the object with the mentioned property. For any of the incorrect representations (rep_2 and rep_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 object with the mentioned property by accident. The numerator represents all the outcomes where the learner looks to the object with the mentioned property, while the denominator also includes the two additional outcomes where the learner looks to the other object (rep_2 and rep_3 with incorrect behavior).

D.2 $p_{rep|beh}$

Given that the referential expression is *another one* (**R=*another one***), the pronoun is *one* (**Pro=*one***), the syntactic environment indicates the pronoun is smaller than an NP (**env=<NP**), a property was mentioned (**m=yes**), an object present has the mentioned property (**o-m=yes**), AND the child has looked at the object with the mentioned property (**O=O-M**), what is the probability that the representation is the adult representation, where the antecedent = e.g., *red bottle* (**A=*red bottle***)? This would mean that the antecedent includes the property (**i=yes**), the antecedent does not include the determiner (**det=no**), the antecedent includes a modifier (**mod=yes**), and the antecedent category is N′ (**C=N′**). This can be calculated by doing probabilistic inference over the graphical model in Figure 2, as shown in (10).

$$\begin{aligned} p_{rep|beh} = p(&A = \textit{red bottle}, i = yes, det = no, mod = yes, C = N'| \\ &R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes, O = \textsc{o-m}) \end{aligned} \quad (10a)$$

$$= \frac{p(A{=}\textit{red bottle}, i{=}yes, det{=}no, mod{=}yes, C{=}N', R{=}\textit{another one}, Pro{=}one, env{=}{<}NP, m{=}yes, \textit{o-m}{=}yes, O{=}\textsc{o-m})}{\sum_{A,i,det,mod,C} p(R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes, O = \textsc{o-m})} \quad (10b)$$

When $p_{rep|beh}$ is calculated, it can be shown that it is equal to (11).

$$p_{rep|beh} = \frac{rep_1}{rep_1 + rep_2 + rep_3} \quad (11)$$

where rep_1, rep_2, and rep_3 are calculated as in (3), but with s = 2 (again, because there are only two salient objects to choose from in the LWF experimental setup). More specifically, given that the object with the mentioned property has been looked at (whether on purpose (rep_1) or by accident (rep_2 and rep_3)), we calculate the probability that the look is due to the target representation (rep_1).

E Simulation results for different values of s

Table 9 shows the results of the learning simulations over the different input sets with different values of s (the number of properties salient to the learner when interpreting the data point) ranging from 2 to 49, with averages over 1000 runs reported and standard deviations in parentheses.

A few observations can be made about this range of results. First, with the exception of the DirectUnamb learner, the performance of the learners depends to some degree on the value of s. This is to be expected as the DirectUnamb learner uses only unambiguous <NP data in its intake, and since these data were not found in our dataset, this learner effectively learns nothing.

When we examine the results for the +OtherPro learner, we see fairly consistent overall behavior, though the exact values of each probability increase slightly as s increases. Thus, the qualitative behavior we observed before does not change – this learner decides that the antecedent should include the mentioned property (p_{incl}>0.99) and has a moderate dispreference for believing *one* is N′ when it is smaller than an NP ($p_{N'}$=0.34−0.38), no matter what the value of s.

Table 9: Probabilities after learning, using different values of s, which is the number of properties salient to the learner when interpreting a data point.

	Prob	**DirectUnamb**	**DirectFiltered**	**DirectEO**	**+OtherPro**
$s = 2$	$p_{N'}$	0.50 (<0.01)	0.34 (<0.01)	0.14 (<0.01)	0.34 (0.03)
	p_{incl}	0.50 (<0.01)	0.02 (<0.01)	<0.01 (<0.01)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.50 (<0.01)	0.50 (<0.01)	0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	<0.01 (<0.01)	<0.01 (<0.01)	0.99 (<0.01)
$s = 5$	$p_{N'}$	0.50 (<0.01)	0.94 (<0.01)	0.16 (0.02)	0.36 (0.04)
	p_{incl}	0.50 (<0.01)	0.68 (<0.01)	0.04 (0.01)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.70 (<0.01)	0.50 (<0.01)	>0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	0.58 (<0.01)	<0.01 (<0.01)	>0.99 (<0.01)
$s = 7$	$p_{N'}$	0.50 (<0.01)	0.98 (<0.01)	0.18 (0.03)	0.37 (0.04)
	p_{incl}	0.50 (<0.01)	0.91 (<0.01)	0.10 (0.05)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.88 (<0.01)	0.50 (<0.01)	>0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	0.87 (<0.01)	0.01 (0.01)	>0.99 (<0.01)
$s = 10$	$p_{N'}$	0.50 (<0.01)	0.99 (<0.01)	0.25 (0.06)	0.37 (0.04)
	p_{incl}	0.50 (<0.01)	0.96 (<0.01)	0.38 (0.18)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.95 (<0.01)	0.53 (0.04)	>0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	0.95 (<0.01)	0.11 (0.11)	>0.99 (<0.01)
$s = 20$	$p_{N'}$	0.50 (<0.01)	0.99 (<0.01)	0.34 (0.05)	0.37 (0.04)
	p_{incl}	0.50 (<0.01)	0.99 (<0.01)	0.93 (0.03)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.98 (<0.01)	0.79 (0.07)	>0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	0.98 (<0.01)	0.72 (0.11)	>0.99 (<0.01)
$s = 49$	$p_{N'}$	0.50 (<0.01)	>0.99 (<0.01)	0.37 (0.05)	0.38 (0.05)
	p_{incl}	0.50 (<0.01)	0.99 (<0.01)	0.99 (<0.01)	>0.99 (<0.01)
	p_{beh}	0.56 (<0.01)	0.99 (<0.01)	0.94 (0.02)	>0.99 (<0.01)
	$p_{rep\vert beh}$	0.23 (<0.01)	0.99 (<0.01)	0.94 (0.02)	>0.99 (<0.01)

For both the DirectFiltered and DirectEO learners, we find the results depend non-trivially on the value of s, which determines how suspicious a coincidence it is that the intended referent just happens to have the mentioned property. We examine the DirectFiltered learner first. Previous studies (Regier & Gahl, 2004; Pearl & Lidz, 2009) found that this filtered learner has a very high probability of learning *one* is N′ when it is smaller than NP ($p_{N'} \approx 1$) and a very high probability of including a mentioned property in the antecedent ($p_{incl} \approx 1$), even with s values as low as 2. We find this is true when s=7 or above; however, when s=5, the learner is much less certain that the mentioned property should be included in the antecedent (p_{incl}=0.68); when s=2, the learner is inclined to believe *one* is N^0 ($p_{N'}$=0.34) and is nearly certain that the mentioned property should NOT be included in the antecedent (p_{incl}=0.02). Similarly, when s=7 or above, the learner reliably reproduces the observed infant behavior (p_{beh}=0.88−0.99) and likely has the target representation when doing so ($p_{rep|beh}$=0.87−0.99). Yet, when s has lower values, the results are quite different (s=5: p_{beh}=0.70, $p_{rep|beh}$=0.58; s=2: p_{beh}=0.50, $p_{rep|beh}$=0.02).

If we examine the DirectEO learner, we again find variation in the overall pattern of behavior. Pearl and Lidz (2009) found that this learner has a very low probability of learning *one* is N′ when it is smaller than NP ($p_{N'} \approx 0$), and a very high probability of including a mentioned property in the antecedent ($p_{incl} \approx 1$), even with s values as low as 5. When s=20 or 49, we see something close to this behavior where a dispreference for *one* as N′ ($p_{N'}$=0.34$-$0.37) occurs with a strong preference for including the mentioned property in the antecedent (p_{incl}=0.93$-$0.99). However, for $s \leq 10$, low values of $p_{N'}$ occur with low values of p_{incl} ($p_{N'}$=0.14$-$0.25, p_{incl}=<0.01$-$0.38). Though Pearl and Lidz (2009) don't assess this learner's ability to generate the LWF experimental results, it is likely their learner would behave as we see the learners with s=20 or 49 do here – specifically, because p_{incl} is so high, there is a high probability of generating the LWF behavior (p_{beh}=0.79$-$0.94) and a strong probability of having the target representation when doing so ($p_{rep|beh}$=0.72$-$0.94). This is the same behavior we found in the +OtherPro learner. However, the DirectEO learner differs by failing to exhibit this behavior this when $s \leq 10$: The learner is instead at chance for generating the LWF behavior (p_{beh}=0.50$-$0.53) and is unlikely to have the target representation if it happens to do so ($p_{rep|beh}$=<0.01$-$0.11).

Why do we see these differences in learner behavior, compared to previous studies? The answer appears to lie in the probabilistic learning model. In particular, recall that there is a tight connection between syntactic and semantic information in the model (Figure 2), as both are used to determine the linguistic antecedent. In particular, each ALWAYS impacts the selection of the antecedent when a property is mentioned, which was not true in the previous probabilistic learning models used by R&G and P&L. This is reflected in the update equations for the Sem-Syn ambiguous data, where both ϕ_{incl} and $\phi_{N'}$ involve the current values of p_{incl} and $p_{N'}$, as do all the equations corresponding to the probabilities of the different representations (recall the equations in (3)). This means that there is an inherent linking between these two probabilities when Sem-Syn data are encountered.

For example, if p_{incl} is very high (as it would be for high values of s), it can make the value of $\phi_{N'}$ higher for Sem-Syn ambiguous data (and so increase $p_{N'}$ more).[27] This subsequently gives a very large boost to $p_{N'}$, thus increasing the power of these kind of data. In other words, when s is high enough, the suspicious coincidence is very strong, and thus both p_{incl} and $p_{N'}$ benefit strongly – each Sem-Syn ambiguous data point effectively functions as if it were an unambiguous <NP data point.

However, the opposite problem strikes when s is low and the coincidence is not suspicious enough. When this occurs, p_{incl} is actually decreased slightly if $p_{N'}$ is not high enough. For example, in the initial state when $p_{N'}$=0.5, p_{incl}=0.5, and s=2, seeing a Sem-Syn ambiguous data point leads to a p_{incl} of 0.409. This causes subsequent Sem-Syn ambiguous data points to have even less of a positive effect on p_{incl} – which eventually drags down $p_{N'}$. For example, if this same learner encounters 20 Sem-Syn ambiguous data points in a row initially, its p_{incl} will then be 0.12 and its $p_{N'}$ 0.48. Thus, when s is low, the power of Sem-Syn ambiguous data is significantly lessened, and can even cause these data to have a detrimental effect on learning. This is why the DirectFiltered learner fails for low s values. The situation is worse when Syn ambiguous data are included in the mix, as for the DirectEO learner – not only are the Sem-Syn ambiguous data insufficiently powerful, but the Syn ambiguous data cause $p_{N'}$ to plummet.

[27] See (3) for a reminder of why this happens.

Notably, when unambiguous NP data are added into the mix for the +OtherPro learner, p_{incl} is only ever increased every time one of these data points is encountered. Thus, even if s is very low, these data points compensate for the insufficiently helpful Sem-Syn ambiguous data. Due to the linking between p_{incl} and $p_{N'}$ in the Sem-Syn ambiguous data update, the high p_{incl} value will cause Sem-Syn ambiguous data points to act as if they were unambiguous <NP data points, and so $p_{N'}$ is also increased. This is why the +OtherPro learner is not susceptible to changes in its behavior when s changes. Still, because this benefit to $p_{N'}$ only occurs when Sem-Syn ambiguous data are encountered, and these are relatively few, the final $p_{N'}$ value is still fairly low (0.34−0.38). If we remove the Sem-Syn ambiguous data from the +OtherPro learner's dataset (i.e., it only encounters Syn ambiguous and unambiguous NP data points, as well as uninformative data points), we can see a final $p_{N'}$ that is much lower ($p_{N'}$=0.13), even though p_{incl}=>0.99.

To summarize, the behavior of the learner that uses indirect positive evidence is robust because it can leverage unambiguous NP data to compensate for (or further enhance the effectiveness of) the Sem-Syn ambiguous data. In contrast, learners who are restricted to only direct positive evidence and indirect negative evidence are greatly affected by how suspicious a coincidence Sem-Syn ambiguous data points are. Our results are similar to previous results for the DirectFiltered and DirectEO learners for certain values of s. However, because of the way semantic and syntactic information are integrated in the probabilistic learning model we present here (i.e., both information types are given equal weight), our results deviate from prior results with these learners for other values of s. In particular, we find a higher $p_{N'}$ than (Pearl & Lidz, 2009) did with their integrated probabilistic learning model for the DirectEO learner with high values of s. We also find low values of $p_{N'}$ and p_{incl} for the DirectFiltered learner when s is very low.

We additionally note that these results are not due to the particular duration of the learning period we chose. Figure 5 shows a sample trajectory for the +OtherPro learner with s=7, which converges on its final probabilities fairly quickly – little change occurs after the first few hundred data points. This is true for all the learners and all s values. Thus, we would not predict the behavior of any of the learners to alter appreciably if they were exposed to more data, unless those data were very different from the data they had been learning from already or they were able to use those data in a very different way.

F Other learning strategies

F.1 Using data more effectively

The probabilistic 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 salient properties it could have had, implemented with parameter s).

However, there may be more nuanced ways to interpret how suspicious a coincidence is.[28] For

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

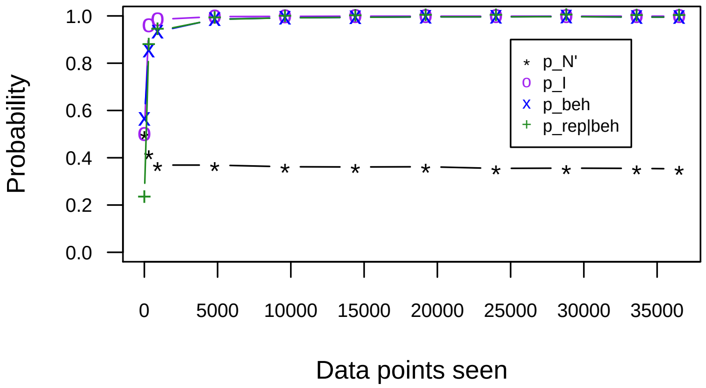

Figure 5: A sample trajectory of +OtherPro probabilities over the learning period, with s=7.

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 <NP data (which has ϕ_{incl} = $\phi_{N'}$ = 1), since it is more likely that the mentioned property is included in the antecedent (p_{incl}) 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, this effect can be simulated somewhat with high values of s, since high values of s cause Sem-Syn ambiguous data to have an impact more like unambiguous <NP data. If we assume that ALL Sem-Syn ambiguous data are these "more influential" kind (clearly an overestimate), we can look to the results in table 9 in Appendix E to see the impact. The qualitative behavior of the learners able to use some of the data (i.e., not the DirectUnamb learner) differs for only one learner when s=49, the DirectEO learner. For this learner, instead of completely failing to learn the adult representation, it behaves similarly to the +OtherPro learner by learning a context-dependent representation (see Appendix E for more details). Thus, there is clearly some benefit a Bayesian learner can reap from these data, though the qualitative effect for learners restricted to direct evidence will depend on exactly how frequent these data are in the input.

F.2 Using sophisticated contextual cues

Another source of information involves more sophisticated contextual cues. Some examples are shown below in (12):

(12a) *I hate that red bottle– do you have another one?*
(12b) *I want this* ***red*** *bottle, and you want* ***that*** *one.*
(**boldface** indicates emphasis)

Many adults would interpret the referent of *one* in both cases as a BOTTLE that is not red. For (12a), this is perhaps based on the verb *hate*, and the inference that someone would not ask for another of something they hate. For (12b), this is perhaps based on the contrastive focus that occurs between *red* and *that*. In both cases, this involves an inference that draws from information beyond the default syntactic and semantic representation. In (12a), this is an inference about when a speaker would use *hate* in this way; in (12b), this is an inference about when speakers use contrastive focus. The default interpretation of *one* seems to include the modifier (see 13). In (13a), it seems the speaker is requesting another red bottle. In (13b), while there is contrastive focus with *that*, it doesn't interfere with the interpretation of *one*'s antecedent as *red bottle*.

(13a) *I love that red bottle – do you have another one?*
(13b) *I want* ***this*** *red bottle, and you want* ***that*** *one.*
(**boldface** indicates emphasis)

We note that we did not find any occurrences of data like (12a) in our corpus analysis, which suggests that young children probably do not encounter these data very often.[29] 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 assume that the initial target state for learning is the default interpretation where the mentioned property is included in the antecedent. It would be useful to assess when children have the adult interpretations for non-default anaphoric *one* examples like those in (12), as this would allow us to further fine-tune the acquisition trajectory.

[29] Our corpus was not marked for contrastive focus, so it is unclear how often data like (12b) appear.