The role of indirect positive evidence in syntactic acquisition: A look at anaphoric *one*

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

February 2, 2015

Abstract

Language learners are often faced with a scenario where the data allow multiple generalizations, even though only one is actually correct. One promising solution to this problem is that children are equipped with helpful learning strategies that guide the types of generalizations made from the data. Two successful approaches in recent work for identifying these strategies have involved (i) expanding the set of informative data to include INDIRECT POSITIVE EVIDENCE, and (ii) using observable behavior as a target state for learning. We apply both these ideas to the case study of English anaphoric *one*, using computationally modeled learners who assume *one*'s antecedent is the same syntactic category as *one* and form their generalizations based on realistic data. We demonstrate that a learner who is biased to include indirect positive evidence coming from other pronouns in English can generate 18-month-old looking preference behavior. Interestingly, we find that the knowledge state responsible for this target behavior is a context-dependent representation for anaphoric *one*, rather than the adult representation, but this immature representation can suffice in many communicative contexts involving anaphoric *one*. More generally, these results suggest that children may be leveraging broader sets of data to make the syntactic generalizations leading to their observed behavior, rather than selectively restricting their input. We additionally discuss the components of the learning strategies capable of producing the observed behavior, including their possible origin and whether they may be useful for making other linguistic generalizations.

Keywords: anaphoric *one*; acquisition; computational modeling; indirect positive evidence; induction problems; online probabilistic learning

1 Introduction

Language acquisition, as with many other kinds of knowledge acquisition, involves making generalizations from data. One recurring issue is that many generalizations may be possible from the data available, but often only one is the target generalization, representing the knowledge adults have. This scenario describes an induction problem, sometimes referred to in the language acquisition literature as the "Poverty of the Stimulus" (e.g., Chomsky, 1980a, 1980b; Crain, 1991; Lightfoot, 1989), the "Logical Problem of Language Acquisition" (e.g., Baker & McCarthy, 1981; Hornstein & Lightfoot, 1981; Pinker, 2004), or "Plato's Problem" (e.g. Chomsky, 1988; Dresher, 2003). One promising solution to induction problems is that the language learner is equipped with helpful learning strategies that guide the types of generalizations made from the data. Traditionally, proposals for the strategies necessary for making correct syntactic generalizations have involved fairly specific (and often linguistic) prior knowledge. Some examples include the following:

(i) knowing syntactic rules are structure-dependent (Chomsky, 1980a; Anderson & Lightfoot, 2000; Fodor & Crowther, 2002; Berwick, Pietroski, Yankama, & Chomsky, 2011; Anderson, 2013)

(ii) knowing certain dependencies are limited to spanning no more than a single specific, abstract linguistic structure (Chomsky, 1973; Huang, 1982; Lasnik & Saito, 1984)

(iii) knowing certain syntactic category assignments are illicit for certain words in a language (Baker, 1978)

However, recent investigations have suggested that learning strategies involving less specific knowledge may be sufficient to learn the target syntactic generalizations in several cases (e.g., Regier & Gahl, 2004; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Pearl & Lidz, 2009; Pearl & Mis, 2011; Perfors, Tenenbaum, & Regier, 2011; Pearl & Sprouse, 2013b, 2013a). Interestingly, a common successful approach in some of the most recent work (Pearl & Mis, 2011; Perfors, Tenenbaum, & Regier, 2011; Pearl & Sprouse, 2013b, 2013a) involves expanding the set of informative data to include INDIRECT POSITIVE EVIDENCE (discussed in more detail below in section 2). In addition, several recent computational approaches have focused on learning syntactic generalizations that lead to observed behavior (e.g., Pearl & Mis, 2011; Perfors, Tenenbaum, & Regier, 2011; Pearl & Sprouse, 2013b, 2013a), with the idea that observable behavior is a more direct empirical checkpoint than the knowledge state responsible for that behavior.

Here, we apply both these ideas to the case study of English anaphoric *one*, using computationally modeled learners who form their generalizations based on realistic input data (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; Pearl, 2011; Pearl & Mis, 2011; Perfors, Tenenbaum, & Regier, 2011; Yang, 2012; Sakas & Fodor, 2012; Legate & Yang, 2013; Pearl & Sprouse, 2013b, 2013a). We demonstrate that a learner who assumes *one*'s antecedent is the same syntactic category as *one* and is biased to include indirect positive evidence coming from other pronouns in English can generate the looking preference behavior observed in 18-month-olds (Lidz, Waxman, & Freedman, 2003). Interestingly, we find that the knowledge state responsible for this target behavior in this learner is a context-dependent representation for anaphoric *one*, rather than the adult representation. Nonetheless, the linguistic generalizations made by this learner can suffice in many communicative contexts involving anaphoric *one*, highlighting their utility even

though they lead to immature representations of *one*. More generally, these results suggest that children may be leveraging broader sets of data to make the syntactic generalizations leading to their observed behavior, rather than selectively restricting their input.

In the remainder of this paper, we first discuss different types of evidence available in principle to the learner, including indirect positive evidence. We then describe how to define learning problems in general, using components that can be specified for any particular learning problem by drawing on theoretical, experimental, and computational results. We subsequently describe the details of the English anaphoric *one* learning problem we investigate, including relevant aspects of adult knowledge, young children's observed behavior, the data available for learning, and several proposed learning strategies for solving this learning problem, including a new one that relies on indirect positive evidence. We test the effectiveness of the strategies by embedding them in an online probabilistic learning model that is based on a formal model of understanding a referential expression, incorporating both syntactic and referential information. We investigate the ability of each strategy to learn the target generalizations and generate the observed toddler behavior. The modeling results demonstrate that an immature context-dependent representation of *one* is compatible with observed toddler behavior. We conclude by discussing the components of the learning strategies capable of producing the observed behavior, including their origin and whether they are useful for making other linguistic generalizations.

2 Types of evidence

There are at least two dimensions that seem relevant when describing the types of evidence available to a learner (Figure 1):

(i) POSITIVE vs. NEGATIVE: Is the evidence about items that are present in the language (*positive*) or about items that are absent in the language (*negative*)?

(ii) DIRECT vs. INDIRECT: Is it certain that the items are (un)grammatical (*direct*) or does it require inference on the learner's part (*indirect*)?

To illustrate the four evidence types captured by these distinctions, consider the utterances in (1) with respect to learning about anaphoric *one* in English:

(1) a. Jack already has a red cup but he wants another one.
b. * Jack drank from the edge of the cup while Lily drank from the one of the bowl.
c. Jack has a red cup and Lily wants it.

DIRECT POSITIVE evidence would correspond to items such as (1a) appearing in the input, an indication that they are grammatical because they are used by speakers. Direct positive evidence has traditionally been assumed to be available to learners, often as the *only* evidence available (Chomsky, 1980a; Baker & McCarthy, 1981; Bowerman, 1988; Wexler & Culicover, 1980; Crain, 1991; Hornstein & Lightfoot, 1981; Roeper, 1981; Lightfoot, 1982b; Pinker, 1984, 1989; Anderson & Lightfoot, 2000, 2002; Crain & Pietroski, 2002; Legate & Yang, 2002; Lidz et al., 2003; Gualmini, 2007; Crain & Thornton, 2012; Anderson, 2013).

DIRECT NEGATIVE evidence would correspond to the learner being explicitly informed that

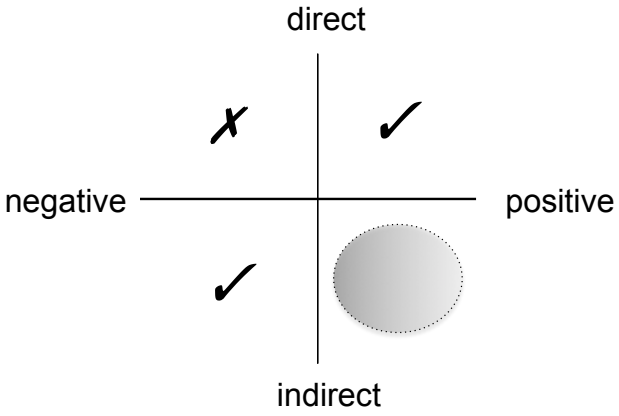

Figure 1: Evidence types available to a learner in principle, along with some indicators of whether they are believed to be available in practice. The circle in the indirect positive evidence quadrant highlights that this type has been under-investigated.

items like (1b) are ungrammatical. Given that children are notoriously resistant to being corrected (e.g McNeill, 1996; Zwicky, 1970; Braine, 1971), particularly about their syntactic generalizations, direct negative evidence for syntactic knowledge has typically been assumed to be unavailable to the learner or ignored (Brown & Hanlon, 1970; Braine, 1971; Baker & McCarthy, 1981; Bowerman, 1983; Fodor & Crain, 1987; Bowerman, 1988; Grimshaw & Pinker, 1989; Lasnik, 1989; Marcus, 1993, 1999; Anderson & Lightfoot, 2000; Crain & Pietroski, 2002; Legate & Yang, 2002; Lidz et al., 2003; Crain & Thornton, 2012; Anderson, 2013).

INDIRECT NEGATIVE evidence[1] would correspond to the learner noticing that items like (1b) are absent from the input, and so inferring that these items are absent because they are ungrammatical. Indirect negative evidence has been argued to be available, particularly to statistical learners that form expectations about how frequently items should appear in the input (e.g., via some form of entrenchment: Rohde & Plaut, 1999; Regier & Gahl, 2004; Clark & Lappin, 2009; Foraker et al., 2009; Perfors, Tenenbaum, & Wonnacott, 2010; Perfors, Tenenbaum, & Regier, 2011; Ambridge et al., 2013; Ramscar, Dye, & McCauley, 2013) and learners that use statistical pre-emption to recognize when an alternative semantically and pragmatically related item is used instead of the item in question (e.g., Boyd & Goldberg, 2011; Goldberg, 2011; Ambridge et al., 2013).

INDIRECT POSITIVE evidence would correspond to the learner observing the presence of items like (1c) – which do not actually involve *one* – and using those data to make inferences about (1a) and (1b). For example, the learner can form expectations that (1a) should appear while (1b) should not, even if neither (1a) or (1b) has appeared yet.[2] More formally, examples involving linguis-

[1]This is sometimes called *implicit negative* evidence (Rohde & Plaut, 1999), and can be implemented via *entrenchment* (Ambridge, Pine, Rowland, Chang, & Bidgood, 2013) or *statistical pre-emption* (Boyd & Goldberg, 2011; Goldberg, 2011; Ambridge et al., 2013), among other ways.

[2]We note that some expectations formed on the basis of the indirect positive evidence from (1c) can be used as indirect negative evidence for (1b), since they are expectations about (1b)'s absence. Thus, indirect positive evidence may lead to indirect negative evidence.

tic knowledge L1 appear in the input (e.g., how to interpret the pronoun *it*) and allow the learner to learn about knowledge L2 (e.g., how to interpret the pronoun *one*). Indirect positive evidence seems to have only recently been recognized either implicitly (e.g., Reali & Christiansen, 2005; Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2008; Foraker et al., 2009; Perfors, Tenenbaum, & Regier, 2011) or explicitly (e.g., Pearl & Mis, 2011; Pearl & Sprouse, 2013b, 2013a) as a type of informative data for syntax acquisition. Interestingly, it corresponds quite well to the ideas behind linguistic PARAMETERS in generative linguistic theory and OVERHYPOTHESES in Bayesian inference. In particular, both parameters and overhypotheses allow positive evidence about items besides the specific items of interest to be leveraged by the learner. For parameters, 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 knowledge yields information about others (e.g., Chomsky, 1981; Viau & Lidz, 2011; Pearl & Lidz, 2013). For example, if parameter P controls knowledge L1 and L2, data about knowledge L1 can set the value of P, which then provides information about knowledge L2. Similarly for overhypotheses, if hypotheses H1 and H2 are instances of overhypothesis O, data for H1 can help determine O, which in turn helps the learner infer something about H2 (Kemp, Perfors, & Tenenbaum, 2007; Perfors, Tenenbaum, Griffiths, & Xu, 2011). Thus, while indirect positive evidence has rarely been explicitly recognized in prior syntactic acquisition investigations, it seems to be a natural consequence of both linguistic parameters and Bayesian overhypotheses. Here, we investigate its application for learning syntactic knowledge related to English anaphoric *one*.

3 Defining learning problems

One way to characterize the language learning process is that the learner starts in some initial state, having at her disposal prior knowledge, learning abilities, and learning biases (some of which form a specific learning strategy). As she encounters input over time, she applies her learning abilities to that input in order to update her knowledge state, and this process is guided by her learning strategy. Eventually, she updates her knowledge state to the target knowledge state, which allows her to generate target linguistic behavior. This description allows us to identify four important components of the learning problem: the INITIAL STATE of the learner, the DATA INTAKE used by the learner, the LEARNING PERIOD during which the learner is updating her knowledge, and the TARGET STATE the learner is trying to reach. For any given learning problem, we can attempt to specify these components by using theoretical, experimental, and computational methods.

3.1 Initial state

The INITIAL STATE consists of the child's initial knowledge state, the child's existing learning capabilities, and the child's learning biases. 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 syntactic categories? Similarly, experimental methods can also be used to assess what learning capabilities and biases children have, e.g., whether they *can* use different inference procedures and whether they actually *do* in realistic learning scenarios.

We note that we are allowing a broad definition of "learning bias", where "bias" simply represents a preference of some kind. Under this view, a learning bias can pertain to either the hypothesis space or the learning mechanism in some way. An example bias about the hypothesis space might involve viewing the learning problem as a decision between two syntactic categories instead of three. An example bias about the learning mechanism might involve what update procedure to use, such as probabilistic inference (e.g., Pearl & Lidz, 2009; Yang, 2012) vs. a random step algorithm (e.g., Gibson & Wexler, 1994; Niyogi & Berwick, 1996; Sakas, 2003).

3.2 Data intake

The DATA INTAKE (sometimes called *acquisitional intake*) for a learning problem refers to the data children use for learning (Fodor, 1998; Pearl & Weinberg, 2007; Pearl & Lidz, 2009; Gagliardi & Lidz, 2014; Omaki & Lidz, 2014; Lidz & Gagliardi, 2015), and is often a subset of the available input. In particular, the data intake is the subset of the available input that the child views as relevant or informative for the learning task at hand. This is defined by the prior knowledge and biases the child has in the initial state. For example, if children are biased to assume only direct evidence is relevant, they may ignore indirect evidence that could otherwise be informative. Once the information children use is defined, corpus analysis methods can often provide realistic estimates of the input children encounter.

3.3 Learning period

The LEARNING PERIOD defines how long children have to reach the target state. Experimental methods can provide information about the beginning and ending of the learning period, usually by assessing the knowledge children have at a particular age, as demonstrated by their behavior. For example, if the child's initial state must contain knowledge of syntactic categories, the learning period could not begin before children attain this knowledge. Similarly, target linguistic behavior is often used to assess whether children have learned the target knowledge – once children display this behavior, this marks the end of the learning period. Often in computational studies, the learning period is implemented as children receiving a specific amount of data, which is the amount they would encounter between the relevant ages. After that quantity of data, they should then reach the target state.

3.4 Target state

The TARGET STATE is often defined in terms of the knowledge children are trying to attain, though it is typically inferred from observable linguistic behavior. For example, Lidz et al. (2003) assessed knowledge of English anaphoric *one* in toddlers by measuring their looking preferences, which were similar to adult looking preferences. The basic idea is that when the observed behavior

matches the target (adult) behavior in properly controlled experiments, it is because the underlying knowledge generating that behavior matches the target knowledge generating the adult behavior. This is an assumption, of course, but it allows empirical results pertaining to the target behavior to be a proxy for the target knowledge, whose exact form is specified by theoretical methods. Relatedly, it is useful to determine which knowledge states can generate the target behavior, as the target knowledge state may not be the only knowledge state capable of doing so. Thus, the target state can be specified by using both theoretical and experimental methods, and this is the approach we pursue here for learning about English anaphoric *one*.

4 Defining the English anaphoric *one* learning problem

A learning problem concerning a specific aspect of knowledge about English anaphoric *one* has been vigorously debated in the literature (e.g., Baker (1978); Hornstein and Lightfoot (1981); Lightfoot (1982b); Crain (1991); Ramsey and Stich (1991); 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); Pearl and Mis (2011); Payne, Pullum, Scholz, and Berlage (2013), among others). We first define this learning problem in terms of the components described above, and then review the learning strategies that have been investigated previously for this learning problem. We then present a new strategy that relies on indirect positive evidence.

4.1 Specifying the target state

4.1.1 Adult behavior and knowledge

Consider the scenario and utterance in (2).

(2) Situation: The speaker sees a red bottle.
Utterance: *Look – a red bottle!*
Situation: The speaker then sees a purple bottle and a second red bottle.
Utterance: *Oh, look – another one!*

In this scenario, an available interpretation is that *one* refers to the second red bottle present, rather than the purple bottle (i.e., the referential expression in the second utterance is interpreted as *another red bottle*). Syntactically and semantically, this means that the linguistic antecedent of *one* is the string *red bottle*. Referentially, because the antecedent includes the property red, this means the referent of *one* needs to be a RED BOTTLE (which the red bottle is), and not just a BOTTLE (which both the purple and red bottles are). Thus, the representation of *one* in this utterance requires both syntactic/semantic and referential components.

4.1.1.1 Underlying structure: Syntactic vs. semantic An important assumption for interpreting anaphoric elements is that the anaphor has the same structure as its antecedent. Traditionally, this was assumed to be a syntactic structure (specifically, a particular syntactic category)

(Jackendoff, 1977; Baker, 1978) and many subsequent theoretical, psycholinguistic, and computational studies have adopted this assumption (e.g., Hornstein & Lightfoot, 1981; Lightfoot, 1982b; Lidz et al., 2003; Regier & Gahl, 2004; Foraker et al., 2009; Pearl & Lidz, 2009). Recently however, Payne et al. (2013) have argued that it is instead only the semantic structure (specifically, a particular semantic type) that *one* and its antecedent have in common, since antecedents for *one* that adults allow do not always correspond to syntactic constituents.

We investigate the traditional syntactic instantiation here, with learners assuming *one* and its antecedent have a syntactic category in common and the structural part of the learning problem is to determine which category that is. However, if Payne et al. (2013) are correct, this is not the ultimate target knowledge state for *one*'s structure – instead, learners using this approach would need to shift from a syntactic strructural representation to a semantic one at some point (presumably upon discovering sufficient evidence of non-constituent antecedents). In contrast, if children begin with the assumption that *one* and its antecedent have only a semantic type in common, no shift would be necessary to reach the adult knowledge state. Currently, it is unclear which assumption young children have – that is, if they initially rely on syntactic or semantic structure when learning to interpret anaphors. Importantly, for questions of syntactic knowledge acquisition, only the syntactic instantiation we investigate has anything concrete to offer (as Payne et al. (2013) note), though both instantiations are worth investigating for the more general issue of how children acquire linguistic knowledge of any kind.

4.1.1.2 The syntactic instantiation The string *a red bottle* can be described as having the syntactic structure in Figure 2, shown in bracket notation in (3) (Chomsky, 1970; Jackendoff, 1977).

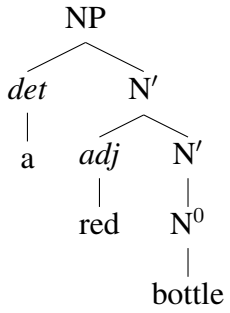

Figure 2: Phrase structure tree for *a red bottle*.

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

The syntactic category N^0 contains noun strings (e.g., *bottle*) only, and the category NP contains any noun phrase (e.g., *a red bottle*). The syntactic category N' can contain both noun strings (e.g., *bottle*) and modifier+noun strings (e.g., *red bottle*).[3]

[3]We note that while we use the labels N' and N^0, other theoretical implementations may use different labels to distinguish these hierarchical levels. The actual labels themselves are immaterial – it is only relevant for our purposes

Since *one*'s antecedent can be *red bottle* in (2), then *one* must be category N′ in this context. 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 strings like *bottle*, and we would only be able to interpret the second utterance in (2) as *Oh, look – another bottle!*

One way to represent this adult knowledge of *one* for data like (2) is as in (4). On the syntactic side, the syntactic category of *one* is N′ and so *one*'s antecedent is also N′. On the referential side, the referent has the property mentioned in the potential antecedent (e.g., *red*). 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*).

(4) Adult anaphoric *one* knowledge in utterances like
Look – a red bottle! Oh, look – another one! when *one* is interpreted as *red bottle*

a. Syntactic category of *one*: N′

b. Referent and antecedent: The referent of *one* has the mentioned property (*red*). So, *one*'s antecedent is $[_{N'}$ *red* $[_{N'}$ $[_{N^0}$ *bottle*]]] rather than $[_{N'}$ $[_{N^0}$ *bottle*]].

Understanding a referential expression that involves the pronoun *one* draws on this knowledge, and can be formalized as part of a more general model of understanding a referential expression that involves any pronoun having a linguistic antecedent, shown in Figure 3. Notably, both syntactic and referential information can be used by the learner to infer the linguistic antecedent, which identifies the pronoun's referent.

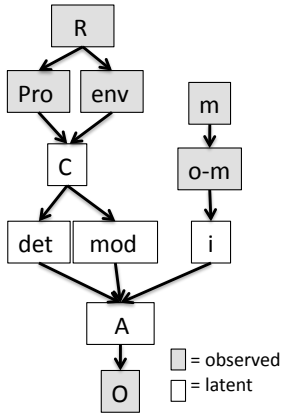

Figure 3: Model of understanding a referential expression that involves a pronoun. The variables correspond to (i) **syntactic** information (R, Pro, env, C, det, mod), (ii) **referential** information (m, o-m, i), (iii) the **linguistic antecedent** (A), and (iv) the **intended referent** (O). All variables are discrete, with binary variables in lowercase.

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 Appendix G for alternate theoretical representations that additionally differentiate *red bottle* from *bottle*, which lead to similar learning results as those presented below.

Beginning with the syntactic information (shown on the lefthand side of Figure 3), **R** is the referential string 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. Specifically, 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 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 pronoun, which could 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 infer 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 referential information (shown on the righthand side of Figure 3), **m** concerns whether a property was mentioned in the potential linguistic antecedent, 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 depend on the referential information available from the data point.

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**:[4]

(a) **det=yes**, **mod=yes**, **i=yes** yielding e.g., **A= *a 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, which is

[4]In particular, **i** and **mod** must agree. If **i=yes** and **mod=no**, the referential intention 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**). This would be the case for category N^0. If **i=no** and **mod=yes**, the referential intention 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 **i** (and so **mod**) = **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 2's phrase structure tree).

why we have indicated it as an observed variable in Figure 3. That is, the learner can often observe what object is the intended referent, even if the linguistic antecedent is ambiguous. For example, consider an utterance like "*Look – a red bottle! Oh, look – another one!*" in a scenario with two red bottles present. Even if it is unclear whether the antecedent is *red bottle* or *bottle*, since both are compatible with the second object present (a RED BOTTLE), the basic point is that the listener knows which object is intended as *one*'s referent (the second 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 not.

The values that each of the variables in the model can take on are summarized in Table 1.

Table 1: Variable values in referential data points, with syntactic variables on the left and referential variables on the right. Observable variables are in **bold**. 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**).

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

When interpreting a referential expression involving *one*, such as the utterance in (2), adults can use both their acquired syntactic and referential knowledge. On the syntactic side, they know that *one*'s category is N′ (**C=N′**) when it is used this way and on the referential side, they know that a mentioned property should be included in the linguistic antecedent (**i=yes**). This combined knowledge then yields the antecedent (e.g., **A=*red bottle***), and the knowledge that the referent should have the mentioned property (e.g., **O=RED BOTTLE**).

4.1.2 Child behavior and knowledge

To assess child knowledge of anaphoric *one* in these scenarios, Lidz et al. (2003) (henceforth **LWF**) observed the behavior of 18-month-olds in experimental scenarios designed to reveal how they were interpreting *one*. Using an intermodal preferential looking paradigm (Spelke, 1979; Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987), LWF examined the looking behavior of 18-month-olds using the setup in (5):

(5) LWF experimental setup

a. **Habituation**
Example scenario: A red bottle appears on the screen.
Example utterance: "Look, a red bottle!"

b. **Test**
Example scenario: A red bottle and a purple bottle appear on the screen.
Example utterance: "Now look..."
(i) **Neutral**: "What do you see now?"
(ii) **Noun-only**: "Do you see another bottle?"
(iii) **Anaphoric**: "Do you see another one?"
(iv) **Adjective-noun**: "Do you see another red bottle?"

For each of the test conditions, LWF measured the amount of time infants looked at the familiar bottle vs. the novel bottle (e.g., the red bottle vs. the purple bottle in (5b)). For both the Neutral condition (5b-i) and the Noun-only condition (5b-ii), 18-month-olds had a novelty preference, and looked to the familiar bottle 45.9% of the time, which is significantly below chance. This indicated that their default preference was the same as their preference when asked to look for a *bottle*: to look at the object that was new (e.g., the purple bottle in (5b)). In contrast, for both the Anaphoric condition (5b-iii) and the Adjective-noun condition (5b-iv), 18-month-olds had a familiarity preference, and looked to the familiar bottle 58.7% of the time, which was significantly above chance. This indicated that their preference when asked to interpret anaphoric *one* was the same as their preference when asked to explicitly look for another *red bottle*, and markedly different from their default novelty preference.[5]

LWF interpreted this to mean that by 18 months, children have acquired the same representation for anaphoric *one* that adults have.[6] In particular, in this scenario, 18-month-olds interpret the linguistic antecedent of *one* to be the N′ *red bottle*, and the referent of *one* to be a RED BOTTLE, rather than just any BOTTLE.

Importantly for our purposes, these experimental results also provide a useful specification of the target state behavior. In particular, when presented with the LWF experimental paradigm, the learner should display the same familiarity preference that 18-month-olds do when hearing an utterance containing anaphoric *one* like (5b-iii).

4.2 Specifying the learning period

The LWF results suggest that learners should acquire this aspect of *one* interpretation by 18 months. But when would learning begin? Pearl and Lidz (2009) assumed that children would need to know syntactic categories before they would be able to learn about the representation of anaphoric *one*. They estimated this knowledge to be in place at 14 months at the earliest, based on experimental

[5]These probabilities were calculated by estimating the looking times from the figures in LWF, described below:
(a) Neutral $\approx$ 2.0 seconds for the familiar bottle, 2.5 seconds for the novel bottle
(b) Noun-only $\approx$ 2.65 seconds for the familiar bottle, 2.95 seconds for the novel bottle
(c) Anaphoric $\approx$ 2.75 seconds for the familiar bottle, 1.95 seconds for the novel bottle
(d) Adjective-noun $\approx$ 3.0 seconds for the familiar bottle, 2.1 seconds for the novel bottle.
An average was taken of the percentage of the time spent looking at the familiar bottle for the conditions causing a novelty preference (Neutral and Noun-only) and the conditions causing a familiarity preference (Anaphoric and Adjective-noun).

[6]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).

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). We adopt their assumptions, and specify the learning period as being between 14 months and 18 months.

4.3 Specifying the data intake

The data intake is defined as any data the learner views as informative. Clearly, this determination must depend on the biases in the learner's initial state, which cause the learner to perceive some data as relevant and other data as irrelevant. It is useful to review the different data available to get a sense of what data might be considered informative, before describing the data that different learning proposals suggest is informative. A formal description of the properties of each data type with respect to the model of understanding a referential expression in Figure 3 is provided in Appendix A.

4.3.1 Direct positive evidence

There are several types of direct positive evidence that have been considered informative by prior learning strategies. The first is unambiguous data using anaphoric *one* (6), which are rare because they require a specific conjunction of situation and utterance, in addition to a potentially sophisticated reasoning process on the learner's part.

(6) Direct positive unambiguous (DIRUNAMB) 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 (6), 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. 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′.

Another type of direct positive evidence involves *one* data that are ambiguous both with respect to *one*'s referent and to the syntactic category of *one*.

(7) Direct positive referentially and syntactically ambiguous (DIRREFSYNAMB) example
Situation: There are two red bottles present.
Utterance: *Look, a red bottle! Oh look– another one!*

Referentially and syntactically ambiguous data like (7) are unclear about both the properties of the referent and the category of *one*. In (7), if the child believed that the referent was simply a 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 18-month-old interpretation in the LWF experiment), 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 referentially. This data point is ambiguous syntactically because of the possibility that the antecedent could be *bottle*, which is either N^0 or N'.

A third type of direct positive evidence involves *one* data that are ambiguous with respect only to the syntactic category of *one*.

(8) Direct positive syntactically ambiguous (DIRSYNAMB) example
Situation: There are two bottles present.
Utterance: *Look, a bottle! Oh look – another one!*

Syntactically ambiguous data like (8) do not clearly indicate the category of *one*, even though the referent is clear. In (8), the referent must be a BOTTLE since the antecedent can only be *bottle*. But, is the syntactic structure $[_{N'}\ [_{N^0}\ \textit{bottle}]]$ or just $[_{N^0}\ \textit{bottle}]$? Notably, if the child believed that *one* was category N^0, this data point would not conflict with that hypothesis since it is compatible with the antecedent being $[_{N^0}\ \textit{bottle}]$.

4.3.2 Indirect positive evidence

A type of indirect positive evidence available comes from data containing other pronouns (e.g., *it*, *him*, *her*) that have a linguistic antecedent. More specifically, because the ability for a linguistic element to be interpreted as another string is not unique to *one*, a learner may be able to learn something about how to interpret *one* by observing how to interpret these other pronouns. We note that while this is only one type of potential indirect positive evidence, we have chosen to focus on its impact on acquisition because of its similarity to the direct positive evidence previously assumed to be part of the learner's intake. Given this and the model of understanding a referential expression we use, we have a natural way to formally describe how a learner would leverage this type of evidence using other pronouns. Notably, these other pronouns would unambiguously be category NP,[7] since they replace an entire noun phrase (NP) when they are used, as in (9):

(9) Indirect positive unambiguous (INDIRUNAMB) example
Look at the cute penguin. I want to hug it.
antecedent of *it* = $[_{\text{NP}}\ \textit{the}\ [_{\text{N}'}\ \textit{cute}\ [_{\text{N}'}\ [_{\text{N}^0}\ \textit{penguin}]]]]$

The utility of these indirect positive data relates to the learner's preferences when encountering pronouns that have more than one potential antecedent, such as in DirRefSynAmb data like (7). In particular, if the learner tracks how often referents in general have the mentioned property, these indirect positive data will increase the learner's bias for a referent having the property. This is because all IndirUnamb data by necessity include the mentioned property in the NP antecedent (e.g., in (9), *cute* is included) and so the referent must have that property (e.g., in (9), the referent is a CUTE PENGUIN). This in turn could cause the learner to prefer that referents generally have the mentioned property and so, in ambiguous cases, the learner would then prefer an antecedent that includes that modifier (e.g., selecting *red bottle* instead of just *bottle* for the antecedent in (7)).

[7] In fact, it turns out that *one* can also have an NP antecedent. See Appendix B for discussion.

4.3.3 Corpus analysis of data types

We conducted a corpus analysis of the Brown-Eve corpus (Brown, 1973) from the CHILDES database (MacWhinney, 2000), since it included naturalistic speech directed to a fairly young child (starting at the age of 18 months and continuing through 27 months). The 17,521 child-directed utterances included 2,874 that contained a pronoun, with the distribution shown in Table 2. For each of these 2,874 data points, we identified whether it was one of the four data types described above (Table 3), or was instead uninformative for our learners. Uninformative data include ungrammatical uses of anaphoric *one*, uses of *one* where no potential antecedent was mentioned in the previous linguistic context (e.g., *Do you want one?* with no previous linguistic context), and uses of pronouns as NPs where the antecedent did not contain a modifier (e.g., *Mmm – a cookie. Do you want it?*). This last kind of data is viewed as uninformative because NP data points can only help indicate whether a mentioned property is included in the antecedent (see discussion above in 4.3.2). If no property is mentioned, then the data point is uninformative as to whether the antecedent must contain the mentioned property.

Notably, we did not find any DirUnamb data, which accords with Baker's original intuition that such data are scarce. This is also in line with the corpus analysis of Lidz et al. (2003), who found that 0.2% of anaphoric *one* data points were DirUnamb data points – interestingly, rarer even than the ungrammatical uses, which comprised 0.4%. The DirRefSynAmb data are fairly rare as well (again aligning with the corpus analysis of Lidz et al. (2003)), while the DirSynAmb and IndirUnamb data appear much more frequently. Still, the majority of the data would be viewed as uninformative about the aspects of anaphoric *one* under consideration.

Table 2: Pronoun frequencies in Brown-Eve corpus utterances.

Pro	it	one	he	them	she	they	her	him	ones	his	its	itself	their	himself	**Total**
Freq	1536	347	321	183	165	142	80	76	9	6	3	3	2	1	2874

Table 3: Data type frequencies. Percentages are calculated with respect to all data points containing a pronoun in the corpus (2874).

Data type	**Brown-Eve**
DirUnamb	0.00%
DirRefSynAmb	0.66%
DirSynAmb	7.52%
IndirUnamb	8.42%
Uninformative	83.4%

4.4 Specifying the initial state

The initial state for the English anaphoric *one* learner has traditionally been thought to include the following basic syntactic knowledge (e.g., Baker, 1978; Hornstein & Lightfoot, 1981; Lightfoot, 1982a; Crain, 1991):

(10) Prior knowledge in the initial state when learning about English anaphoric *one*

a. **SynCat**: Syntactic categories exist, in particular N^0, N', and NP.

b. **A=SameCat**: Anaphoric elements like *one* take linguistic antecedents of the same category.

Each proposed learning strategy has then added additional biases and/or capabilities. We first review prior strategies and then describe the indirect positive evidence strategy we propose.

4.4.1 Prior learning strategy proposals

The original strategy considered for this problem (Baker, 1978) assumed that only direct positive evidence was relevant, and that only unambiguous data were informative. This direct positive unambiguous strategy (**DirUnamb**) added the following to the initial state:

(11) DirUnamb updated initial state

a. **DirPos**: Use direct positive evidence for learning *one*.

b. **Unamb**: Only unambiguous evidence for *one* is useful.

Baker (1978) assumed these data were too sparse for a learner to make the correct generalization about *one*, and subsequent corpus analyses (LWF's and our own in section 4.3.3) verified that these data were far below what theory-neutral estimates would suggest is necessary for acquisition by 18 months (Legate & Yang, 2002; Yang, 2004, 2012).

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, utilizing this prior linguistic knowledge.[8] We call this the **DirUnamb + N'** strategy, and it updates the initial state as follows:

(12) DirUnamb + N' updated initial state

a. **DirPos**: Use direct positive evidence for learning *one*.

b. **Unamb**: Only unambiguous evidence for *one* is useful.

c. ***one*$\neq$N^0**: *One* is not category N^0.

Regier and Gahl (2004) investigated a learning strategy that assumed children used probabilistic inference, and so were not restricted to learning only from unambiguous data. Instead, this learner leveraged DirRefSynAmb data by tracking how often the referent had the property that was mentioned (e.g., when *red* was mentioned, was the referent just a BOTTLE or specifically

[8] 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*].

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*). More specifically, the direct positive evidence of DirRefSynAmb data provides indirect negative evidence about *one* because a data point where the referent does not have the property mentioned in the potential antecedent (e.g., *"Look – a red bottle! Look – another one!"*, where *one*'s referent is a purple bottle) keeps NOT appearing. A probabilistic learner can take advantage of this suspicious coincidence.

From a learning standpoint, if the learner determines that the antecedent includes the modifier (e.g., *red bottle*), this indicates that *one*'s antecedent is N′, since N^0 cannot include modifiers. *One* would then be N′ too, since it is the same category as its antecedent. The probabilistic learning strategy of Regier and Gahl (2004) did quite well, quickly converging on the adult generalizations when only DirRefSynAmb data were available in the input.

Pearl and Lidz (2009) noted that since children were learning the syntactic category of *one*, an "equal-opportunity" (**EO**) probabilistic learner able to extract information from ambiguous data would also view DirSynAmb data as informative. Interestingly, they found that a probabilistic learner utilizing both DirRefSynAmb and DirSynAmb data (a **DirEO** learner) makes the wrong generalization about *one*'s syntactic category, preferring it to be N^0. Since the harmful DirSynAmb data far outnumber the helpful DirUnamb and DirRefSynAmb data combined (about 20 to 1 in Pearl and Lidz (2009)'s corpus analysis and 11 to 1 in ours), Pearl and Lidz (2009) proposed that a successful probabilistic learner would need to filter out the DirSynAmb data. We term this the **DirFiltered** learner, since it learns from direct positive evidence but filters out some of the ambiguous data. The initial state updates for the successful DirFiltered and unsuccessful DirEO strategies are shown in (13) and (14)

(13) DirFiltered updated initial state
   a. **DirPos**: Use direct positive evidence for learning *one*.
   b. **ProbInf**: Use the probabilistic inference capability so that indirect negative evidence can be leveraged.
   c. **–DirSynAmb**: Do not learn from DirSynAmb data.

(14) DirEO updated initial state
   a. **DirPos**: Use direct positive evidence for learning *one*.
   b. **ProbInf**: Use the probabilistic inference capability so that indirect negative evidence can be leveraged.

4.4.2 Current proposal: Indirect positive evidence

Here we consider a learning strategy that expands the data intake, rather than restricting it. In particular, we propose a probabilistic learning strategy that uses both direct positive evidence and indirect negative evidence, while also learning from the indirect positive evidence that comes from other pronoun data that have linguistic antecedents (**IndirPro**).

(15) IndirPro updated initial state
   a. **DirPos**: Use direct positive evidence for learning *one*.

b. **ProbInf**: Use the probabilistic inference capability so that indirect negative evidence can be leveraged.
c. **+OtherPro**: Use the indirect positive evidence coming from other pronoun data.

4.4.3 Learning strategy comparison

The knowledge, biases, and capabilities for all strategies are summarized in Table 4, and the data each strategy uses is summarized in Table 5. Table 6 illustrates how much data each learning strategy would view as informative, based on the corpus analysis in Table 3. This analysis draws on the estimated number of sentences children hear from birth until 18 months (Akhtar et al., 2004), which is approximately 1,000,000. From this, we calculate that the learner hears approximately 36,500 data points containing a referential pronoun between 14 and 18 months.[9] Perhaps most strikingly, the strategies relying only on direct positive unambiguous data have no data to learn from at all.

Table 4: Knowledge, capabilities, and biases that differ in the learner's initial state for each learning strategy described. Knowledge and learning biases shared by all strategies (**SynCat**, **A=SameCat**, **DirPos**) are not shown.

	Unamb	***one*≠N^0**	**ProbInf**	**–DirSynAmb**	**+OtherPro**
DirUnamb	✓				
DirUnamb + N′	✓	✓			
DirFiltered			✓	✓	
DirEO			✓		
IndirPro			✓		✓

5 Learning about *one*

We now present an online probabilistic learning framework that uses the different kinds of information available in the data types described above. We will use this framework to evaluate the different proposed learning strategies.

5.1 Formalizing target knowledge

The two components of the target knowledge for interpreting anaphoric *one* can be formalized using the model of understanding a referential expression in Figure 3.

[9] Specifically, 2,874 of the 17,521 utterances from the Eve corpus were referential data points containing a pronoun (≈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 2,874/17,521 to get the number of referential pronoun data points heard during this period, which is 36,451, and we round that to 36,500.

Table 5: Data intake for different learning strategies.

Data type	Example	Learning strategies using these data
DirUnamb	*Look– a red bottle! There isn't another one here, though.*	DirUnamb, DirUnamb + N′, DirFiltered, DirEO, IndirPro
DirRefSynAmb	*Look– a red bottle! Oh, look– another one!*	DirFiltered, DirEO, IndirPro
DirSynAmb	*Look– a bottle! Oh, look– another one!*	DirEO, IndirPro
IndirUnamb	*Look a red bottle! I want it/one.*	IndirPro

Table 6: Data intake for different learning strategies, derived from the Brown-Eve corpus analysis.

	DirUnamb	**DirUnamb + N′**	**DirFiltered**	**DirEO**	**IndirPro**
DirUnamb	0	0	0	0	0
DirRefSynAmb	0	0	242	242	242
DirSynAmb	0	0	0	2743	2743
IndirUnamb	0	0	0	0	3073
Uninformative	36500	36500	36258	33515	30442

(16) Target state knowledge

a. **Syntactic**: When the syntactic environment indicates *one* is smaller than an NP (**env=<NP**), it is category N′ (**C=N′**).

b. **Referential**: When an object in the current context has the mentioned property (**o-m=yes**), that property is included in the antecedent of *one* (**i=yes**).

Importantly for the update equations we will use in the online probabilistic learning framework, the variables of interest (**C** and **i**) can only take on two values in these situations: **C** $\in \{\text{N}', \text{N}^0\}$ when **env=<NP** and **i** $\in \{\text{yes}, \text{no}\}$ when **o-m=yes**. Our modeled learner will determine the probability associated with both syntactic and referential knowledge, specifically $p(C{=}N' \mid env{=}{<}NP)$ and $p(i{=}yes \mid o\text{-}m{=}yes)$. We represent the probability of the syntactic category being N′ as $p_{N'}$ and the probability of the antecedent including the mentioned property as p_{incl}. If the target representation of *one* has been learned for the intended context, both probabilities should be near 1.

5.2 Learning target knowledge

We follow the update methods in Pearl and Lidz (2009), and use equation (17) 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 $p_{N'}$ and p_{incl} as parameters in binomial distributions, given that each variable takes on only two values, as noted above.

$$p_x = \frac{\alpha + d_x}{\alpha + \beta + D_x}, \alpha = \beta = 1 \tag{17}$$

Parameters α and β represent a very weak prior when set to 1.[10] 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 (18), and then p_x is updated using equation (17). 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 $\phi_x < 1$.

$$d_x = d_x + \phi_x \tag{18a}$$
$$D_x = D_x + 1 \tag{18b}$$

Probability $p_{N'}$ is updated for DirUnamb data, DirRefSynAmb ambiguous data, and DirSynAmb data only (IndirUnamb data indicate the category is not <NP (**env=NP**), and so are uninformative for $p_{N'}$). Probability p_{incl} is updated for DirUnamb data, DirRefSynAmb data, and IndirUnamb data only (DirSynAmb data do not mention a property, and so are uninformative for p_{incl} since **o-m=*N/A***).

The value of ϕ_x depends on data type. We can derive the values of $\phi_{N'}$ and ϕ_{incl} by doing probabilistic inference over the graphical model in Figure 3. The details of this inference are described in Appendix C. Both $\phi_{N'}$ and ϕ_{incl} 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 Pearl and Lidz (2009) used for m and n, which are m = 1 and n = 2.9.[11]

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., Pearl and Lidz (2009) 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

[10] 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.

[11] The actual numbers Pearl and Lidz (2009) 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.

variety of values ranging from 2 to 49 in Appendix F. A value of $s = 10$ makes the learner believe it is a very suspicious coincidence that the referent just happens to have the mentioned property.

Table 7 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 $s = 10$.

Table 7: The value of $p_{N'}$ and p_{incl} after one data point is seen at the beginning of learning when $p_{N'}$ = p_{incl} = 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	$p_{N'}$	p_{incl}
DirUnamb	0.67	0.67
DirRefSynAmb	0.59	0.53
DirSynAmb	0.48	0.50
IndirUnamb	0.50	0.67

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

For DirRefSynAmb data, both $p_{N'}$ and p_{incl} 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 $p_{N'}$ and p_{incl} (which are both 0.50 initially). After one DirRefSynAmb data point, $p_{N'}$ increases to 0.59, and p_{incl} increases to 0.53. 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 a DirUnamb data point.

DirSynAmb 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 DirSynAmb data that Pearl and Lidz (2009) discovered, where these data cause the learner to believe that *one* is not category N′ when it is smaller than NP. The formal details of why this occurs are described in Appendix D.

IndirUnamb 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}.

5.3 Formalizing and generating target behavior

Previous investigations have focused on learning the target knowledge for anaphoric *one* (Regier & Gahl, 2004; Foraker et al., 2009; Pearl & Lidz, 2009). However, we have empirical data about target behavior in 18-month-olds which we can also use to compare the different learning strategies. A successful learner will generate a familiarity preference in the anaphoric context ("*Look – a red bottle! Now look – do you see another one?*"), and look to the familiar bottle with probability

0.587. This contrasts with the baseline novelty preference when hearing "*Now look – what do you see now?*", where 18-month-olds look to the familiar bottle with probability 0.459.

We can use almost the same graphical model shown in Figure 3 to calculate the probability of the learner looking at the referent that has the mentioned property (e.g., the familiar bottle) in the LWF experimental setup, which we represent as 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. More specifically, given the utterances in the anaphoric context (e.g., "*Look – a red bottle! Now look – do you see another one?*") and two objects present (a familiar one with the mentioned property and a novel one without), we can calculate the probability that the learner looks to the familiar object. This probability depends on the learned values for $p_{N'}$ and p_{incl}.

We describe the formal details of the probabilistic inference involved in calculating p_{beh} in Appendix E.1. This inference involves four free parameters: (i) the two described previously that are related to the syntactic information concerning modifier+noun and noun-only N′ strings, m and n, and (ii) two new parameters that correspond to the baseline and adjusted familiarity looking preferences of 18-month-olds, b and a. The syntactic parameters retain the same empirically-derived values as before (m=1, n=2.9). The looking preference parameters are empirically derived from the LWF experiment, given baseline looking preferences with no referential expression or a noun-only expression like *bottle* and adjusted looking preferences with an anaphoric expression like *another one* or a modifier+noun expression like *red bottle* (b = 0.459, a = 0.587). A learner who can generate the observed toddler behavior should look to the familiar bottle in the anaphoric condition with p_{beh}=0.587.

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 in our modeled learners. It is possible, for example, that children have a different representation, but look at the correct object by chance.

To formally answer this question, we can calculate the probability that the learner has the target representation, given that the learner has produced the target behavior 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. Probability $p_{rep|beh}$ can be calculated by using probabilistic inference over the slightly modified graphical model in Figure 3 that was used for calculating p_{beh}. The formal details of calculating $p_{rep|beh}$ are discussed in Appendix E.2. A learner who has the target representation when generating the target behavior should have $p_{rep|beh}$=1.

6 Results

Table 8 shows the results of the learning simulations over the different input sets with s (the number of properties salient to the learner when interpreting a data point during learning) set to 10. Each learner's input was drawn from the distribution in Table 6. Averages over 1000 runs are reported

for each learning strategy, with standard deviations in parentheses.

Table 8: Probabilities after learning, with s=10. Note that the target value of p_{beh} = 0.587, while all other target values are 1.000.

	DirUnamb	**DirUnamb + N′**	**DirFiltered**	**DirEO**	**IndirPro**
$p_{N'}$	0.500 (<0.01)	1.000	0.991 (<0.01)	0.246 (0.06)	0.368 (0.04)
p_{incl}	0.500 (<0.01)	0.500 (<0.01)	0.963 (<0.01)	0.379 (0.18)	1.000 (<0.01)
p_{beh}	0.475 (<0.01)	0.492 (<0.01)	0.574 (<0.01)	0.464 (0.04)	0.587 (<0.01)
$p_{rep\|beh}$	0.158 (<0.01)	0.306 (<0.01)	0.918 (<0.01)	0.050 (0.11)	0.998 (<0.01)

6.1 Previous learning strategies

A few observations can be made. First, since the DirUnamb learner uses only DirUnamb data in its intake and since these data were not found in our dataset, this learner effectively learns nothing. Thus, the DirUnamb learner remains completely uncertain whether *one* is N′ when it is smaller than NP ($p_{N'}$=0.500) and whether the antecedent includes the mentioned property (p_{incl}=0.500). Given these general non-preferences, it does not generate the target adjusted looking time preference for the LWF experiment (p_{beh}=0.475 instead of 0.587) – it instead retains its novelty preference, and looks less frequently at the familiar bottle. If it happens to look at the familiar bottle, it is fairly unlikely to have the target representation ($p_{rep|beh}$=0.158). Specifically, if the DirUnamb learner looks at the bottle with the mentioned property, it has only a 15.8% of doing so because it has the same antecedent as adults do. Thus, learning from DirUnamb data alone runs into an induction problem, as Baker (1978) (and many others) supposed and we affirm here.

Baker's solution was that the learner had a learning bias involving the knowledge that *one* was not category N^0, which would make it N′ in this context. Thus, the DirUnamb + N′ learner already knows $p_{N'}$=1.000. While this learner has the correct syntactic representation, it still has no data to learn from, and so it learns nothing about whether the antecedent includes the mentioned property (p_{incl}=0.5). Because of this, like the DirUnamb learner, it also does not generate the target familiarity preference for the LWF experiment (p_{beh}=0.492 instead of 0.587) and is fairly unlikely to have the target representation if it happens to do so ($p_{rep|beh}$=0.306). So, this learning strategy appears insufficient to generate the target behavior observed at 18 months, even though it has the target syntactic knowledge.

For the DirFiltered learner, previous studies (Regier & Gahl, 2004; Pearl & Lidz, 2009) found that this learner has a very high probability of acquiring the target representation. We replicate this qualitative result here ($p_{N'}$=0.991, p_{incl}=0.963). In addition, we also observe that this learner can generate a familiarity preference that is nearly as strong as the observed familiarity preference in 18-month-olds (p_{beh}=0.574, which is close to 0.587), and is quite likely to have the target representation when doing so ($p_{rep|beh}$=0.918). This new finding suggests that not only can a learner using this strategy learn the target knowledge state, but it can generate the target behavior and have the target representation when doing so.

For the DirEO 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 ($p_{N'}$=0.246, p_{incl}=0.379). In addition, we also observe that this learner does not generate a familiarity preference (p_{beh}=0.464 instead of 0.587), and is very unlikely to have the target representation if it happens to look to the familiar bottle ($p_{rep|beh}$=0.050). This new finding suggests that not only can a learner using this strategy not learn the target knowledge state, but it also fails to generate the target behavior and does not have the target representation if it happens to do so.

6.2 The indirect positive evidence learning strategy

Turning now to the IndirPro learner, we see that including the indirect positive evidence of IndirUnamb data allows this learner to learn that the antecedent should include the mentioned property (p_{incl}=1.000). This seems intuitively satisfying as this probability is exactly what IndirUnamb data boost. However, this learner also has a moderate dispreference for believing *one* is N′ when it is smaller than an NP ($p_{N'}$=0.368). That is, this learner is inclined to incorrectly believe that *one* is category N^0 in general, which is not the target syntactic knowledge.

Interestingly, this lack of the target syntactic knowledge does NOT prevent the IndirPro learner from generating the observed toddler familiarity preference (p_{beh}=0.587) and having the target representation when doing so ($p_{rep|beh}$=0.998). How can this be?

This behavior 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 IndirPro 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.

In effect, LWF's strict interpretation of their results does not hold – generating target behavior in this context does not necessarily indicate that the learner has the target knowledge in general. Nonetheless, LWF were not mistaken in assuming that learners should have the target representation in this context when they generate the target behavior, as this probability is very high for the IndirPro learner ($p_{rep|beh}$=0.998).

What exactly does this learning outcome mean for the IndirPro learner? First, this learner will succeed in having the target representation whenever a property is mentioned in the potential antecedent (e.g. "*Look – a red bottle!*"). These data include the LWF experimental setup, as well as DirUnamb, DirRefSynAmb, and IndirUnamb data points.

However, when no property is mentioned in the potential antecedent, such as in DirSynAmb data points (e.g., "*Look – a bottle!*"), this learner will not have the target representation. While it will believe the antecedent is, e.g., *bottle*, it will assume that string is category N^0 instead of N′, due to the low probability of $p_{N'}$ and the fact that the high probability of p_{incl} cannot help since no property was mentioned. Nonetheless, this mistake will NOT impede communicative success, since the referent is the same in either case (a BOTTLE). Thus, this mistake is unlikely to be detected by either the learner or the people the learner communicates with.

Still, there are scenarios when the mistake would be detected. In particular, this learner would be perfectly fine with utterances that use *one* as an N^0, such as * *I drank from the edge of the cup while you drank from the one of the bowl*. In contrast, adults who only allow *one* as N′ when it is smaller than NP will not find this grammatical. It is currently unknown when children attain this specific linguistic knowledge about *one*, though grammatical judgment methodology (Ambridge & Rowland, 2013) could likely be used to find out. Once experimental methods identify when children attain this knowledge, we can investigate learning strategies that will allow successful acquisition of that knowledge.

Since it seems that the immature representation would only rarely fail for communication purposes (in particular, for the scenario described above), it may be that children do not attain this knowledge for quite some time. Foraker et al. (2009) demonstrate a successful probabilistic learning strategy for learning that *one* is N′ in general, which is the key difference between the immature and adult representations. This strategy relies on fairly sophisticated conceptual knowledge linked to syntactic representations and draws on indirect negative evidence about how *one* is used when compared to nouns like *edge*. If it turns out that children do not attain the adult representation of *one* until significantly later in development, it may be that they have acquired the conceptual knowledge and links to syntactic representation necessary to use this strategy. So, before 18 months, children could use the IndirPro learning strategy to learn an immature representation, and then switch to Foraker et al. (2009)'s strategy to subsequently learn the adult representation once they attain the knowledge that strategy relies on.

6.3 The impact of s

Interestingly, we find there is a qualitative difference between the behavior of the DirFiltered and DirEO learners and that of the IndirPro learner with respect to s, which determines how suspicious a coincidence a DirRefSynAmb data point is. Results for a range of s values are presented and discussed in more detail in Appendix F, but in brief, there are some values of s which qualitatively change the results for the DirFiltered and DirEO learners. Notably, the DirFiltered learner fails when $s <= 5$, a situation where a DirRefSynAmb data point isn't all that suspicious a coincidence. In contrast, the DirEO learner can succeed when $s >= 20$, a situation where a DirRefSynAmb data point is a very suspicious coincidence. This fluctuating behavior contrasts with the IndirPro learner, whose behavior remains invariant across all s values investigated. Thus, the IndirPro strategy seems more robust to variation in the learning environment. If all other factors are equal, this may be a reason to prefer this strategy. However, if empirical evidence about s's true value can be determined in the future, any strategy that yields success with that s value would be viable.

6.4 Summary of results

Two strategies that are always unsuccessful are those that use only direct positive unambiguous data (DirUnamb, DirUnamb + N′), while the strategy that leverages information from all direct positive data (DirEO) is typically unsuccessful. In contrast, the strategy that filters the data intake down to a subset of the direct positive data (DirFiltered) is typically successful at both reaching

the target knowledge state and generating the target behavior. Still, the DirFiltered strategy's performance does have some variation, unlike the strategy that expands the data intake to include indirect positive evidence coming from other pronouns (IndirPro). The IndirPro strategy always generates the target behavior, though it learns an immature content-sensitive representation of *one* that nonetheless suffices in many contexts.

7 Discussion

7.1 General discussion of results

Through this empirically-grounded computational modeling study, we have identified two learning strategies capable of generating the observable anaphoric *one* behavior in 18-month-olds. One strategy (DirFiltered) restricts the data intake of learners to a subset of the direct positive data and generates this behavior from a knowledge state similar to that of adults, though it is less robust to different learning scenarios. The other strategy (IndirPro) expands the data intake to include all direct positive data as well as some indirect positive evidence coming from other pronouns, and is able to generate the observable behavior without having the adult knowledge state. While this strategy is robust to different learning scenarios, an immature context-dependent representation of anaphoric *one* underlies the observable behavior. This underscores that even if children demonstrate they have the adult interpretation in some contexts, they do not necessarily have the adult representation. Nonetheless, both these learners have clearly made useful syntactic generalizations since they lead to target behavior in 18-month-olds, and it is worthwhile to consider the components of the learning strategies that allowed them to do so.

7.2 Strategy components

In addition to a learning strategy's ability to generate observable behavior, another way to evaluate it is by the components it requires. First, where do these components come from? Second, how task-specific are these components? Theoretically, we may prefer strategies that require fewer components that are both innate and domain-specific, as such components commit us to finding an explanation of how they arose in human biology. We may also prefer strategy components that are useful for learning other knowledge. With this in mind, we discuss possible origins and the general utility of the required components for each successful strategy.

7.2.1 Possible origins

One approach is to begin by assuming all strategy components are innate, and then demonstrate via existence proof how a particular component could arise from other knowledge and experience. That is, "innate" serves as a placeholder until we have a precise model of the process that generates that necessary component (Pearl, 2014). We consider different strategy components below and present some concrete suggestions for how they might be derived.

The two successful strategies share several components while differing on a single component. For the shared components, they each (i) have knowledge of certain syntactic categories, (ii) have

knowledge that anaphoric elements take linguistic antecedents of the same category, (iii) learn from the available direct positive evidence, and (iv) use the probabilistic inference learning ability to leverage indirect negative evidence. For the syntactic category knowledge, it may be possible to derive the appropriate categories using distributional clustering strategies (e.g., frequent frames (Mintz, 2003, 2006)) or other distributional cues (Gerken, Wilson, & Lewis, 2005). It may also be possible to derive the knowledge about anaphoric element antecedents through distributional learning techniques. For example, perhaps a learner could observe the linguistic antecedents of anaphoric elements where the antecedent is unambiguous (e.g., "*Those two penguins are cute – I like them a lot*", with *them* unambiguously referring to *those two penguins*). From this, the learner might determine that anaphoric elements and their antecedents share distributional environments, and so are the same category.

The DirFiltered learner's strategy incorporates an additional bias to filter out a certain kind of ambiguous direct positive data: the DirSynAmb data, such as "*Look – a bottle! Oh look – another one!*". Pearl and Lidz (2009) suggest that this bias could come from a preference for learning only when there is uncertainty about the referent, as opposed to when there is uncertainty about the syntactic category. This preference would cause the learner to ignore these data, since the referent is clear (BOTTLE above), even if the syntactic category is not. One idea for the origin of this bias is that it is derived from some more general principle of communicative efficacy where the learner is particularly attentive when there is ambiguity in comprehension. In particular, if comprehension is "good enough" (Ferreira, Bailey, & Ferraro, 2002), then learners would be unconcerned about improving linguistic knowledge about the utterance. In this case, "good enough" means the correct referent is understood, even if the syntactic category is incorrect.

The IndirPro learner's strategy incorporates an additional bias to expand the data intake to include a type of indirect positive data involving other pronouns: the IndirUnamb data, such as "*Look – a blue bottle! Do you want it?*". To do this, this learner must know that *one* is similar to other pronouns, even though it appears in a syntactic environment that they do not (*another one*, but **another it*). One idea for the origin of this bias is that the learner develops an overhypothesis (Kemp et al., 2007) about how pronouns are used, with *one* being one specific instantiation and other pronouns being other related instantiations of that overhypothesis.

A very important question for future research is clear from the description of these strategy components: for each component that is *possibly* derivable, can we find a way to *actually* derive it from realistic data? This requires us to create a concrete learning model whose target state is the appropriate knowledge or bias in each case (Kol, Nir, & Wintner, 2014; Pearl, 2014). If we can demonstrate how a given component is derived, we can then ask what knowledge, capabilities, and learning biases were necessary to do so – and then investigate where those components might come from, until we identify the core un-derivable components. These are then the innate components necessary for making this linguistic generalization.

In short, at the heart of every learning strategy component is some innate core. An interesting question is then what kind of innate core it is. If it is language-specific, it becomes a concrete proposal for a piece of Universal Grammar that demonstrably helps acquisition (Ambridge, Pine, & Lieven, 2014; Pearl, 2014). If it is domain-general, it is likely to be something that affects cognitive development of all kinds.

7.2.2 Utility

Could a learner use these learning strategy components to construct successful learning strategies for acquiring other linguistic knowledge besides anaphoric *one*? No matter a component's origins, we can still explore whether it would be useful for learning other things by identifying successful learning strategies for other linguistic phenomena and seeing if they make use of that component. We now speculate briefly about the utility of the different strategy components, drawing on empirical evidence where available.

For the shared components, the syntactic category knowledge, though specifically about NP, N′ and N^0 here, seems a fairly fundamental component for learning syntactic knowledge more generally since representations of syntactic knowledge typically assume the syntactic category of the word has already been identified (e.g., any knowledge based on phrase structure). It simply may be that other categorical distinctions are relevant, depending on the specific syntactic knowledge to be learned. Similarly, knowing that anaphoric elements take antecedents of the same category seems fundamental for learning about referential elements more generally. The biases to use direct positive evidence and probabilistic inference have already been shown to be very useful for learning other linguistic knowledge (e.g., Tenenbaum & Griffiths, 2001; Yang, 2004; Xu & Tenenbaum, 2007; Pearl, 2011; Perfors, Tenenbaum, & Regier, 2011).

For the DirFiltered learner, the bias to shrink the data intake and ignore data that are not referentially ambiguous may be a specific instantiation of a bias for communicative efficacy, where learning only occurs when comprehension is not "good enough". This approach works for English anaphoric *one* by filtering out misleading syntactically ambiguous data, and it could possibly allow learners to filter out potentially misleading data for other syntactic phenomena as well.

The bias to expand the data intake and learn from other pronoun data is a specific instantiation of a bias to learn from all informative data, including indirect positive evidence. For this bias, recent studies have already suggested that using indirect positive evidence is a crucial component of successful strategies for learning about both fundamental and fairly sophisticated aspects of syntactic knowledge (hierarchical structure of syntactic representations: Perfors, Tenenbaum, and Regier (2011); syntactic islands: Pearl and Sprouse (2013b, 2013a)). Thus, this component already has demonstrable utility for learning syntactic knowledge more generally.

7.3 Further expansion of the data intake

One useful extension of the current work is to consider if the learner's data intake could be expanded still further to leverage other types of indirect positive evidence. It is certainly possible that many different types of data involving pronouns may be viewed by the learner as relevant, including the evidence we considered uninformative in the current model context (e.g., uses of *one* without a linguistic antecedent like *Do you want one?*). To leverage data of this kind, it is crucial to be very precise about how these data are used. Our model of understanding a referential expression with a pronoun could easily incorporate the indirect positive evidence we examined, as that evidence impacted relevant variables in the model. In general, for any proposal of indirect positive evidence, there must be an explicit linking hypothesis about how that evidence will impact the learner's beliefs, typically instantiated as model variables.

For Bayesian learning models, this part is carried out in the model specification, which defines exactly how a given data point impacts the learner's beliefs. This then determines which data are viewed as relevant and how relevant those data are. Simply put, data that relate to model variables are viewed as relevant and data that do not are effectively ignored. For example, if a model of pronoun interpretation does not include any variables that are impacted by pronouns without linguistic antecedents, an utterance like *Do you want one?* is uninformative. In contrast, if the model includes a variable about how often pronouns appear as category NP in general, this same utterance is informative because it impacts that variable.

We consider the potential expansion of the learner's data intake an exciting area for future research on this acquisition problem, particularly as the full knowledge about how to interpret *one* is far more complex than the one aspect we have focused on here. Still, we note that either intake expansion or intake restriction may be reasonable acquisition approaches, depending on exactly how the intake specification is implemented. In general, a specification that derives from general-purpose learning biases may be theoretically preferable to a specification that requires additional task-specific learning biases. For intake expansion when learning about *one*, a general-purpose bias to learn from informative data can be coupled with a precisely defined learning model to yield a very large intake, as described above. For intake restriction when learning about *one*, a general-purpose bias for communicative efficacy may naturally cause the learner to filter out data that might otherwise be viewed as informative given the learning model. We believe the best way to investigate either approach is similar to what we have done here: implement learning strategies using each approach to see if they work and, when they do, identify what makes them work (Pearl, 2014).

8 Conclusion

We have investigated how children make syntactic generalizations, using the acquisition of knowledge about English anaphoric *one* as a case study. We have applied two core ideas. First, if children leverage any data deemed informative, they may draw on indirect positive evidence during acquisition, expanding their data intake beyond the direct evidence available. Second, we can empirically ground the target state of learning by drawing on behavioral data from children, with the idea that a successful learning strategy should allow the learner to acquire linguistic knowledge capable of generating that target behavior. We have demonstrated that one successful and robust strategy for acquiring certain knowledge about English anaphoric *one* is a probabilistic learning strategy using indirect positive evidence coming from other pronouns. Interestingly, the knowledge underlying this learner's target behavior is an immature context-dependent representation that nonetheless functions quite well in many communicative contexts. Whether the knowledge representations are the target ones or are instead transitory ones, it is important to understand what components comprise the learning strategies that lead to children's observable behavior. To this end, we have provided a concrete framework for investigating learning strategies that draws on empirical results in theoretical, experimental, and computational research. By identifying precisely what children are learning, when they are learning it, and what they're learning it from, we can better understand how they are able to do it so well.

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, Greg Carlson, 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 Stanford 2013 workshop on Cognition and Language, the NYU 2012 Linguistics Colloquium, CogSci2011, and the UChicago 2011 workshops on Language, Cognition, and Computation and Language, Variation, and Change. All errors are, of course, are own and not at all their fault. In addition, this research was supported by NSF grants BCS-0843896 and BCS-1347028 to LP.

References

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

Ambridge, B., Pine, J., & Lieven, E. (2014). Child language acquisition: Why Universal Grammar doesn't help. *Language*, *90*(3), e53–e90.

Ambridge, B., Pine, J., Rowland, C., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47–62.

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *Wiley Interdiscplinary Reviews in Cognitive Science*, *4*(2), 149–168.

Anderson, S. R. (2013). 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., & McCarthy, J. (1981). *The logical problem of language acquisition*. Cambridge, MA: 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.

Bowerman, M. (1983). How do children avoid constructing an overly general grammar in the absence of feedback about what is not a sentence? *Papers and Reports on Child Language Development*, *22*, 23–25.

Bowerman, M. (1988). The 'No Negative Evidence' problem: How do children avoid constructing an overly general grammar? In J. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Oxford, England: Blackwell.

Boyd, J., & Goldberg, A. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, *87*(1), 55–83.

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar: a theoretical symposium* (pp. 153–186). New York: Academic Press.

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

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition on child speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). New York: Wiley.

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.

Clark, A., & Lappin, S. (2009). Another Look at Indirect Negative Evidence. In *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition* (pp. 26–33). Stroudsburg, PA: Association for Computational Linguistics.

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.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

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

Fodor, J. D., & Crain, S. (1987). Simplicity and generality of rules in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 35–63). Hillsdale, NJ: Erlbaum.

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.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

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

Goldberg, A. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, *22*(1), 131–153.

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.

Grimshaw, J., & Pinker, S. (1989). Positive and negative evidence in language acquisition. *Behavioral and Brain Sciences*, *12*, 341.

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.

Huang, C.-T. J. (1982). *Logical relations in Chinese and the theory of grammar*. Unpublished doctoral dissertation, Cambridge, MA.

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.

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language*, *41*(1), 176–199.

Lasnik, H. (1989). On certain substitutes for negative data. In R. Matthews & W. Demopoulos (Eds.), *Learnability and linguistic theory* (pp. 89–105). Dordrecht: Kluwer/Academic Press.

Lasnik, H., & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry*, *15*, 235–289.

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* (pp. 168–182). Oxford, UK: Oxford University Press.

Lidz, J., & Gagliardi, A. (2015). How Nature Meets Nurture: Universal Grammar and Statistical

Learning. *Annual Review of Linguistics*, *1*(1), 333–352.
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.
Marcus, G. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.
Marcus, G. (1999). Language acquisition in the absence of explicit negative evidence: can simple recurrent networks obviate the need for domain-specific learning devices? *Cognition*, *73*, 293–296.
McNeill, D. (1996). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The genesis of language* (pp. 15–84). Cambridge, MA: MIT Press.
Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.
Mintz, T. (2006). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). Oxford: Oxford University Press.
Niyogi, P., & Berwick, R. C. (1996). A language learning model or finite parameter spaces. *Cognition*, *61*, 161–193.
Omaki, A., & Lidz, J. (2014). Linking Parser Development to Acquisition of Syntactic Knowledge. *Language Acquisition*, *xx*, xxx–xxx. doi: 10.1080/10489223.2014.943903
Payne, J., Pullum, G., Scholz, B., & Berlage, E. (2013). Anaphoric *one* and its implications. *Language*, *90*(4), 794–829.
Pearl, L. (2007). *Necessary Bias in Natural Language Learning*. Unpublished doctoral dissertation, University of Maryland, College Park, College Park, MD.
Pearl, L. (2011). When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition*, *18*(2), 87–120.
Pearl, L. (2014). Evaluating learning strategy components: Being fair. *Language*, *90*(3), e107–e114.
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. (2013a). Computational Models of Acquisition for Islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental Syntax and Islands Effects* (pp. 109–131). Cambridge: Cambridge University Press.

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

Pearl, L., & Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development*, *3*(1), 43–72.

Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction ot bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

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.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure.* Cambridge, MA: MIT Press.

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*, *31*, 949–953.

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

Ramscar, M., Dye, M., & McCauley, S. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*(4), 760–793.

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.

Roeper, T. (1981). On the deductive model and the acquisition of productive morphology. In C. Baker & J. McCarthy (Eds.), *The logical problem of language acquisition.* Cambridge, MA: MIT Press.

Rohde, D., & Plaut, D. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*, 67–109.

Sakas, W. (2003). A Word-Order Database for Testing Computational Models of Language Ac-

quisition. 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*.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*(2), 245-272.

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. (2012). Computational models of syntactic acquisition. *WIREs Cognitive Science*, *3*, 205–213.

Zwicky, A. (1970). A double regularity in the acquisition of English verb morphology. In *Working Papers in Linguistics 4.* Columbus, Ohio: The Ohio State University.

A Formal description of data types

Table 9 formalizes the properties of each of the data types with respect to the model of understanding a referential expression in Figure 3. It can be easily observed where the ambiguities arise for each data type, based on the variables that have more than one value.

Table 9: Data types and variable values. Observable variables are in **bold**. Multiple values indicate ambiguity for that variable.

Variable	DirUnamb	DirRefSynAmb	DirSynAmb	IndirUnamb
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

DirUnamb data like "*Look – a red bottle. There doesn't seem to be another one here, though*" when a red bottle and a purple bottle are present 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**) – specifically, the intended referent has the mentioned property, even though it's not physically present (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 DirUnamb data, DirRefSynAmb data like "*Look – a red bottle. Oh, look – another one*" when two red bottles are present 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**) – specifically, 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.

DirSynAmb data like "*Look – a bottle. Oh, look – another one*" 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.

IndirUnamb data like "*Look – a red bottle. I want it*" 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**) – specifically, 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.

B Data that use *one* as an NP

There are data demonstrating that *one* can also have an NP antecedent, as in (17). Though these data do involve *one*, they have not traditionally been considered as part of the direct positive evidence when learning whether *one* is N′ or N^0 in certain contexts because *one*'s syntactic category is unambiguously NP in these data. In fact, some theoretical analyses have considered these uses a different instance of *one* altogether (i.e., the determinative usage, rather than the regular common count noun used in our other anaphoric examples (Payne et al., 2013)). We classify them as indirect positive evidence here, since they will function the same way as indirect positive evidence coming from other pronouns.

(17) Indirect positive unambiguous (INDIRUNAMB) example involving *one*
Look! A red bottle. I want one.
antecedent of *one* = [$_{NP}$ *a* [$_{N'}$ *red* [$_{N'}$ [$_{N^0}$ *bottle*]]]]

We emphasize 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 the utterances in (2), (6), and (7), and (8)). This shows that *one* clearly has some categorical flexibility, since it can function as both NP and smaller than NP (or at least has instances that can do one or the other (Payne et al., 2013)). 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 *one* can alternate between NP and N′ in a particular context. Instead, in (17) it is always NP, while in direct positive unambiguous (DirUnamb) utterances like (6), 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 (2), (7), and (8).

C Deriving $\phi_{N'}$ and ϕ_{incl}

The values $\phi_{N'}$ and ϕ_{incl} are used for updating $p_{N'}$ and p_{incl}, respectively, which are the probabilities associated with the target syntactic ($p_{N'}$) and referential (p_{incl}) representation for anaphoric *one*. We can derive the values of $\phi_{N'}$ and ϕ_{incl} by doing probabilistic inference over the graphical model in Figure 3.

C.1 $\phi_{N'}$

$\phi_{N'}$ uses the expanded equation in (1), which calculates the probability that the syntactic category is N′ (**C=N′**) when the syntactic environment indicates the 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**.

$$\phi_{N'} = p(C = N' | env =< NP) \tag{1a}$$

$$= \frac{p(C = N', env =< NP)}{p(env =< NP)} \tag{1b}$$

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

The value of $\phi_{N'}$ depends on data type. When $\phi_{N'}$ is calculated for DirUnamb data using equation (1), it can be shown that $\phi_{N'}$=1, which is intuitively satisfying since these data unambiguously indicate that the category is N′ when the syntactic environment is <NP. When $\phi_{N'}$ is calculated for DirRefSynAmb data using (1), it can be shown that $\phi_{N'}$ is equal to (2):

$$\phi_{N' DirRefSynAmb} = \frac{rep_1 + rep_2}{rep_1 + rep_2 + rep_3} \tag{2}$$

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, 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*].

When $\phi_{N'}$ is calculated for DirSynAmb data using equation (1), it can be shown that $\phi_{N'}$ is equal to (4):

$$\phi_{N' DirSynAmb} = \frac{rep_4}{rep_4 + rep_5} \tag{4}$$

where

$$rep_4 = p_{N'} * \frac{n}{m+n} \tag{5a}$$

$$rep_5 = 1 - p_{N'} \tag{5b}$$

The quantities in (5) 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 (4) contains the representation that has *one*'s category as N′, while the denominator contains both possible representations.

C.2 ϕ_{incl}

ϕ_{incl} uses the expanded equation in (6), 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**.

$$\phi_{incl} = p(i = yes | \textit{o-m} = yes) \tag{6a}$$

$$= \frac{p(i = yes, \textit{o-m} = yes)}{p(\textit{o-m} = yes)} \tag{6b}$$

$$= \frac{\sum_{O,A,det,mod,C,Pro,env,R,m} p(i = yes, \textit{o-m} = yes)}{\sum_{O,A,det,mod,C,Pro,env,R,i,m} p(\textit{o-m} = yes)} \tag{6c}$$

The value of ϕ_{incl} also depends on data type. When ϕ_{incl} is calculated for DirUnamb and IndirUnamb data using (6), 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 DirRefSynAmb data using (6), it can be shown that ϕ_{incl} is equal to (7):

$$\phi_{incl} = \frac{rep_1}{rep_1 + rep_2 + rep_3} \tag{7}$$

where rep_1, rep_2, and rep_3 are the same as in (3). Equation (7) is intuitively satisfying as only rep_1 corresponds to a representation where the property is included in the antecedent.

D DirSynAmb data effects

Pearl and Lidz (2009) discovered that DirSynAmb 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 $p_{N'}$ = p_{incl} = 0.50) because the representation using syntactic category N^0 (rep_5 above in section C.1) at that point has a higher probability than the representation using category N′ (rep_4 above in section C.1).

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 probabilistic 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 DirSynAmb data cause a probabilistic learner to (incorrectly) favor the N^0 category over the N′ category.

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

E.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 3 modified to have **O** as an inferred variable, as shown in the equations in (8).

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

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

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

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

$$p_{beh} = \frac{rep_{1f} + rep_{2f} + rep_{3f}}{rep_{1f} + rep_{1n} + rep_{2f} + rep_{2n} + rep_{3f} + rep_{3n}} \quad (9)$$

where rep_{1f}, rep_{1n}, rep_{2f}, rep_{2n}, rep_{3f}, and rep_{3n} are defined as in (10).

$$rep_{1f} = p_{N'} * \frac{m}{m+n} * p_{incl} * a \quad (10a)$$

$$rep_{1n} = p_{N'} * \frac{m}{m+n} * p_{incl} * (1-a) \quad (10b)$$

$$rep_{2f} = p_{N'} * \frac{n}{m+n} * (1-p_{incl}) * b \quad (10c)$$

$$rep_{2n} = p_{N'} * \frac{n}{m+n} * (1-p_{incl}) * (1-b) \quad (10d)$$

$$rep_{3f} = (1-p_{N'}) * (1-p_{incl}) * b \quad (10e)$$

$$rep_{3n} = (1-p_{N'}) * (1-p_{incl}) * (1-b) \quad (10f)$$

m = 1 and n = 2.9, as before. The variables a and b correspond to the adjusted and baseline familiarity preferences, respectively, of toddlers in the LWF experiment, with a=0.587 and b=0.459.

Adjusted refers to the preference when the referring expression itself involves a modifier (*Look, a red bottle – do you see another red bottle?*) or the potential antecedent involves modifier (*Look, a red bottle – do you see another one?*). Baseline refers to the preference when the referring expression itself does not involve a modifier (*Look, a red bottle – do you see another bottle?*) or no referring expression is used (*Look, a red bottle – what do you see now?*). The looking time results demonstrate 18-month-olds had a baseline novelty preference which was overcome when a referring expression was used that contained a modifier or whose potential antecedent contained a modifier.

As before, the quantities in (10) 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 looks to the object with the mentioned property (the familiar object) with probability a (rep_{1f}) and looks to the object without the mentioned property (the novel object) with probability $1 - a$ (rep_{1n}). For the incorrect representations (rep_2 and rep_3) where the antecedent string is just the noun (e.g., *bottle*), the learner can believe the category is either N' (rep_2) or N^0 (rep_3). In either case, the learner uses the baseline preferences, and looks to the familiar object with probability b (rep_{2f}, rep_{3f}) and the novel object with probability $1 - b$ (rep_{2n}, rep_{3n}). The numerator of (9) represents all the outcomes where the learner looks to the object with the mentioned property (the familiar object), while the denominator also includes the three outcomes where the learner looks to the novel object.

E.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 target 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 3, as shown in (11).

$$p_{rep|beh} = p(A = \textit{red bottle}, i = yes, det = no, mod = yes, C = N' | R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes, O = \text{O-M}) \quad (11a)$$

$$= \frac{p(A=\textit{red bottle}, i=yes, det=no, mod=yes, C=N', R=\textit{another one}, Pro=one, env=<NP, m=yes, \textit{o-m}=yes, O=\text{O-M})}{\sum_{A,i,det,mod,C} p(R = \textit{another one}, Pro = one, env =< NP, m = yes, \textit{o-m} = yes, O = \text{O-M})} \quad (11b)$$

When $p_{rep|beh}$ is calculated, it can be shown that it is equal to (12).

$$p_{rep|beh} = \frac{rep_{1f}}{rep_{1f} + rep_{2f} + rep_{3f}} \quad (12)$$

where rep_{1f}, rep_{2f}, and rep_{3f} are calculated as in (10). More specifically, given that the object with the mentioned property has been looked at (whether the relevant antecedent includes the modifier (rep_{1f}) or not (rep_{2f} and rep_{3f})), we calculate the probability that the look is due to the target representation (rep_{1f}).

F Simulation results for different values of s

Table 10 shows the results of the learning simulations over the different input sets with 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 DirUnamb and DirUnamb + N′ learners, the performance of the learners depends to some degree on the value of s. This is to be expected as both of the DirUnamb learners use only DirUnamb data in their intake, and since these data were not found in our dataset, this learner effectively learns nothing no matter what the value of s.

When we examine the results for the IndirPro 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}=0.998−1.000) and has a moderate dispreference for believing *one* is N′ when it is smaller than an NP ($p_{N'}$=0.342−0.376), no matter what the value of s.

For both the DirFiltered and DirEO 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 DirFiltered 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}=0.683); when s=2, the learner is inclined to believe *one* is N^0 ($p_{N'}$=0.340) and is nearly certain that the mentioned property should NOT be included in the antecedent (p_{incl}=0.020). Similarly, when s=7 or above, the learner reliably reproduces the observed infant behavior (p_{beh}=0.557−0.585) and likely has the target representation when looking to the familiar bottle ($p_{rep|beh}$=0.807−0.985). Yet, when s has lower values, the results are quite different (s=5: p_{beh}=0.511, $p_{rep|beh}$=0.468; s=2: p_{beh}=0.459, $p_{rep|beh}$=0.002).

If we examine the DirEO 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′ ($p_{N'}$=0.344−0.366) occurs with a strong preference for including the mentioned property in the antecedent (p_{incl}=0.931−0.987). However, for $s \leq 10$, low values of $p_{N'}$ occur with low values of p_{incl} ($p_{N'}$=0.136−0.246, p_{incl}=<0.010−0.379). 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

Table 10: Probabilities after learning, using different values of s, which is the number of properties salient to the learner when interpreting a data point. Note that the target value of p_{beh} = 0.587. All other target values are 1.000.

	Prob	**DirUnamb**	**DirUnamb + N′**
$s = 2, 5, 7, 10, 20, 49$	$p_{N'}$	0.500 (<0.01)	1.000
	p_{incl}	0.500 (<0.01)	0.500 (<0.01)
	p_{beh}	0.475 (<0.01)	0.492 (<0.01)
	$p_{rep\|beh}$	0.158 (<0.01)	0.306 (<0.01)

	Prob	**DirFiltered**	**DirEO**	**+IndirPro**
$s = 2$	$p_{N'}$	0.340 (<0.01)	0.136 (<0.01)	0.342 (0.03)
	p_{incl}	0.020 (<0.01)	0.010 (<0.01)	0.998 (<0.01)
	p_{beh}	0.459 (<0.01)	0.459 (<0.01)	0.584 (<0.01)
	$p_{rep\|beh}$	0.002 (<0.01)	0.000 (<0.01)	0.980 (<0.01)
$s = 5$	$p_{N'}$	0.942 (<0.01)	0.159 (0.02)	0.362 (0.04)
	p_{incl}	0.683 (<0.01)	0.037 (0.01)	0.999 (<0.01)
	p_{beh}	0.511 (<0.01)	0.459 (<0.01)	0.586 (<0.01)
	$p_{rep\|beh}$	0.468 (<0.01)	0.002 (<0.01)	0.992 (<0.01)
$s = 7$	$p_{N'}$	0.984 (<0.01)	0.185 (0.03)	0.367 (0.04)
	p_{incl}	0.906 (<0.01)	0.102 (0.05)	0.999 (<0.01)
	p_{beh}	0.557 (<0.01)	0.460 (<0.01)	0.586 (<0.01)
	$p_{rep\|beh}$	0.807 (<0.01)	0.007 (0.01)	0.993 (<0.01)
$s = 10$	$p_{N'}$	0.991 (<0.01)	0.246 (0.06)	0.368 (0.04)
	p_{incl}	0.963 (<0.01)	0.379 (0.18)	1.000 (<0.01)
	p_{beh}	0.574 (<0.01)	0.464 (0.04)	0.587 (<0.01)
	$p_{rep\|beh}$	0.918 (<0.01)	0.050 (0.11)	0.998 (<0.01)
$s = 20$	$p_{N'}$	0.994 (<0.01)	0.344 (0.05)	0.373 (0.04)
	p_{incl}	0.987 (<0.01)	0.931 (0.03)	1.000 (<0.01)
	p_{beh}	0.582 (<0.01)	0.532 (0.07)	0.587 (<0.01)
	$p_{rep\|beh}$	0.971 (<0.01)	0.626 (0.11)	1.000 (<0.01)
$s = 49$	$p_{N'}$	0.995 (<0.01)	0.366 (0.05)	0.376 (0.05)
	p_{incl}	0.993 (<0.01)	0.987 (<0.01)	1.000 (<0.01)
	p_{beh}	0.585 (<0.01)	0.573 (0.02)	0.587 (<0.01)
	$p_{rep\|beh}$	0.985 (<0.01)	0.912 (0.02)	1.000 (<0.01)

do here – specifically, because p_{incl} is so high, there is a higher probability of generating the LWF familiarity preference (p_{beh}=0.532−0.573) and a stronger probability of having the target representation when looking at the familiar bottle ($p_{rep|beh}$=0.626−0.912). This is the same qualitative behavior we found in the IndirPro learner. However, the DirEO learner differs by failing to exhibit this behavior this when $s \leq 10$: The learner does not generate the LWF behavior

(p_{beh}=0.459−0.460) and is unlikely to have the target representation if it happens to look at the familiar bottle ($p_{rep|beh}$=<0.000−0.050).

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 referential information in the model (Figure 3), 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 Regier and Gahl (2004) and Pearl and Lidz (2009). This is reflected in the update equations for the DirRefSynAmb data, where both $\phi_{N'}$ and ϕ_{incl} involve the current values of $p_{N'}$ and p_{incl}, as do all the equations corresponding to the probabilities of the different antecedent representations (recall equation (3)). This means that there is an inherent linking between these two probabilities when DirRefSynAmb 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 DirRefSynAmb data (and so increase $p_{N'}$ more). 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 $p_{N'}$ and p_{incl} benefit strongly – each DirRefSynAmb data point functions almost as if it were a DirUnamb 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 DirRefSynAmb data point leads to a p_{incl} of 0.409. This causes subsequent DirRefSynAmb 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 DirRefSynAmb 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 DirRefSynAmb data is significantly lessened, and can even cause these data to have a detrimental effect on learning. This is why the DirFiltered learner fails for low s values. The situation is worse when DirSynAmb data are included in the mix, as for the DirEO learner – not only are the DirRefSynAmb data insufficiently powerful, but the DirSynAmb data cause $p_{N'}$ to plummet.

Notably, when IndirUnamb data are added into the mix for the IndPro 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 DirRefSynAmb data. Due to the linking between p_{incl} and $p_{N'}$ in the DirRefSynAmb data update, the high p_{incl} value will cause DirRefSynAmb data points to act as if they were DirUnamb data points, and so $p_{N'}$ is also increased. This is why the IndirPro learner is not susceptible to changes in its behavior when s changes. Still, because this benefit to $p_{N'}$ only occurs when DirRefSynAmb data are encountered, and these are relatively few, the final $p_{N'}$ value is still fairly low (0.342−0.376). If we remove the DirRefSynAmb data from the IndirPro learner's dataset (i.e., it only encounters DirSynAmb and IndirUnamb data points, as well as uninformative data points), we can see a final $p_{N'}$ that is much lower ($p_{N'}$=0.130), even though p_{incl}=1.000.

To summarize, the behavior of the learner that uses indirect positive evidence is robust because it can leverage IndirUnamb data to compensate for (or further enhance the effectiveness of) the DirRefSynAmb data. In contrast, learners who are restricted to only direct positive data are greatly

affected by how suspicious a coincidence DirRefSynAmb data points are. Our results are similar to previous results for the DirFiltered and DirEO learners for certain values of s. However, because of the way referential 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 and Lidz (2009) did with their integrated probabilistic learning model for the DirEO learner with high values of s. We also find low values of $p_{N'}$ and p_{incl} for the DirFiltered 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. For all learners and all s values, the probabilities converge to their final values within the first few hundred data points. 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.

G A different knowledge representation

Another theoretical representation of noun phrase syntax assumes different syntactic categories than the ones in the representation we examined here. In particular, our representation (Chomsky, 1970; Jackendoff, 1977) incorporated 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 (18a). However, an alternate representation of noun phrases is available (Bernstein, 2003; Longobardi, 2003), shown in (18b) 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′.

(18) Theoretical representations for noun phrase syntax

a. [$_{NP}$ *a* [$_{N'}$ *delicious* [$_{N'}$ [$_{N^0}$ *bottle*] [$_{PP}$ *of wine*]]]]

b. [$_{DP}$ *a* [$_{NP}$ *delicious* [$_{N'}$ [$_{N'}$ [$_{N^0}$ *bottle*]] [$_{PP}$ *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 (19)), when it is smaller than DP.

(19) *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*. To match the observed toddler behavior in the LWF experiment, a learner 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).

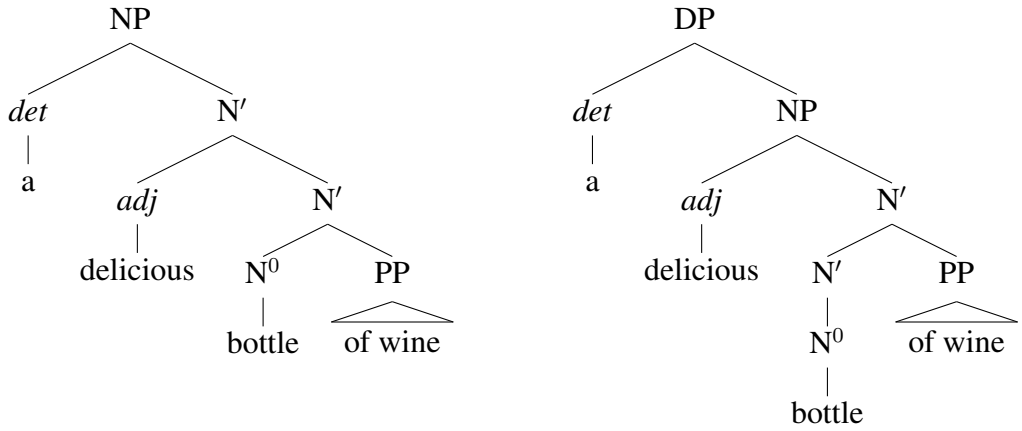

Figure 4: Phrase structure trees corresponding to the bracket notation in examples (18a) and (18b) for *a delicious bottle of wine*.

Therefore, the target knowledge state of the learner should be adjusted with respect to syntactic category (NP instead of N′), though the target referential knowledge (include the modifier in the antecedent) and target behavior (look to the familiar object) remain the same. Similarly, the initial state of the learner is adjusted so that categorical knowledge includes DP, NP, N′, and N^0.

While we have not implemented a learning strategy that uses this syntactic representation, we can easily describe the expected results for the indirect positive evidence strategy (IndirPro) that learns from data containing other pronouns, as there are still many similarities to the learning scenario already implemented. We describe the impact of the four data types in turn.

DirUnamb data (e.g., *Look – a red bottle! Oh, but I don't see another one here* when a red bottle and purple bottle are present) still indicate that the antecedent should include the modifier, so the probability is increased of the correct referential interpretation (p_{incl}). Because only category NP can include a modifier, the probability of the correct syntactic category (NP) also increases. This is qualitatively similar to our current implementation.

DirRefSynAmb data (e.g., *Look – a red bottle! Oh, look – another one* when two red bottles are present) are still ambiguous between three antecedents – here, [$_{NP}$ *red bottle*], [$_{N'}$ *bottle*], and [$_{N^0}$ *bottle*]. When the suspicious coincidence of the referent just happening to have the mentioned property is high enough ($s > 5$), these data will cause the learner to believe the antecedent includes the modifier. So, the probability of the correct referential interpretation is increased (p_{incl}) and the probability of syntactic category NP is also increased since this is the only category that allows a modifier. This is again qualitatively similar to our current implementation.

DirSynAmb data (e.g., *Look – a bottle! Oh, look – another one!*) retain their two-way ambiguity (N′ vs. N^0). When given data compatible with two hypotheses, our probabilistic 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 DirSynAmb data will cause the learner to prefer N^0, as our learner did here (though perhaps not as quickly, depending on the frequency of noun+complement N′ strings in the input).

Still, DirSynAmb data remain misleading about the syntactic category of *one* (i.e., category = N^0), similar to our current implementation.

IndirUnamb data (e.g., *Look – a red bottle! I want it*) are still informative about p_{incl}, as they indicate the modifier is included in the antecedent. It is simply that the syntactic category is DP instead of NP, as in our current implementation. Thus, again, the effect of these data on learning is the same as in our current implementation.

Because no data favor N′, we would expect that the learner disprefers *one* as N′ at the end of learning. Instead, the learner would assume *one* is NP (e.g., antecedent = *red bottle*) in contexts like the LWF experiment that have a property mentioned and *one* is N^0 in general when no property is mentioned. This is the same result that we have found here with our current implementation. Thus, altering the theoretical representation this way does not qualitatively alter the results we have found with respect to the indirect positive evidence strategy.