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

Lisa S. Pearl, Benjamin Mis

LANGUAGE
A JOURNAL OF THE LINGUISTIC
SOCIETY OF AMERICA

Language, Volume 92, Number 1, March 2016, (Article)

Published by Linguistic Society of America
DOI: https://doi.org/10.1353/lan.2016.0022

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

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

A Formal description of data types

Table 1 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 1: 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 (1). 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, Pullum, Scholz, & Berlage, 2013)). We classify them as indirect positive evidence here, since they will function the same way as indirect positive evidence coming from other pronouns.

(1) 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), (7), and (8) in the main text). 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 (1) it is always NP, while in direct positive unambiguous (DirUnamb) utterances like (6) in the main text, 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) in the main text.

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**.

$$
\begin{aligned}
\phi_{incl} &= p(i = yes | o\text{-}m = yes) && \text{(6a)}\\
&= \frac{p(i = yes, o\text{-}m = yes)}{p(o\text{-}m = yes)} && \text{(6b)}\\
&= \frac{\sum_{O,A,det,mod,C,Pro,env,R,m} p(i = yes, o\text{-}m = yes)}{\sum_{O,A,det,mod,C,Pro,env,R,i,m} p(o\text{-}m = yes)} && \text{(6c)}
\end{aligned}
$$

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).

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

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

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 2 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 2: 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 1, and shown in bracket notation in (2a). However, an alternate representation of noun phrases is available (Bernstein, 2003; Longobardi, 2003), shown in (2b) and the right side of Figure 1. 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′.

(2) Theoretical representations for noun phrase syntax

a. $[_{\text{NP}}$ *a* $[_{\text{N'}}$ *delicious* $[_{\text{N'}}$ $[_{\text{N}^0}$ *bottle*] $[_{\text{PP}}$ *of wine*]]]]
b. $[_{\text{DP}}$ *a* $[_{\text{NP}}$ *delicious* $[_{\text{N'}}$ $[_{\text{N'}}$ $[_{\text{N}^0}$ *bottle*]] $[_{\text{PP}}$ *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 (3)), when it is smaller than DP.

(3) *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).

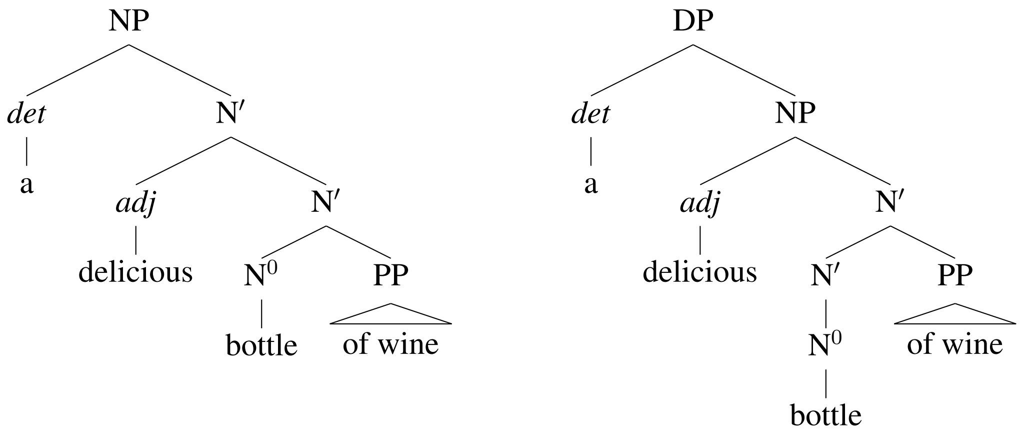

Figure 1: Phrase structure trees corresponding to the bracket notation in examples (2a) and (2b) for *a delicious bottle of wine*.

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.

References

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.

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

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

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.

Payne, J., Pullum, G., Scholz, B., & Berlage, E. (2013). Anaphoric *one* and its implications. *Language*, *90*(4), 794–829.

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.

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

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