Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics

K.J. Savinelli
UCI Cognitive Sciences
ksavinel@uci.edu

Gregory Scontras
UCI Linguistics
g.scontras@uci.edu

Lisa Pearl
UCI Linguistics
& Cognitive Sciences
lpearl@uci.edu

Abstract

Behavioral data suggest that both children and adults struggle to access the `inverse` interpretation of scopally-ambiguous utterances in certain contexts. To determine whether the causes of both child and adult difficulty are similar, we extend an existing computational model of children's scope ambiguity resolution in context. We find that the same utterance-disambiguation mechanism is active in both children and adults, supporting the theory of developmental continuity. Moreover, because adult behavior requires an `exact` semantics for numerals, we also provide empirical support for this theory of linguistic representation.

Keywords: ambiguity resolution, developmental continuity, language acquisition, numerals, pragmatics, processing, Rational Speech Act model, scope, semantics

1 Introduction

Consider a scenario where two out of three horses jump over a fence. Is the utterance in (1) a reasonable description?

(1) *Every horse didn't jump over the fence.*
   a. $\forall \gg \neg$ (`surface` scope):
   None of the horses jumped over the fence.
   b. $\neg \gg \forall$ (`inverse` scope):
   Not all of the horses jumped the fence.

Adults typically endorse the *every-not* utterance as true, while children typically do not (Musolino, 1998; Lidz and Musolino, 2002; Musolino and Lidz, 2006; Musolino, 2006; Viau et al., 2010). This utterance is scopally ambiguous, involving multiple quantifiers (i.e., *every* and *n't*). Children's behavior is non-adult-like at five years old: though the `inverse` interpretation in (1b) is true, five-year-olds still do not endorse the utterance.

Now, consider a scenario with only two horses, one of which successfully jumps. Is the *two-not* utterance in (2) a reasonable description?

(2) *Two horses didn't jump.*
   a. $\exists 2 \gg \neg$ (`surface` scope):
   There are two horses that didn't jump.
   b. $\neg \gg \exists 2$ (`inverse` scope):
   It's not the case that there are two horses that jumped.

Most adults would *not* endorse the utterance, despite the `inverse` interpretation in (2b) being true (Musolino and Lidz, 2003)—that is, it is not the case that two horses jumped (only one did).

This pair of findings underscores that not endorsing a scopally-ambiguous utterance when only the `inverse` interpretation is true occurs in both children and adults in different contexts. We might therefore wonder about continuity in the development of scope ambiguity resolution: is the cause of child utterance non-endorsement in an *every-not* scenario qualitatively similar to the cause of adult non-endorsement in the *two-not* scenario? If so, this similarity supports developmental continuity: children use the same mechanism as adults when understanding ambiguous utterances in context. The only difference would be that adults are better-equipped to deploy this mechanism, owing perhaps to increased domain-general knowledge and/or cognitive capacities, or to language-specific experience. In contrast, if the underlying causes are different for child and adult utterance non-endorsement, this would suggest developmental discontinuity: children are engaging in a fundamentally different process as they understand ambiguous utterances. So, the development of adult-like behavior would involve ac-

Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 67–75,
Salt Lake City, Utah, USA, January 7, 2018.

quiring a new mechanism for resolving ambiguity.

To choose between these accounts, we must understand utterance (non-)endorsement behavior. To that end, Savinelli et al. (2017) articulated a computational model of ambiguity resolution within the Rational Speech Act (**RSA**) framework (Goodman and Frank, 2016). The model demonstrated the central role of pragmatic factors over processing factors in explaining children's non-adult-like behavior in *every-not* contexts like (1). Here, we extend this same model to capture *two-not* utterance endorsement behavior in adults, identifying the factors that yield the experimentally-observed patterns of behavior.

We begin by reviewing the scope ambiguity resolution findings from Savinelli et al. (2017), together with the experimental results that informed the design of the computational model. Next, we consider the experimental findings from Musolino and Lidz (2003), where adults seem to behave like children in specific contexts. We then extend the model from Savinelli et al. (2017) to capture these new data, and demonstrate support for developmental continuity, with the same utterance-disambiguation mechanism active in both children and adults. Importantly, the complete range of experimentally-observed behavior can only be captured if adults represent *two* with an `exact` interpretation, an unexpected finding that informs the debate on numeral semantics.

2 Previous work: Modeling *every-not*

In the basic truth-value judgment task (**TVJT**) meant to assess children's scope disambiguation behavior, children first watch a scene acted out and hear a puppet produce a scopally-ambiguous utterance; then they are asked whether they would endorse the utterance as a true description of the scenario. Children typically do not endorse the ambiguous *every-not* utterance in the critical context where the `surface` interpretation is false but the `inverse` interpretation is true (e.g., a NOT-ALL scenario where two out of three horses jumped over a fence). This behavior has been interpreted as children failing to access the `inverse` scope interpretation that would make the utterance true.

Interestingly, various alterations to the task setup have yielded more adult-like behavior in children, with higher rates of endorsement for the *every-not* utterance. These experimental manipulations highlight at least three core factors (two pragmatic, one processing) that underlie children's behavior in the TVJT: (i) *pragmatic*: expectations about the experimental world (e.g., how likely successful outcomes are), (ii) *pragmatic*: expectations about the Question Under Discussion (**QUD**; e.g., were `all` outcomes successful?), and (iii) *processing*: the accessibility of the `inverse` scope (i.e., the ease by which the logical form is either derived or accessed in real time).

To capture and independently manipulate the contributions of each of these factors, Savinelli et al. (2017) modeled ambiguity resolution for *every-not* utterances within the Bayesian RSA framework (Goodman and Frank, 2016). They found that when it comes to understanding non-adult-like behavior in the TVJT, there is likely a stronger role for the pragmatics of context management (as realized in prior beliefs about world state and QUD) than for grammatical processing (as realized in the prior on scope interpretations), although there may be a role for both. So, children's failure to endorse scopally-ambiguous *every-not* utterances in NOT-ALL contexts likely stems from their beliefs about the experimental world (e.g., whether actors are *a priori* likely to succeed) and about the topic of conversation (e.g., whether the conversational goal is to determine if `all` the actors succeeded), rather than an inability to grammatically derive or access the inverse scope interpretation in real time.

Perhaps most interesting was the prediction that the highest rates of utterance endorsement (i.e., adult-like behavior) occur when resolving the scope ambiguity is *irrelevant* for communicating successfully about the NOT-ALL world. This occurs when expectations about the world state favor total success, or when the QUD asks if `all?` of the actors succeeded. In either case, *both* scope interpretations serve to inform a listener, either that the *a priori* likely total-success world state does not hold or that the answer to the `all?` QUD is *no*.

The explanation for utterance non-endorsement (i.e., non-adult-like behavior) is similar: Savinelli et al. (2017)'s model predicts the lowest rates of utterance endorsement in NOT-ALL scenarios when neither interpretation is useful for successful communication, either because the interpretation is false (`surface`) or because beliefs about the pragmatic context render the interpretation uninformative (`inverse`). Thus, the TVJT utterance non-endorsement data previously used to demon-

strate children's difficulty with `inverse` scope calculation in fact require no disambiguation at all if the goal is informative communication. Instead, children simply need the ability to manage the pragmatic context so they can recognize the potential informativity of these ambiguous utterances. Notably, considerations of pragmatic context have long played a role in the design and interpretation of the TVJT (e.g., Crain et al., 1996). Savinelli et al. (2017) take the extra step of formally articulating specific pragmatic factors and the role they play in children's apparent difficulty with ambiguous utterances in the TVJT.

3 Experimental *two-not* results

Musolino and Lidz (2003) (**ML2003**) demonstrated that adults are sensitive to some of the same experimentally-manipulated factors as children when it comes to endorsing scopally-ambiguous utterances. Like us, ML2003 were interested in developmental continuity: are child and adult ambiguity resolution behavior in context qualitatively similar? To investigate this, they conducted three TVJTs.

The goal of the first TVJT was to determine which interpretation adults preferred when they endorsed a scopally-ambiguous utterance in context. For example, adults heard *"Cookie Monster didn't eat two pizza slices"* in a context where both interpretations were true, such as Cookie Monster eating one of three available pizza slices (`surface`: *it's not the case he ate two* = true; `inverse`: *there are two he didn't eat* = true). Importantly, they were then asked to explain *why* they endorsed the utterance so that their preferred scope interpretation could be inferred. For example, if their answer referred to Cookie Monster eating only one slice, then it was assumed that they accessed the `surface` interpretation (`surface`: *he only ate one, so it's not the case he ate two*). However, if their answer referred to the two slices Cookie Monster did not eat, then it was assumed that they accessed the `inverse` interpretation (`inverse`: *there are two he didn't eat*). All participants endorsed the utterance, and their explanations indicated a strong surface scope bias (75% `surface`, 7.5% `inverse`, 17.5% unclear from explanation). ML2003 interpreted this finding as evdence that adults prefer the `surface` scope interpretation when both interpretations are true in context. It could then be that children's non-endorsement behavior, if due to a preference for the `surface` scope interpretation, is driven by a stronger version of this same preference.

In the second TVJT, adults heard an utterance like (2) (e.g., *Two frogs didn't jump over the rock*) in two different contexts. The first context included two actors (e.g., frogs), with one actor successfully completing the action (e.g., $frog_1$ jumping over the rock while $frog_2$ does not). In this 1-OF-2 context, the `surface` interpretation is false (only $frog_2$ did not jump, so it is false that two frogs didn't jump), but the `inverse` interpretation is true (only $frog_1$ did jump, so it is indeed not the case that two frogs jumped). Yet, adults had low endorsement (endorsement rate: 27.5%).

In the second context, there were four actors. For example, four frogs attempted to jump over a rock; two jumped ($frog_1$, $frog_2$) and two did not ($frog_3$, $frog_4$). In this 2-OF-4 context, the `surface` interpretation of the scopally-ambiguous utterance is true because $frog_3$ and $frog_4$ did not jump. However, the `inverse` interpretation is false because $frog_1$ and $frog_2$ did indeed jump. Here, adults had an endorsement rate of 100%.

ML2003 interpreted this asymmetry of endorsement between the two contexts as a strong `surface` scope preference in adults. According to this explanation, non-endorsement occurs in the 1-OF-2 context because only the `inverse` scope is true; in contrast, endorsement occurs in the 2-OF-4 context because only the `surface` scope is true. That is, both these patterns would result because adults favor the `surface` interpretation. While we find this account compelling, we note that there are other differences between the two contexts that might lead to the observed asymmetry. For example, it could be that the seemingly benign change from two to four total actors affects the pragmatic context. Another variable is the potential ambiguity present in the numeral semantics, which only occurs in the 2-OF-4 context.[1] In either case, exploring the effects of these factors in a formal model of TVJT behavior can clarify the process underlying utterance disambiguation.

Returning to the question of continuity, while the observable behavior appears qualitatively the same in children and adults (i.e., a non-endorsement preference when only the `inverse` scope is true), it remains unclear whether the underlying cause of this behavior is the same. To

[1] A topic discussed in more detail in the following section.

evaluate this, ML2003 conducted a third TVJT with adults in 1-OF-2 contexts, involving an experimental manipulation from Lidz and Musolino (2002) that children are known to be sensitive to. This manipulation is implemented as an explicit linguistic contrast clause before the scopally-ambiguous utterance, such as the bolded material in (3).

(3) **Two frogs jumped over the fence but** two frogs didn't jump over the rock.

Adults responded the same way as the children from Lidz and Musolino (2002), shifting to strong endorsement in the 1-OF-2 context (endorsement rate: 92.5%; cf. 27.5% endorsement without the explicit contrast). Yet, as ML2003 note themselves, it is not obvious *why* the adult endorsement rate increases when the linguistic contrast is present. According to ML2003, the linguistic contrast creates the positive expectation necessary to make the negation in the later clause felicitous (Wason, 1965; Musolino and Lidz, 2003). However, it remains unclear *how* exactly the context creates the positive expectation. There are multiple ways this information could impact the context. For example, the positive expectation could arise because of a change *either* in the pragmatic factor of world knowledge or in the pragmatic factor of the QUD. Specifically, the affirmative statement could alter the listener's beliefs about how successful frogs are known to be in the experimental world. This affirmative statement also potentially changes the listener's expectations about the QUD: because both frogs were successful before, the topic of conversation might now be focused on whether both frogs were successful again. Both these effects could generate a context that makes the negated clause more informative.

Without knowing the factors responsible for endorsement behavior, it is difficult to determine whether the same factors are operating in both children and adults, and whether the underlying representation of *two* matters. Computational modeling can help determine why these two behavioral patterns occur: (i) adult sensitivity to the pragmatic contrast manipulation, and (ii) asymmetry in endorsement behavior between 1-OF-2 and 2-OF-4 contexts in the absence of that pragmatic contrast. In the next section, we extend Savinelli et al. (2017)'s model of utterance disambiguation to handle these empirical data.

4 Modeling *two-not*

Savinelli et al. (2017)'s model of ambiguity resolution is conceived within the Bayesian Rational Speech Act (RSA) framework (Goodman and Frank, 2016), which views language understanding as a social reasoning process. A *pragmatic listener* L_1 interprets an utterance by reasoning about a cooperative *speaker* S_1 who is trying to inform a *literal listener* L_0 about the world. The model is a "lifted-variable" extension in which the ambiguous utterance's literal semantics gets parameterized by interpretation-fixing variables (e.g., the relative scope of the quantificational elements; Bergen et al., 2012; Lassiter and Goodman, 2013; Scontras and Goodman, 2017). Hearing an ambiguous utterance, the pragmatic listener L_1 reasons jointly about the true state of the world (e.g., how many frogs successfully jumped), the scope interpretation speaker S_1 had in mind (i.e., `surface`, `inverse`), as well as the likely QUD that the utterance addresses (e.g., did `all` frogs succeed?). To generate testable predictions, participant TVJT behavior is modeled as a *pragmatic speaker* S_2's (relative) endorsement of an utterance about an observed situation (cf. Degen and Goodman, 2014; Tessler and Goodman, 2016). That is, this model predicts whether a speaker S_2 would endorse the scopally-ambiguous utterance as a description of the observed state. S_2 decides this by reasoning about whether a pragmatic listener L_1 (who is reasoning about a speaker S_1 reasoning about a literal listener L_0) would arrive at the correct world state after hearing the utterance.

We take world states $w \in W$ to consist of a collection of n individuals (e.g., frogs), each of which either succeeds or fails at the relevant task (e.g., jumping over a rock). The world success baserate b_{suc} determines the probability that an individual will succeed. We assume a simple truth-functional semantics where an utterance u denotes a mapping from world states to truth values ($Bool = \{\texttt{true}, \texttt{false}\}$). We parameterize this truth function so that it depends on the scope interpretation $i \in I = \{\texttt{inverse}, \texttt{surface}\}$, $[\![u]\!]^i$: $W \rightarrow Bool$. We consider two alternative utterances $u \in U$: the `null` utterance (i.e., saying nothing at all, and so choosing *not* to endorse the utterance) and the scopally-ambiguous utterance `amb` (e.g., *"Two frogs didn't jump over the rock"*).

To fix the utterance semantics, we must consider potential ambiguity introduced by the nu-

meral in cases where the number of relevant individuals n exceeds the numeral's value. For example, consider the positive utterance *"Two frogs jumped over the rock."* If we assign an `exact` (=) semantics to *two*, the sentence will be true only when two frogs succeeded. If we assign an `at-least` ($\geq$) semantics, the sentence will be true when two or more frogs succeeded. In worlds with only two frogs, the $=$ vs. $\geq$ distinction makes no difference: the sentence will be true in the world where both frogs succeed, and false in all other worlds. However, in a world with four frogs, the numeral semantics will define different truth-functional mappings. With the $=$ semantics, the sentence is true in any world where two frogs—but not more—succeed. With the $\geq$ semantics, the sentence is true in a larger set of worlds, where two or more frogs succeed.

To evaluate the potential contribution of utterance semantics to the 1-OF-2 vs. 2-OF-4 asymmetry, we consider two different sets of utterance alternatives, one with $\texttt{amb}_{=}$ and another with $\texttt{amb}_{\geq}$. So, $U_{=} = \{\texttt{null}, \texttt{amb}_{=}\}$ and $U_{\geq} = \{\texttt{null}, \texttt{amb}_{\geq}\}$. The utterance semantics in (4) shows that scope parameterization i only impacts the truth conditions for `amb` utterances.[2]

(4) *Utterance semantics* $[\![u]\!]^i$:

a. $[\![\texttt{null}]\!]^i = \texttt{true}$

b. $[\![\texttt{amb}_{=/\geq}]\!]^i =$ if $i =$ `inverse` then $[\![\texttt{inverse}_{=/\geq}]\!]$ else $[\![\texttt{surface}_{=/\geq}]\!]$

where:

$[\![\texttt{inverse}_{=}]\!] =$
λw. $\neg\exists$!x: $|\text{x}| = 2 \wedge \text{x} \subseteq \texttt{success}(\text{w})$

$[\![\texttt{surface}_{=}]\!] =$
λw. $\exists$!x: $|\text{x}| = 2 \wedge \text{x} \not\subseteq \texttt{success}(\text{w})$

$[\![\texttt{inverse}_{\geq}]\!] =$
λw. $\neg\exists$x: $|\text{x}| = 2 \wedge \text{x} \subseteq \texttt{success}(\text{w})$

$[\![\texttt{surface}_{\geq}]\!] =$
λw. $\exists$x: $|\text{x}| = 2 \wedge \text{x} \not\subseteq \texttt{success}(\text{w})$

We consider five potential QUDs $q \in Q$, three from the original Savinelli et al. (2017) model: (i) "What happened with the frogs?" (`what-happened?`), (ii) "Did all the frogs succeed?" (`all?`), and (iii) "Did none of the frogs succeed?" (`none?`). We also consider two additional QUDs specific to the *two-not* utterance: (iv) "Did exactly two frogs succeed?" ($\texttt{two}_{=}\texttt{?}$), and (v) "Did at least two frogs succeed?" ($\texttt{two}_{\geq}\texttt{?}$). The QUDs serve as projections from the inferred world state to the relevant dimension of meaning, so that $q : W \rightarrow X$ (Kao et al., 2014a,b). In practice, the QUDs establish partitions on the possible world states, as shown in (5). For example, the `all?` QUD partitions the world space in two: the unique world in which all frogs succeeded (`true`) and all other possible worlds (`false`).

(5) *QUD semantics* $[\![q]\!]$:

a. $[\![\texttt{what-happened?}]\!] = \lambda$w. w

b. $[\![\texttt{all?}]\!] = \lambda$w. $\texttt{success}(\text{w}) = \text{w}$

c. $[\![\texttt{none?}]\!] = \lambda$w. $\texttt{success}(\text{w}) = \emptyset$

d. $[\![\texttt{two}_{=}\texttt{?}]\!] = \lambda$w. $|\texttt{success}(\text{w})| = 2$

e. $[\![\texttt{two}_{\geq}\texttt{?}]\!] = \lambda$w. $|\texttt{success}(\text{w})| \geq 2$

Literal listener L_0 has prior uncertainty about the true state, $P(w)$. L_0 updates beliefs about w conditioned on the the literal semantics, and restricts prior beliefs to those worlds that $[\![u]\!]^i$ maps to `true`. The function $\delta_{[\![u]\!]^i(w)}$ maps the Boolean truth value to a probability, 1 or 0.

$$P_{L_0}(w|u,i) \propto \delta_{[\![u]\!]^i(w)} \cdot P(w)$$

To capture the notion that communication proceeds relative to a specific QUD q, L_0 must infer not only the true world state w, but also the value of the QUD applied to that world state, $[\![q]\!](w) = x$.

$$P_{L_0}(x|u,i,q) \propto \sum_{w} \delta_{x=[\![q]\!](w)} \cdot P_{L_0}(w|u,i)$$

Speaker S_1 chooses an utterance u in proportion to its utility in communicating about the true world state w with respect to the QUD q, $[\![q]\!](w) = x$. Thus, the speaker maximizes the probability that L_0 arrives at the intended x from u. This selection is implemented via a softmax function (exp) and free parameter α, which controls how rational the speaker is in utterance selection.

$$P_{S_1}(u|w,i,q) \propto exp(\alpha \cdot log(L_0(x|u,i,q)))$$

Utterance interpretation happens at the level of the pragmatic listener L_1, who interprets an utterance u to jointly infer the world state w, the interpretation i, and the QUD q. We model ambiguity resolution as pragmatic inference over an underspecified utterance semantics (i.e., the interpretation variable i). To do this, L_1 inverts S_1's model, and so the joint probability of w, i, and q is proportional to the likelihood of S_1 producing utterance u given world state w, interpretation i, and QUD q, as well as the priors on w, i, and q.

$$P_{L_1}(w,i,q|u) \propto P_{S_1}(u|w,i,q) \cdot P(w) \cdot P(i) \cdot P(q)$$

[2] The `success()` function in (4) returns the set of successful outcomes in a world w.

To model the utterance endorsement implicit in TVJT, we need an additional level of inference. Pragmatic speaker S_2 observes the true world state w and selects u by inverting the L_1 model, thus maximizing the probability that a pragmatic listener would arrive at w from u by summing over possible interpretations i and QUDs q for world w.

$$P_{S_2}(u|w) \propto exp(log \sum_{i,q} P_{L_1}(w,i,q|u))$$

To generate model predictions for adult sensitivity to the pragmatic contrast manipulation and the 1-OF-2 vs. 2-OF-4 asymmetry, we fix various model parameters. For 1-OF-2 data, we set the number of individuals n to 2; for 2-OF-4 data, we set n to 4. The S_1 speaker rationality parameter $\alpha > 0$ is set to 2.5 (i.e., the same value in the *every-not* simulations in Savinelli et al., 2017). The priors $P(w)$ and $P(q)$ correspond to expectations for the discourse context (i.e., likely world states or QUDs). In the default case, we set these priors to be uniform over their possible values, with the individual success baserate b_{suc} set to 0.5 and the relevant QUDs having equal probability. The interpretation prior $P(i)$ corresponds to how easy it is to access the `inverse` scope interpretation. In the default case, P(`inverse`) = P(`surface`) = 0.5. Importantly, to better understand utterance endorsement behavior with scopally-ambiguous utterances, we can independently manipulate the values of the priors on W, Q, and I, and observe their impact on utterance endorsement.

5 Results

Recall the empirical phenomena we are trying to capture: (i) the dramatic increase in endorsement rates in the 1-OF-2 context when an explicit contrast is present, and (ii) the stark asymmetry in utterance endorsement rates between 1-OF-2 and 2-OF-4 contexts in the absence of that explicit contrast. We report results for each in turn.

5.1 The explicit contrast effect for 1-OF-2

Following Savinelli et al. (2017), we attempt to capture the increase in ambiguous utterance endorsement rates by systematically manipulating the pragmatic and processing factors, as implemented in the relevant priors.

For the world state prior (Figure 1, *left*), we manipulate baserate b_{suc}, which determines an actor's chance of success. Holding the QUD and scope priors at their default values, we see a marked increase in endorsement of the ambiguous utterance in the 1-OF-2 context as prior beliefs about frog success increase. Utterance endorsement is at its lowest (33%) when prior knowledge suggests that frogs are particularly unlikely to succeed; endorsement is at its highest (86%) when frogs are very likely to succeed.

For the QUD prior (Figure 1, *center*), we selectively favor specific QUDs by assigning a 0.9 probability to the favored QUD and dividing the remaining probability equally among the others. Since the `two?` QUDs are equivalent to the `all?` QUD in the 1-OF-2 context, we omitted the `two?` QUDs in the 1-OF-2 context. Holding the other priors at their default values, endorsement rates increase from favoring the `none?` QUD (35%) to favoring the `what-happened?` QUD (46%) to favoring the `all?` QUD (64%).

For the scope prior (Figure 1, *right*), we manipulate the prior probability of the `inverse` interpretation while holding the other factors at their default values. We see an increase in utterance endorsement as the probability of `inverse` increases, from a low of 40% to a high of 57%.

Each manipulation qualitatively captures the response pattern from ML2003, and replicates the results of Savinelli et al. for *every-not*. However, as observed by Savinelli et al., the pragmatic factors controlling world and QUD beliefs have a much more pronounced effect than the processing factor controlling scope access; the model's world prior baserate manipulation comes closest to capturing the experimentally-observed effect of explicit contrast manipulation (i.e., 27.5% base endorsement vs. 92.5% endorsement with the explicit contrast). We can amplify the effect of the world baserate manipulation by allowing it to interact with the other factors.

As discussed in Section 3, the early success explicit contrast manipulation possibly affects two aspects of the disambiguation calculus: it could increase expectations for success and shift the topic of conversation to whether total success was achieved again. Figure 2 plots the interaction of the world and QUD priors, together with the effect of scope. The low-endorsement baseline (27.5%) most likely results from low expectations for success ($b_{suc} = 0.1$) and QUD uncertainty (QUD: `uniform`), together with a moderate to low probability of accessing the `inverse` scope ($P(inv) =$ 0.1 or 0.5). From this baseline, we implement

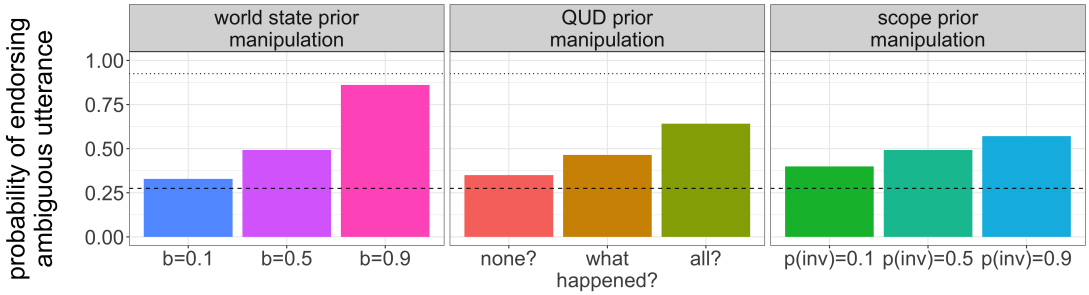

Figure 1: Model predictions for ambiguous *two-not* utterance endorsement (e.g., *Two frogs didn't jump over the rock*) in a 1-OF-2 context. Dotted lines represent experimentally-observed endorsement behavior in the absence (lower) and presence (upper) of an explicit contrast.

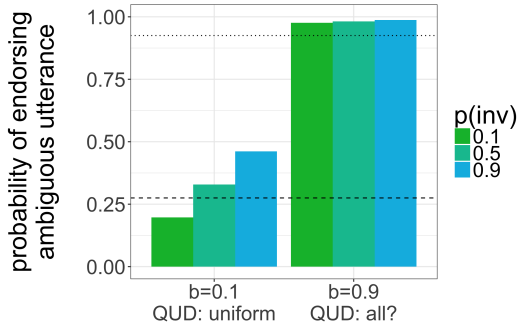

Figure 2: Model predictions for ambiguous *two-not* utterance endorsement in a 1-OF-2 context when multiple factors interact. Dotted lines represent experimentally-observed endorsement behavior in the absence (lower) and presence (upper) of an explicit contrast.

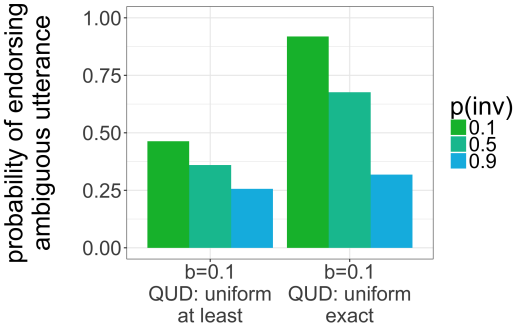

Figure 3: Model predictions for ambiguous *two-not* endorsement in a 2-OF-4 context.

the effect of the explicit contrast manipulation by increasing success expectations ($b_{suc} = 0.9$) and shifting the topic of conversation to whether total success occurred (QUD: `all?`). This manipulation results in a dramatic increase in utterance endorsement, irrespective of scope.

To summarize, if the explicit contrast clause impacts a listener's beliefs about the frogs' chance of success (increasing b_{suc}) or the QUD (favoring `all?`), then the model predicts the endorsement rate should increase. Notably, both of these manipulations make the *two-not* scopally-ambiguous utterance more informative for a listener. In the case of the the world state manipulation, *two-not*—under either scope interpretation—informs the listener that her prior beliefs about total frog success do not hold. Similarly with the QUD manipulation favoring `all?`, both scope interpretations answer this question in the negative (i.e., it is not the case that all (two) frogs succeeded).

5.2 The 1-OF-2 vs. 2-OF-4 asymmetry

If the factors identified for capturing the experimentally-observed effect of the explicit contrast are indeed active in utterance disambiguation (i.e., to validate their explanatory power), we would expect the very same factors and values to additionally capture the ceiling-level endorsement rate in the 2-OF-4 context without the explicit contrast.

Recall the baseline 1-OF-2 values from Figure 2: low expectations for success ($b_{suc} = 0.1$) and QUD uncertainty (QUD: `uniform`). To model the 2-OF-4 context, we change the number of actors n to 4 and additionally manipulate whether the `exact` ($=$) or `at-least` ($\geq$) semantics applies, as they diverge when there are more than two actors in the context (see section 4). This decision impacts both the utterance semantics and the relevant set of QUDs (e.g., if $\geq$ semantics gets used, then the $\texttt{two}_{\geq}$? QUD is included in the set of potential QUDs). As shown in Figure 3, we do indeed predict high endorsement with the same parameter value baseline, but only with `exact` utterance semantics and a low probability of accessing the inverse scope ($P(inv) = 0.1$). In this case, we find an endorsement rate of 92%.

6 Discussion

Our model of ambiguity resolution in context captures the effect of the explicit contrast manipulation observed in adults in ML2003, and notably

also captured the same effect in children (Savinelli et al., 2017). This parallelism—sensitivity to the pragmatic context in both children and adults across different contexts—suggests that the same disambiguation mechanism is active in both children and adults. Adults seem better able to charitably interpret less supportive pragmatic contexts (i.e., the original *every-not* scenarios); yet, there remain scenarios (i.e., certain *two-not* contexts) where even adult abilities are exceeded. We interpret the common underlying mechanism as support for developmental continuity in scope ambiguity resolution, with no qualitative shift required.

In addition to supporting the developmental continuity hypothesis, this model also suggests *why* manipulations like the explicit contrast clause work. The pragmatic variables capture the explicit contrast manipulation because they create a situation where the ambiguous *two-not* utterance is still informative *despite* the ambiguity. When the utterance provides the listener with information that diverges from her prior beliefs, the ambiguous *two-not* utterance becomes more informative, more useful, and therefore more endorsable.

The model also seamlessly captures ML2003's results from the 2-OF-4 context: with the very same parameter values that yield low endorsement rates for 1-OF-2 contexts, the model predicts the high endorsement observed for 2-OF-4 contexts. The only change is increasing the number of relevant individuals from two to four. This exploration of the 1-OF-2 vs. 2-OF-4 contexts allows us to refine our understanding of the potential sources of child and adult behavior. Savinelli et al. (2017)'s findings suggested that pragmatic factors alone are capable of capturing the non-adult-like behavior in children and the extension in the current model captures the explicit contrast effect in adults; however, the processing factor of scope (in particular, disfavoring the inverse scope) is needed to account for ML2003's 2-OF-4 results. This finding supports ML2003's conclusion, namely that adults have a strong preference for surface interpretations of *two-not* utterances. Combined with the appropriate pragmatic context, that preference has the potential to drive the endorsement asymmetry between the 1-OF-2 and 2-OF-4 contexts. Whether this surface interpretation preference in *two-not* contexts is also something children share remains an open empirical question; experimental results for *every-not* do not answer this question definitively (Viau et al., 2010; Savinelli et al., 2017).

Importantly, the present model requires one more ingredient to account for the 1-OF-2 vs. 2-OF-4 difference in adult behavior: an `exact` numeral semantics (in contrast to an `at-least` semantics; cf. Geurts, 2006; Breheny, 2008; Spector, 2013; Kennedy, 2015). While the underlying utterance semantics is not something easy to manipulate in an experiment, it is exactly the kind of variable we can systematically explore in a computational model. By doing so here, we are able to show the necessity of an `exact` semantics in generating observable adult behavior. This provides empirical support, coming from computational modeling, for theories about the semantics of numerals. In particular, the only way to account for the observed adult behavior is if adults interpret *two* utterances as meaning `exactly` two.

To sum up, these findings underscore the complexity of information involved in interpreting scopally-ambiguous utterances, including the literal semantics of the utterances involved, processing factors that affect interpretation accessibility, pragmatic factors that affect the potential informativity of the utterance, and the recursive social reasoning between speakers and listeners. Here, we find evidence for the impact of both pragmatic and processing factors, and in particular how a specific confluence of values for these factors yields the observed adult utterance endorsement behavior in multiple contexts. The fact that pragmatic factors can have such a pronounced effect on their own accords with previous computational findings about the cause of children's utterance endorsement behavior in context, thereby highlighting the developmental continuity in pragmatic reasoning from childhood to adulthood. Moreover, the fact that the processing factor of scope access is crucial for explaining adult behavior in certain contexts motivates experimental work with children to see if their behavior is likewise affected by this processing factor in similar contexts. The fact that only the `exact` utterance semantics is capable of yielding the observed behavior provides empirical support in favor of this theory of representation for numerals. More broadly, we have demonstrated how computational modeling can help us refine our theories about different aspects of language, including theories of language understanding, language development, and language representation.

References

Leon Bergen, Noah Goodman, and Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Cognitive Science Society*. volume 34, pages 120–125. http://escholarship.org/uc/item/5f03m09d.

Richard Breheny. 2008. A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25(2):93–139. https://doi.org/10.1093/jos/ffm016.

Stephen Crain, Rosalind Thornton, Carole Boster, Laura Conway, Diane Lillo-Martin, and Elaine Woodams. 1996. Quantification without qualification. *Language Acquisition* 5(2):83–153. https://doi.org/10.1207/s15327817la0502_2.

Judith Degen and Noah D Goodman. 2014. Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 36, pages 397–402. http://escholarship.org/uc/item/97t2w1f3.

Bart Geurts. 2006. Take five: The meaning and use of a number word. In Svetlana Vogeleer and Liliane Tasmowski, editors, *Non-Definiteness and Plurality*, Benjamins, Amsterdam, pages 311–329. https://doi.org/10.1075/la.95.16geu.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11):818–829. https://doi.org/10.1016/j.tics.2016.08.005.

Justine T Kao, Leon Bergen, and Noah D Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of Annual Meeting of the Cognitive Science Society*. volume 36, pages 719–724. http://escholarship.org/uc/item/09h3p4cz.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33):12002–12007. https://doi.org/10.1073/pnas.1407479111.

Chris Kennedy. 2015. A "de-Fregean" semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semantics and Pragmatics* 8(1):1–44. https://doi.org/10.3765/sp.8.10.

Daniel Lassiter and Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23*. pages 587–610. https://doi.org/10.3765/salt.v23i0.2658.

Jeffrey Lidz and Julien Musolino. 2002. Children's command of quantification. *Cognition* 84(2):113–154. https://doi.org/10.1016/S0010-0277(02)00013-6.

Julien Musolino. 1998. *Universal Grammar and the Acquisition of Semantic Knowledge: An Experimental Investigation into the Acquisition of Quantifier Negation Interaction in English*. Doctoral dissertation, University of Maryland, College Park. http://ling.umd.edu/assets/publications/Musolino_1998.pdf.

Julien Musolino. 2006. Structure and meaning in the acquisition of scope. In *Semantics in Acquisition*, Springer, pages 141–166. https://doi.org/10.1007/1-4020-4485-2_6.

Julien Musolino and Jeffrey Lidz. 2003. The scope of isomorphism: Turning adults into children. *Language Acquisition* 11(4):277–291. https://doi.org/10.1207/s15327817la1104_3.

Julien Musolino and Jeffrey Lidz. 2006. Why children aren't universally successful with quantification. *Linguistics* 44(4):817–852. https://doi.org/10.1515/LING.2006.026.

K.J. Savinelli, Gregory Scontras, and Lisa Pearl. 2017. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 39, pages 3064–3069. https://mindmodeling.org/cogsci2017/papers/0579/paper0579.pdf.

Gregory Scontras and Noah D. Goodman. 2017. Resolving uncertainty in plural predication. *Cognition* 168:294–311. https://doi.org/10.1016/j.cognition.2017.07.002.

Benjamin Spector. 2013. Bare numerals and scalar implicatures. *Language and Linguistics Compass* 7(5):273–294. https://doi.org/10.1111/lnc3.12018.

Michael Henry Tessler and Noah D. Goodman. 2016. A pragmatic theory of generic language. http://arxiv.org/abs/1608.02926.

Joshua Viau, Jeffrey Lidz, and Julien Musolino. 2010. Priming of abstract logical representations in 4-year-olds. *Language Acquisition* 17(1-2):26–50. https://doi.org/10.1080/10489221003620946.

Peter C Wason. 1965. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior* 4(1):7–11. https://doi.org/10.1016/S0022-5371(65)80060-3.