frontiers

Scalar inferences in the acquisition of *even*

Yadav Gowda [1], **Elise Newman** [1], **Leo Rosenstein** [1] **and Martin Hackl** [1]

[1] *Department of Linguistics and Philosophy, MIT, Cambridge , MA , USA*

Correspondence*:
Yadav Gowda
ysg@mit.edu

Elise Newman
esnewman@mit.edu

ABSTRACT

Scalar inferences are ubiquitous in human reasoning. Correspondingly, language has many means of expressing and encoding them. One of these means is the focus particle *even*, which utilizes scalar inferences to signal the pragmatic status of asserted content as noteworthy. The vehicles that *even* employs to signal noteworthiness are scalar likelihood inferences. A peculiarity of these inferences is that they are presuppositional in nature (not-at-issue) and yet, they are responsive to the polarity of the sentence expressing the proposition whose likelihood is signaled. This property raises intricate questions about what learners might expect scalar operators of this sort to look like (initial hypothesis space) as well as what type of evidence and learning strategies they have access to as they figure out the specific properties of *even* in adult English. This paper presents a detailed study of this development, combining data from a series of comprehension experiments and corpus studies. We find that children are sensitive to the basic scalar nature of *even* much earlier than previous literature has claimed. We additionally find, however, that children sometimes exhibit non-adult-like responses to *even* sentences, which we argue provide insight into their developing grammar. On this view, the child grammar offers a larger option space for *even* than the adult grammar. Becoming adult-like, in turn, involves eliminating some of these options, namely those that are underutilized in production due to their limited conversational value.

Keywords: Language Acquisition, *Even*, Scalar Inferences, Presupposition, Focus Particles, Additive Particles, Polarity

1 INTRODUCTION

This paper investigates the status of scalar inferences in child grammar through an acquisition study of the scalar focus particle *even*. English *even* provides a window through which to study the grammatical encoding of scalar inferences due to its sensitivity to polarity. It triggers a least-likely presupposition in positive sentences, and a most-likely presupposition in negative sentences as illustrated in (1) and (2), (Horn (1969), Karttunen and Peters (1979), a.o.).

(1) Jack had even invited AMY to the party.
- Amy was the least-likely for Jack to invite to the party.
- There was at least one other person (out of a salient set) that Jack had invited to the party.

(2) Kim hadn't even invited SAM to the party.
- Sam was the most-likely for Kim to invite to the party.
- There was nobody else (out of a salient set) that Kim had invited to the party.

The meaning conveyed in (1) and (2) consists of three components: the asserted content, which is just the ordinary meaning of the sentence without *even* (i.e. the meaning of prejacent), and two invited inferences – scalar and additive in nature – which do not have the status of at-issue content. Rather they exhibit signature properties of presuppositions (or conventional implicatures), which persist even when the presupposition trigger occurs in the scope of entailment-cancelling operators. This can be readily seen when transforming these sentences into questions, (3) and (4), which still give rise to the same invited inferences.[1]

(3) Did Jack even invite AMY to the party?
- Amy was the least-likely for Jack to invite to the party.
- There was at least one other person (out of a salient set) that Jack had invited to the party.

(4) Did Kim not even invite SAM to the party?
- Sam was the most-likely for Kim to invite to the party.
- There was nobody else (out of a salient set) that Kim has invited to the party.

Interestingly, even though the invited inferences triggered by *even* project in questions, their precise character is affected by the polarity of the host sentence. Concentrating on the scalar inferences, which are the focus of this paper, we observed in (1) and (2) that the presence or absence of negation corresponds to a "most- or least-likely" inference respectively. The fact that these are different seems to suggest that the content upon which the scalar inference is based *is* visible to negation, though not in a way that would cancel the inference as is the case for asserted content.

This interaction between likelihood presupposition and polarity has generated much discussion in the theoretical literature. It also raises rather interesting questions about how learners acquire the full pattern of invited inferences triggered by *even*: What is the initial hypothesis space that learners need to navigate as they acquire *even*?[2] What is the evidence they have access to and make use of to transform their initial state of the grammar into the adult state, and what are the learning strategies that allow them to do that? We aim to address these questions via a series of comprehension and corpus studies (complemented by adult control experiments) that allow us to establish the developmental path children follow as they learn to track the different likelihood inferences triggered by *even* across different environments. Our findings portray a much richer and more intricate acquisition process than previous work has suggested, making it possible to identify the specific structure of the initial hypothesis space for *even* as well as the grammatical and pragmatic factors that enable and constrain the learning.

To foreshadow, our comprehension studies demonstrate that preschool-aged children show clear evidence of sensitivity to scalar inferences associated with *even*, contrary to what previous research has suggested. They also reveal a rather nuanced developmental trajectory traceable already in 4yos regarding the at times non-adult-like nature of the scalar inferences. Importantly, our comprehension tasks reveal that not all non-adult-like behavior should be given the same analysis. While certain responses appear to indicate true confusion about *even*, other responses appear systematic and informed by their developing grammar, despite being non-adult-like. This is most clearly evident from the fact that these responses are accompanied by reasoned justifications involving reference to scalar properties.

[1] There is a lively debate as to the precise content, origin and status of the scalar and additive inferences. Since much of this debate is orthogonal to our purpose here, we label them either as "invited inferences" or as presuppositions. Moreover, we largely abstract away from the additive inference. See among others Horn (1972, 1989); Karttunen and Peters (1979); Rooth (1985); von Stechow (1991); Krifka (1991); Rullmann (1997, 2007); Herburger (2000); Schwarz (2005); Greenberg (2016, 2018); Francis (2018).

[2] Recent years have seen an impressive growth of cross-linguistic work on additive scalar particles like English *even* (Giannakidou, 2007; Crnič, 2009; Gast and van der Auwera, 2011; Greenberg, 2015, 2018, a.o.). A more complete framing of the initial hypothesis space for *even* would situate the current discussion within that typology. However, since the cross-linguistic inventory of such particles is rather rich and there is no consensus yet on the organizational principles spanning that typology, we will have to leave it to future work.

Our adult control studies present a similar finding. They, too, show a systematic error pattern during real time comprehension (albeit at a lower rate) which appears to be a function of the adult-grammar of *even*. We argue on the basis of their systematicity and their similarities that both the non-adult-like inferences generated by children and the corresponding errors in our adult studies reflect options made available by the basic architecture of the respective grammars of *even*. Hence, they should inform our theories of adult and child *even*, and put together, they should frame accounts of how *even* is acquired.

Our conclusions from the comprehension experiments are further enriched by our corpus studies on child and child-directed adult use of *even*, presented in Section 5. We find that children who produce *even* do so essentially error-free, even at 3-4 years of age. We additionally find that the form of our stimuli in the comprehension studies instantiates a low-frequency use pattern for *even* in both children and adults. The fact that children nevertheless show partial command of *even* when their comprehension is tested on these items suggests that children in our age range already have a quite robust grasp of the fundamental fact the *even* always triggers a scalar inference of some sort.

Most striking about the corpus data is, however, that children (and adults) do not exhibit the non-adult-like behavior that we find in the comprehension experiments. At first sight, this constitutes a rather surprising production-comprehension asymmetry: the child grammar appears adult-like in production but non-adult-like in comprehension. We propose to resolve this puzzle by assuming that it is the comprehension data that faithfully reflect the child grammar of *even*, which is non-adult-like in that it provides a larger space of options for *even* than the adult grammar. The fact that their production data is fully adult-like should, in turn, be seen as the result of child speakers choosing not to realize some of the grammatically licit options. We suggest that they underutilize some of these options because of their limited pragmatic utility. To explain why production is ahead of comprehension, we suggest that the pragmatic oddity of these options is more transparent from the perspective of the (child) speaker than from the perspective of the (child) listener since the speaker knows the intended message while the listener needs to infer it from the form of the utterance and the presumed conversational goals of the speaker.

2 *EVEN*'S SCALAR INFERENCES: THEORY AND ACQUISITION

2.1 Theoretical background

Our study's main focus is the interaction between *even*'s scalar inferences and sentential polarity as it relates to acquisition. As argued in Karttunen and Peters (1979) the scalar inferences have the status of a presupposition/conventional implicature. This can be inferred, for instance, from the fact that, unlike at issue content, they survive embedding inside entailment cancelling environments such as questions, (3-4). Interestingly, and somewhat unexpectedly the precise nature of the scalar inference is sensitive to the presence of negation (an entailment cancelling operator in its own right). Recalling the examples in (1) and (2) we see that the scalar inference in (2) is substantially different from the one in (1).

(1) Jack had even invited AMY to the party.
- Amy was the least-likely for Jack to invite to the party.

(2) Kim hadn't even invited SAM to the party.
- Sam was the most-likely for Kim to invite to the party.

In both (1) and (2), we detect an inference about the likelihood of somebody having been invited to a party. However, the inference in (2) is essentially the opposite of that in (1). While Amy is understood to be someone that was unlikely to be invited, Sam is understood to be someone who was very likely to be

invited. Both (1) and (2) therefore convey something surprising, namely that someone was invited who wasn't expected to be invited, or that someone wasn't invited who *was* expected to be invited.

There are two families of approaches that attempt to capture this property of *even*. The first type of approach (Karttunen and Peters 1979, henceforth the *scope theory*) argues that *even* uniformly triggers a least-likely presupposition, but has a requirement that it outscope (clause-mate) negation[3] resulting in most-likely inference. We can schematize this approach as in (5)-(7). *Even* is assumed to be a clausal operator that combines with a propositional argument (the prejacent) and a set of alternative propositions (derived from the prejacent via focus semantics[4]). Its lexical semantics is that of a filter which passes on the meaning of its propositional argument p unchanged but only if p is the least likely member in C, (6). For positive sentences this delivers a least-likely inference. For negated sentences, however, with *even* scoping over negation, as shown in (7) its prejacent will express a negative proposition resulting in a "least-likely-to-not" inference, which is, of course, equivalent to "most-likely-to" inference.

(5) [Even [(NOT) S]]

(6) $[\![\text{even}]\!]^{w,g}(C)(p) \Leftrightarrow$ p is the least likely member of C . p[5]

(7) [Even [Kim had not invited Sam to the party]]
- Least-likely inference: *That Kim hadn't invited Sam to the party* is the least-likely proposition of the form *that Kim hadn't invited X to the party* $\Leftrightarrow$ Most-likely inference: *That Kim invited Sam to the party* was most-likely.

The second approach (Rooth 1985, henceforth the *ambiguity theory*)[6] proposes that *even* is lexically ambiguous. One lexical entry is the one assumed by the scope theory, which has an (in principle) unrestricted distribution. The other lexical item, however, is a Negative Polarity Item (NPI) that comes with a most-likely presupposition, (9), and is restricted in it's distribution to occur below negation, (8)-(10).

(8) [NOT [even_{NPI} [S]]]

(9) $[\![\text{even}_{NPI}]\!]^{w,g}(C)(p) \Leftrightarrow$ p is the most likely member of C . p

(10) [NOT [even_{NPI} [Kim had invited Sam to the party]]]
- Most-likely inference: *That Kim had invited Sam to the part* is the most likely proposition of the form *that Kim had invited X to the party.*

The prejacent of $\textit{even}_{NPI}$ in (10) is a clausal constituent without negation. The scalar inference will therefore target a positive proposition, which will yield a most-likely inference since the presuppositional requirement of $\textit{even}_{NPI}$ demands of p to be the most likely element in C. On the ambiguity theory, then, the inferences attributed to *even* in various environments are determined by a lexical specification of NPI-hood rather than by scope. $\textit{Even}_{NPI}$ appears in contexts where NPIs are licensed, and is specified

[3] See (Wilkinson, 1996; Lahiri, 1998; Guerzoni, 2004, a.o.). A complete description says that *even* needs to outscope *all* downward-entailing operators, not just negation. Consider example (i) where *even* is inside a conditional. An in situ treatment of *even* makes the wrong (and contradictory) prediction that Mary is less likely to notice one mistake than she is to notice multiple mistakes. In fact, we infer that she is *most-likely* to infer one mistake over many.

(i) If Mary noticed even one$_F$ mistake of yours, it would be a problem. (Guerzoni, 2004)
- Mary noticing one mistake is most-likely compared to noticing multiple mistakes.

On the scope theory, the most-likely inference in (i) is explained if *even* moves outside the scope of the conditional.

[4] Since the fine mechanics of focus prosody and focus semantics are not central to our paper we abstract away from the details here and refer the reader to e.g. Rooth (1996) etc.

[5] Notation: $[\![\alpha]\!] = \phi.\psi$ states that the semantic value of α is defined only if ϕ and when defined $[\![\alpha]\!] = \psi$.

[6] See (Rullmann, 1997, 2007, a.o.).

to trigger a most-likely inference. In a context where an NPI would not be licensed, regular *even* is used, giving rise to a least-likely inference.

Both of these theories can successfully analyze the examples we have seen so far, and much continued debate attempts to distinguish them. Of interest here is that they differ in which parts of the grammar are responsible for a given inference. On the scope theory, examples like (1) carry an unambiguous least-likely inference because *even* always has a least-likely presupposition. On the ambiguity theory, examples like (1) are unambiguous due to an additional component of the grammar, namely a constraint on the distribution of NPIs.

The unambiguous most-likely inference in (2) can likewise be accounted for by both proposals. However, in this case, both proposals depend on an additional component of the grammar, extrinsic to the lexical specification of *even*, to rule out ambiguity. Relying solely on the basic meaning of *even* as a sentential operator, which requires only that *even* combines with a propositional node, makes a scope position below negation (a propositional operator in its own right) in principle suitable on both theories. Thus, both theories need to appeal to a mechanism extrinsic to the lexical specification of *even* that prevents a least-likely inference to surface.

Anticipating the format of our experimental material, we illustrate this point with a sentence that employs *even* in pre-subject position to the left of negation, (11). Both theories can explain the attested most-likely-to/least-likely-to-not inference by assuming that (regular) *even* scopes above negation, (11a). However, both theories also need to explain why a logical form where (regular) *even* scopes below negation, which would give rise to a least-likely-to inference, is not attested, (11b).[7]

(11) Even AMY wasn't invited to the party.

a. [Even [NOT [AMY was invited to the party]]]

b. * [NOT [even [AMY was invited to the party]]]

– Scope Theory: violates scope constraint for *even*

– Ambiguity Theory: Blocked by $even_{NPI}$

On the scope theory (11b) is ruled out by a rather specific prohibition against interpreting *even* below clause-mate negation. On the ambiguity theory, a blocking principle of some sort is required to ensure that the availability of $even_{NPI}$ preempts regular *even* from being inserted in this environment (e.g. because more specified lexical items, $even_{NPI}$ in this case, are prioritized by principles of vocabulary insertion over less specified lexical items, non-NPI *even*, along the lines of Halle and Marantz (1993)).

2.2 Predicting the acquisition profile of *even*

Given two possible inferences (least-likely and most-likely) and two types of environments (positive and negative, or more generally, upward and downward entailing[8]), we might expect a learner to consider four possible ways to use *even*. A learner who hypothesizes that *even* is polysemous between a most/least-likely inference may start with all four use-patterns in the space (c.f. Giannakidou (2007) on Greek) (Table 1A). A learner who entertains only a least-likely inference associated with *even* (along the lines of the scope theory), however, should only consider three at the outset (Table 1B).

[7] The precise mechanism (e.g. some form of reconstruction) by which such a logical form might be generated is not important to our purpose.

[8] In this paper, we abstract way from the interpretation of *even* in non-monotonic contexts, see e.g. Crnič (2014).

The acquisition path that each of these hypothetical learners takes is expected to be different because they have different starting points. When confronted with a positive *even*-sentence, (12), a learner who hypothesizes a uniform least-likely inference is expected to appear adult-like, because they should only detect a least-likely inference in this context. By contrast, a learner who entertains a polysemous *even* should find sentences like (12) ambiguous between a most-likely/least-likely interpretation. The profile of this type of learner is therefore expected to be non-uniform – at-times adult-like and at-times non-adult-like depending on how they choose to resolve the ambiguity.

(12) Even AMY was invited to the party.

– Uniform least-likely *even*: adult-like
– Polysemous *even*: adult-like or non-adult-like depending ambiguity resolution

(13) Even AMY wasn't invited to the party.

– Uniform least-likely *even*: adult-like or non-adult-like depending on scope
– Polysemous *even*: adult-like or non-adult-like depending on ambiguity resolution

When confronted with a negative *even* sentence, (13), however, both learners are predicted to detect an ambiguous most-likely/least-likely inference. For the learner who posits a uniform least-likely inference, this is because *even* has two in principle available scope positions with respect to negation. For the polysemous *even* learner, they always posit these two inferences regardless of scope.

For a uniform least-likely *even* learner, getting to the adult-like pattern, then, amounts to ruling out ambiguity in negative environments, presumably by expunging the low scope option for *even* in those contexts. For a polysemous *even* learner, they must rule out ambiguity in both polarity environments to arrive at the adult pattern. And they must do so in a specific way: in positive environments the most-likely variant of *even* must be expunged while in negative environments the least-likely variant is targeted.[9]

These different learning profiles should be detectable in a well-controlled acquisition study as soon as a child has the machinery to reason about likelihood, and identifies *even* as a linguistic device used to convey likelihood inferences. To our knowledge, however, this has not been successfully investigated. Rather, existing acquisition studies of *even* aim to address a coarser question: *do children know* even? The findings of these studies have been taken to indicate that children struggle so much with *even*'s focus and scalar properties, that it is unreasonable to think one could ever glean the specific shape of the child's hypothesis space for the likelihood inferences associated with *even*.

We will argue, however, that these conclusions are not well supported by the experimental record. Additionally, we will show that controlling for previously uncontrolled experimental factors reveals a more intricate structure of the development of the grammar and learning path of *even*. In particular, we uncover evidence that children as young as 4 years of age *do* entertain a space like those in Table 1, and moreover that they start with a space that resembles the ambiguity theory of *even*.

2.3 Previous work: Kim (2011)

Motivation for our study, as well as inspiration for its design, is due in part to results from Kim (2011), which to our knowledge is the first systematic investigation of *even* in positive and negative environments in child language. Kim framed her interest in *even* within the context of prior studies on the acquisition

[9] There is hypothetically a third possibility, which is that a child first hypothesizes a uniform most-likely inference associated with *even*. The learning path of a learner with this profile is unpredictable because the child would effectively have to start over at some point by adding to their hypothesis space.

of scalar implicatures. Echoing the consensus view of work in that domain, Kim argues that children as old as 5 are essentially ignorant about *even*. However, we find her results to be ambiguous between two interpretations: 1) children ages 4-5 do not detect *even*'s scalar inferences, or 2) children ages 4-5 *do* detect *even*'s scalar inferences, but nevertheless exhibit non-adult-like behavior that is invited by their developing grammar of *even*.

Kim (2011) conducted a comprehension experiment employing a forced-choice task, in which the experimenter tells children stories about 3 characters who are all different sizes. In each story, all of the characters are supposed to do a task, which scales in difficulty with the size of the character. At the end of the stories, either all of the characters succeed or they all fail. The end of the story is accompanied by a prompt of the form, *Even X was/n't able to do Y*. With the help of a puppet, the children are asked to point to X. Figure 1 shows a sample story setup about bears reaching for cookies on a shelf, with both possible story outcomes.

Kim tested 30 children on three positive/negative story pairs of this sort, yielding 90 responses, where each 'response' corresponds to a pair of responses to the positive/negative versions of each story (Table 2). There were 3 distinct response profiles corresponding to a given positive/negative pair. Some were completely adult-like ("target characters for both sentence types"), some gave opposite of adult-like responses by choosing the tallest character in positive environments and the shortest character in negative environments ("opposite characters for both sentence types"), and some uniformly chose either the rightmost or leftmost character, regardless of polarity ("always rightmost or leftmost character"). Here "both sentence types" refers to both positive and negative prompts (i.e. *even X was able to do Y/even X wasn't able to do Y*).

In addition, Kim ran a control version of the experiment in which *even* was removed from the prompts. No other changes were made to the stories or the task in the control experiment. For example, in the bear story the prompt might be *Larry was (not) able to reach the cookie*. The control study has the interesting property that an adult would presumably struggle to find a felicitous answer to the question, "Who is Larry?", given that "Larry was (not) able to reach the cookie" is equally true of every character. Despite the pragmatic oddity of the task, and the lack of an "adult-like" target answer, Kim shows that children performed very similarly in the control experiment as they did in the main experiment with *even* (Table 2).

The fact that children showed low rates of adult-like responses in the *even* experiment, as well as the fact their responses didn't change substantially when *even* was removed in the control experiment, led Kim to conclude that children essentially ignore *even* when participating in these comprehension tasks. She therefore concludes that the children in this age group do not understand *even*.

We think, however, that her results are compatible with an alternative explanation and so are not compelled to accept her conclusion. Notice that in her main experiment, none of the children chose the middle character. All response profiles included one or both of the extrema, but never the middle. This is unexpected if their selection were truly random. While it is possible to conclude with Kim that this is an accidental result, or a product of the pragmatics of the task, in an unpublished manuscript, Kenyon Branan argues that this pattern could also reflect a feature of the developing grammar. In other words, it might be the case that children don't choose the middle character because they know that *even* triggers either a least-likely or a most-likely inference, neither of which supports picking the middle character. What they don't know, on this conjecture, are the grammatical conditions that control in the adult grammar which likelihood inference is triggered in which environment.

Looking closer at Kim's design and procedures reinforces being more cautious in interpreting her results. Note first that Kim's child participants were asked to point out both extrema characters but never the middle

character during the story leading up to the *even* sentence. This introduces a potential bias in favor of the extrema which makes it difficult to assess whether this gap should be seen as an experimental artifact or as a reflection of a non-adult-like grammar.

Second, the specific choices Kim made in the design of her materials introduce a potential confound. Concretely, her experiment employed three story types, 2 of which involved reaching tasks, and one of which was a lifting task. For both reaching and lifting stories, the likelihood of success is *directly* proportional to the size of the character. Larger characters are both taller and stronger, and are thus, all else being equal, more likely to succeed at reaching something tall or lifting something heavy. An adult-like response to these story types thus amounts to choosing the smallest character in positive environments (least likely to succeed), and the largest character in negative environments (least likely to not succeed). However, this set up introduces a confound for children who showed a preference for the rightmost or the leftmost character. Such a preference could be interpreted in multiple ways. One interpretation is that children have an irrelevant preference for either the smallest or the largest character (or the rightmost or leftmost character). This interpretation is compatible with Kim's conclusion that children ignore *even* completely when doing the task.

However, this behavior is also explained if children are accessing likelihood inferences that are not detectable in adult language, but are present in their hypothesis space for *even*. If for any given example, a child is aware that both a most- and a least-likely inference is in principle available, the child might sometimes choose the most-likely character to succeed, and sometimes the least-likely character to succeed, irrespective of the polarity of the sentence. Choosing either the most- or least-likely character uniformly in both polarity environments would amount to a right-most or left-most preference, in which case this response pattern should not be treated as evidence of naivety. In sum, then, Kim's observations are amenable to two quite different explanations and additional work is required to decide which one is on the right track.

3 COMPREHENSION EXPERIMENT 1

Our own comprehension experiments adopt Kim's basic setup, which we think is quite elegant and offers a very natural way of testing children's comprehension of *even*. We are, however, implementing a number of modifications to help overcome the aforementioned limitations in interpreting her results.

3.1 Methods

Experiment 1 adopts the basic format of Kim's set-up, but includes the modifications and amendments summarized below. Examples of stimuli used in Experiment 1 can be seen in Figure 2.

1. All characters are equally prominent in the story dialogue to avoid biasing children towards the extrema. Unlike Kim's dialogue, which singles out the extrema characters ("Who is the shortest [bear]? Who is the tallest [bear]?"), our dialogue gives equal weight to all three characters ("This story is about 3 squirrel brothers. There's a little one, a medium one, and a big one.")
2. We include filler stimuli without *even* where the middle character is the correct answer, both to test children's alertness and to give them an opportunity to see that the middle character is a possible answer (Supplementary Material; Figure S1). The filler stimuli involve matching a character to an object based on attributes like color and size.
3. Children are asked to justify their responses (e.g. "How did you know that was Larry?")

4. There are two additional story types in which likelihood to succeed scales *inversely* with the size of the character (i.e. fitting and filling stories). This distinguishes a preference for the smallest character from a preference for the least-likely character.[10]
5. The age range of our study is 3-6yo to look for a possible developmental trajectory for *even* (in contrast with Kim's range of 4-5yo).

We recruited 91 English-speaking children ages 3;1-6;11 (mean = 5;0, age in years;months format) from Boston-area daycares, preschools, and through the Living Laboratory program at the Museum of Science, Boston. Three subjects were excluded from the analysis either because they did not complete the task, or because their responses on the control items (as well as justifications) suggested they were not actively participating in the task.

In total, the experiment consisted of 4 filler stimuli and 8 target stimuli: 4 positive target stimuli (1 each for reach/lift/fit/fill) and 4 negative target stimuli (1 each for reach/lift/fit/fill). The experimental items were blocked by polarity of the *even* sentences (NEG vs. POS) and presented in one of two orders (NEG-first vs. POS-first). Children who were assigned to the NEG-first order heard all of the negative stimuli before hearing any positive stimuli, and vice versa.

As a control, we ran a version of this experiment with 68 adult subjects on Mechanical Turk. We slightly modified the stimuli from the child study for use on IBEX (Drummond, 2012), creating an introductory slide which introduced the characters and the situation, and a question slide, which introduced the *even* sentence and asked the participant to identify the named character. Participants were given 10 seconds to respond, starting immediately after the question slide was displayed. IBEX recorded question responses, reaction times, as well as justifications. We followed two exclusion criteria: 1) we excluded participants who incorrectly answered more than one filler item, 2) we excluded participants who gave the same answer (e.g. the middle character) throughout the entire study. After these exclusions, 56 subjects remained.

Data were analysed using the MCMCglmm package in R (Hadfield, 2010) with a mixed-effects multinomial logistic regression. We took response type as the dependent variable, and modeled the fixed effect of polarity and age group {3yo, 4yo, 5yo, 6yo}. Additionally, we modeled a random intercept by subject and by story type. We opted to use a Bayesian approach to our data analysis for two main reasons. First, Bayesian approaches make it relatively easy to specify hierarchical models, such as the multinomial mixed-effects model deployed here, as compared to frequentist data analysis. Second, Bayesian models with maximal random effects structures converge with less data than frequentist models. 4 chains were generated per model, and convergence was tested across these chains using the Gelman-Rubin diagnostic (Gelman et al., 1992; Plummer et al., 2006), as well as by visual inspection of the posterior distributions. Credible Intervals were calculated at 95% with Highest Posterior Density intervals.[11,12,13,14]

[10] A keen reader might notice that the filling story in Figure 2 is in fact ambiguous between two possible likelihood inferences: 1) likelihood of the basket to be filled up (smallest basket = most-likely) vs. 2) likelihood of the character to collect enough for their basket (largest character = most-likely). Indeed we will see that children and adults alike are sensitive to this ambiguity, which added noise to our initial results, and prompted a change to these stimuli in Experiment 2.

[11] For further background on the use of Bayesian methods in the context of linguistics, see Nicenboim and Vasishth (2016).

[12] Credible Intervals, used in Bayesian analysis, are intervals which contain a certain percentage (in this paper, 95%) of the values in the posterior distribution for an unobserved parameter. For instance, a 95% Credible Interval of [10%, 22%] indicates that we can be 95% certain that the value of a parameter falls between 10% and 22%. In this paper, we follow a basic decision rule to take the 95% Credible Interval as an indication of whether a given value can be inferred to be a possible "true" value of a parameter (Kruschke et al., 2012). Thus, the CI here serves a similar role to the *p*-value used in frequentist statistics, in giving us a criterion to reject or accept the null hypothesis.

[13] Highest Posterior Density is a method for selecting Credible Intervals which, as might be expected, involves selecting the highest density intervals within the posterior distribution. Note that because HPD intervals are chosen by density of the posterior distribution, they will *not* center around the mean.

[14] All code used for the analyses in this paper can be found here: `https://github.com/MITLanguageAcquisitionLab/even`

3.2 Results

In our original experimental design, polarity (NEG vs. POS) was a within-subject condition and order (NEG-first vs. POS-first) was a between-subject factor. However, our models failed to converge with order as a factor. Because of this, we chose to only take data from the first four target items from each subject, removing order as a condition, and turning polarity into a between-subject factor. This way we can be sure that our reported results on the interaction between response-type and polarity is not affected by possible within experiment learning or other order effects throughout the experiment.[15]

The results of our experiment differ from Kim's in several ways. The children that we tested offered three types of responses, which we call *adult-like*, *middle*, and *opposite*, corresponding to which character they chose. While Kim's results had no middle responses, in our study, subjects ages 3-5 did sometimes choose the middle character for target items. We also see a steady increase in the number of adult-like responses across age, and a stable number of opposite responses. Figure 3 summarizes the results of Experiment 1.

Figure 3B shows the rate of adult-like responses across age, split by polarity. What we find is that this rate increases much more quickly in negative environments than in positive environments. At age 4, there is a clear difference between the positive and negative environments with respect to adult-like responses ([-3%, 74%]). The profile of *even* comprehension in negative environments is quite stable across ages 4-6 (4yo: [56%, 98%], 5yo: [56%, 97%], 6yo: [63%, 99%]) while adult-like behavior lags in positive environments at age 4, rising steadily until age 6 (4yo: [12%, 74%], 5yo: [48%, 100%], 6yo: [73%, 100%]).

A complementary trend can be seen in the rate of middle responses (Figure 3C), which given that they are never invited by the grammar are an indication of confusion. We see that the number of middle responses remains low, decreasing over time (3yo: [0%, 42%], 4yo: [1%, 27%], 5yo: [0%, 30%], 6yo: [0%, 3%]).

Looking at the last response type, opposite responses, reveals a different pattern from both adult-like and middle responses. Opposite responses are stable across the four age groups, hovering at approximately 20-25% (Figure 3D) (3yo: [0%, 49%], 4yo: [5%, 38%] 5yo: [0%, 17%], 6yo: [0%, 22%]). Additionally, we do not see a polarity effect for opposite responses.

We also asked children to justify their answers and coded their responses as *scalar, random* or *none* to reflect whether their justifications contained evidence of scalar reasoning. In general, all justifications that referred to the size of the characters were coded as "scalar", and all responses were coded jointly by the two first authors. Some sample justifications that we coded as "scalar" include: "Because it's rare that a tiny thing can lift a big thing", "teeny one", "small mouses can usually fit", "because it's the biggest", etc.

Children also often provided reasons that did not reference a discernible scale, which we coded as "random". Some sample "random" justifications include: "look at the pink bunny!", "because I just knew it", "he's two (years old)", "that one has a little bow", etc. Some children were too shy to offer a justification, or stated that they didn't know why they chose a particular character, in which case we coded their responses as "none".

[15] Though our experimental design included filler items for the purpose of potentially excluding inattentive participants, a problem with one of them prevented us from using these filler items as grounds for excluding subjects in our analysis of Experiment 1. Our original exclusion criteria would have excluded children who answered more than 1/4 filler items incorrectly (allowing for occasional but not systematic lapses in judgment). However, our child participants systematically struggled with one of the filler items, even when it was clear that they were attentive throughout the experiment. It was therefore clear that we should remove that item from our consideration for exclusion. However, doing so increased the likelihood of exclusion from 2/4 to 2/3, which we felt was too significant a difference to implement without probable cause. We therefore chose to be maximally inclusive of all subjects who completed the study.

Table 3 summarizes the number of justifications of each type that were provided for each response pattern. Notably, scalar justifications were predominantly offered for adult-like and opposite responses, while random justifications were the most frequent justification type for middle responses.

The overall results of the adult control study appear in Table 4. The adults, as expected, predominantly pick the least-likely character in positive environments (88%) and the most-likely character in negative environments (82%). This shows a slight asymmetry in favor of positive stimuli – however, it is not statistically significant.[16]

3.3 Analysis and discussion

We want to draw attention to several features of these results, some of which resolve open questions in Kim's experiment, and some of which raise new ones.

First, we argue that these results refute Kim's conclusion that children ages 4-5 show no evidence of comprehending *even*. Our results show a clear upward trajectory in the rate of adult-like responses between the ages 3-6, as seen in Figure 3A.

Furthermore, our results provide a clear indication as to when children begin to reliably exhibit sensitivity to the scalar properties of *even*. Given that there are three possible response types, chance behavior should be 33%. While 3yos do not give adult-like responses at an above chance rate ([17%, 83%]), starting at 4 years old, children provide adult-like responses at well above chance, nearing adult-like levels by age 6 (4yo: [45%, 89%], 5yo: [60%, 100%], 6yo: [76%, 100%]).

An analysis of the mean difference between the rate of adult-like, middle and opposite responses across ages shows that by age 6, the rate of adult-like responses has increased significantly since ages 3-4, while the rate of middle responses has decreased significantly (Table 5A). These results reinforce our claim that children progressively acquire an adult-like understanding of *even*.

Looking just at the progression of adult-like responses in negative contexts, we see that the only significant difference is between the 3yo on the one hand, and the 4-, 5-, and 6yo on the other (Table 5B). There is no significant difference between the rate at which 4yo provide adult-like responses to negative stimuli and the rate at which 5- and 6yo do, suggesting that children acquire negative *even* sentences earlier than positive ones. This result contradicts the expectation that negation adds a computational cost. Not only are children unphased by the addition of negation, but they appear to have an easier time deducing the right inference in its presence.

Looking at adults in the control study, we do not see a polarity effect. There is no significant difference in rates of different response types between positive and negative stimuli.

Adults and children both show two types of error behavior, with a higher rate of opposite responses. In adults, the rate of middle responses is extremely low, but the rate of opposite responses is nearly as high as the rate of opposite responses in 6yo.[17]

The contrasting behavior of middle vs. opposite responses in both adults and children seems to indicate that the learning trajectory for *even* contains a stage at which elements of the grammar of *even* are already in place, but are not yet fully adult-like. The fact that middle responses disappear over time but opposite

[16] A similar asymmetry in the same direction can be seen in terms of response times, with negative stimuli taking longer than positive stimuli, but this difference is again not statistically significant.

[17] Just as in the child study, we asked adults to justify their responses. Unsurprisingly since almost all errors were opposite responses, all of the justifications were scalar except for two participants that chose the middle character and wrote *Just a random guess* and *I don't know* respectively.

responses are stable through age 6 suggests that opposite responses are principled at this stage, while middle responses are not. In sum, we propose to analyze middle responses as indications of genuine confusion, while opposite responses should be analyzed as licensed by the developing grammar of *even*. Moreover, given that they are detectable at the earliest stages where children exhibit a stable appreciation of *even*, we propose to analyze them as an indication of what the initial hypothesis space for *even* looks like, Table 1A.

The decrease in middle responses over time correlates with the increase in adult-like behavior, which supports our view of middle responses as a measure of confusion. As children become more adult-like, they become less confused. We suggest that by contrast, opposite responses are not an indicator of confusion and thus do not decrease noticeably as adult-like behavior increases. Notice that they persist to some extent even in adults.

This characterization of the middle and opposite responses is supported by the justifications that children provided for each response type. Modeling the rate of justification types by response type shows that children provide significantly more scalar justifications to adult-like and opposite responses as compared to middle responses, as seen in Table 5C. Furthermore, there is no significant difference in the rate of scalar justifications between adult-like and opposite responses. The stability of opposite responses throughout our above-chance-performing age ranges (4-6yos), combined with how well-reasoned their justifications are, suggests that children know that *even* is associated with a space like Table 1A.

Recall from Section 2 that the in principle space of inferences associated with *even* on the *adult* grammar included opposite inferences, which motivated the notion of the hypothesis space explored here. This prediction is readily borne out in our adult study; adults essentially only make errors in the form of opposite responses, which is expected given their grammar. The exciting finding from our child study is that children *also* favor this error behavior over other potential error patterns, suggesting that they too access a space like Table 1A. That said, our results contain some unexplained noise that merits further discussion. We therefore conducted a second comprehension study, whose main purpose was to investigate the potential sources of noise and their impact on our results.

One major source of noise in our data was the amount of variation in adult-like behavior across the different story types. In particular, the filling stories were accompanied by more non-adult-like responses than the other stories (Supplementary Material; Figures S2, S3).

A clue for this pattern comes from the justifications provided by children and adults. Specifically, our subjects frequently indicated that there was a salient alternative interpretation of the filling stories. On this alternative interpretation, apparent "opposite" responses are actually adult-like, suggesting that our initial reported rates of error responses were actually somewhat inflated.

In the filling stories, the smallest character should have the easiest time filling their basket because their basket is the smallest. On an alternative construal, however, the smallest character might have the most difficulty collecting the requisite number of acorns (because of their age, or general inexperience) necessary to fill their basket. Several adult and child subjects interpreted the filling stories on the latter characterization, providing well-reasoned justifications like those in (14).[18]

(14) *Adult-like justifications for opposite responses in filling stories*

a. "The first squirrel is Sammy because the youngest one would probably have the hardest time filling his basket and I assume the smallest squirrel to be the youngest." (adult)

[18] Note that one of the children used *even* in an adult-like manner to justify the opposite response (14b).

b. "because even though his basket is little he can even fit a lot of acorns in there" (child)

Experiment 2 therefore includes new stimuli to replace the ambiguous filling stories, in order to eliminate this source of noise.

An additional concern about our experimental design pertains to the role of the ability modal in the prompts. Recall that our test subjects were asked to respond to prompts of the form *Even* X_F *was/n't able to do Y*. This element of the design was inherited from Kim's experiment for the purpose of attempting to replicate her results. However, now that we have failed to replicate her results, we return to this feature of the design with some scrutiny. Of interest is the fact that ability is itself a gradable notion. To interpret a sentence with an ability modal, one must therefore have access to a kind of scalar reasoning.[19] A question we should ask, then, is whether the overt expression of ability accounts for any of the children's behavior on its own. To investigate the relevance of this possible confound, we removed any overt scalar notions from our stimuli in Experiment 2.[20]

4 COMPREHENSION EXPERIMENT 2

4.1 Methods

Experiment 2 contains two large scale changes to the stimuli from Experiment 1: the removal of the ability modal in the prompts, (15), and the replacement of the filling stories with *spilling* stories (Figure 4).

(15) Sample prompts from Experiment 2 (ability modal removed)

a. Even Benny$_F$ has gotten an apple!

b. Even Franky$_F$ hasn't gotten the socks on!

The spilling stories are about likelihood to spill/drop something heavy. Like the fitting stories, height is inversely proportional to likelihood to spill. The smallest character is the most likely to be the weakest, and therefore the most likely to drop or spill the heavy bucket/basket.

These two changes combined are expected to avoid the problems of the previous filling stories because they remove the ambiguity about whether the likelihood scale should refer to the abilities of the characters vs. properties of their baskets/cups.

In order to make our data comparable to Experiment 1, we made polarity a between factor. Each subject evaluated eight target stimuli and four filler items, where all target stimuli were of the same polarity environment. This doubles the number of data points for each polarity condition.

We presented these stimuli to subjects in either of two orders, one in which the story types proceeded as *fit, reach, lift, spill*, and the other in reverse order. We collected data from 80 children, ages 3;1-5;11 (mean = 4;7). Unlike in Experiment 1 (see fn. 15), where children systematically struggled with a particular filler item, no such difficulty was apparent in any of the modified filler items in Experiment 2. Thus, we were able to use performance on these filler items as an exclusion criterion in Experiment 2. Subjects who

[19] See Greenberg (2015), anticipated by Rullmann (2007), for an account of *even* that replaces the likelihood scale with a scale introduced by a (contextually salient) degree predicate.

[20] In fact, we already have some reason to believe that the scalar properties of *be able* do not play a role in opposite responses. First, opposite responses are present in our adult data. While it is possible that some adults ignore *even* entirely in our experiment, this seems less likely. Additionally, we attempted to run a version of Experiment 1 with children where we did not pronounce *even* (e.g. *Jessiepillar was able to reach the shelf.*). We ultimately stopped the experiment because most of the children became very confused and didn't want to finish the task. This suggests that children who provide opposite responses are indeed sensitive to the presence of *even*.

failed to correctly answer at least 3 out of 4 filler questions were excluded from our results (14 exclusions), as were those whose justifications and behavior during the study suggested that they were not paying attention to the task (4 exclusions). After exclusions, 62 subjects remained. The remaining data is evenly distributed by polarity, block order, and age group, with five or six subjects per age group/polarity/order combination. We focused on these earlier years because two interesting effects apparent in Experiment 1 either disappeared or stabilized by age 6: 1) the stable preference for opposite responses over middles responses, and 2) the polarity asymmetry in the rate of adult-like responses.

As with Experiment 1, we performed an adult control study with 85 participants on Mechanical Turk with amended stimuli on IBEX (see section 3.1 for information on the changes made). After exclusion criteria were applied, data from 60 participants remained.

Data for the child and adult experiments were again analysed using a mixed-effects multinomial logistic regression. We modeled the fixed effect of polarity and age group as well as order (forward and reverse). Additionally, we modeled a random intercept by subject and a random intercept and slope by story type. See Section 3.1 for additional information on the statistical analysis.

4.2 Results

Figure 5A summarizes the rate of each response type by age for Experiment 2. The rate of adult-like responses again increases steadily with age (3yo: [8%, 62%], 4yo: [53%, 91%], 5yo: [79%, 100%]), while middle responses decrease steadily with age (3yo: [11%, 81%], 4yo: [0%, 22%], 5yo: [0%, 3%]). We also see a fairly stable population of opposite responses (3yo: [3%, 37%], 4yo: [5%, 33%], 5yo: [0%, 18%]).

Looking at each response type individually by polarity reveals that the polarity asymmetry from Experiment 1 is far less pronounced. The rate of adult-like responses is roughly equal between the two polarity conditions in 4yos, which was where the asymmetry was most pronounced in Experiment 1. 5yos do perform noticeably better in negative environments than positive environments, but this result is not statistically significant (Figure 5B). The same pattern holds for middle responses (Figure 5C). There is no polarity asymmetry at ages 3 or 4, but a slight asymmetry becomes visible at age 5. Opposite responses show no sensitivity to polarity, and are even more stable than in Experiment 1 (Figure 5D).

Finally, the justifications for each response type, coded as *scalar/random/none*, are summarized in Table 6. As in Experiment 1, scalar justifications were offered far more for adult-like and opposite responses than for middle responses. Justifications for middle responses were most often random.

A comparison of adult-like behavior across each of the story types shows that there is less variation across items compared to Experiment 1 for both adults and children (Supplementary Material; Figures S4, S5). Most notably, the profile of responses is no longer substantially different for spilling stories than for the other story types.

The results of the adult control study are summarized in Table 7. As in Experiment 1, adults were slightly less error-prone in positive environments.[21] However, this asymmetry is, again, not statistically significant.

4.3 Analysis and discussion

Experiment 2 confirms the main finding from Experiment 1, namely that children begin to comprehend *even* earlier than Kim (2011) suggests. Once again, 3yos do not choose the adult-like response at an above chance rate [8%, 62%], while 4yos [53%, 91%] and 5yos [79%, 100%] perform at well above chance.

[21] As in the adult control version of Experiment 1, there is also a non-statistically significant polarity asymmetry in terms of reaction time.

Looking at mean differences between age groups, 4- and 5-year-olds give adult-like responses at a significantly higher rate than 3-year-olds (Table 8A). However, opposite responses remain steady throughout age groups.

The opposite responses show essentially the same profile in Experiment 2 as they did in Experiment 1. Many of the children gave opposite responses, including the older ones, and these were often accompanied by scalar justifications. The middle responses, on the other hand, showed some sensitivity to age and polarity, and were primarily given random justifications, Table 8B.

Experiment 2 was successful in that it replicated these results from Experiment 1 with substantially less noise. The justifications indicate that both our adult and child participants interpreted *even* primarily based on the size of the characters and not based on any less obvious criteria such as perceived age or general competence. Figures S4 and S5 in the Supplementary Material show that story type also no longer correlates with any particular error pattern.

Experiment 2 also revealed that the polarity asymmetry in child comprehension observed in Experiment 1 was detectable in Experiment 2 only as a trend that did not reach statistical significance. 5yos, in particular, exhibited an advantage for negative environments on adult-like comprehension similar to that observed in Experiment 1 for 4yos. Determining what might be responsible for the similarities and differences of this environmental effect across the two experiments is unclear to us and deserves further investigation.

Interestingly, opposite responses in our adult control exhibit a noticeable sensitivity to polarity. Specifically, opposite responses are observed primarily in negative environments while positive environments generated vanishingly few.[22] After filtering out answers associated with justifications that indicated guessing or another interpretation of the story, there were 10 opposite responses observed in negative environments, but only 2 opposite responses in positive environments.[23,24]

To summarize, both Experiments 1 and 2 have enabled us to identify two distinct error profiles in adults and children. Middle responses indicate simple confusion or inattention. Opposite responses appear to be licensed by the grammar.

An important difference between adult and child error behavior is the polarity sensitivity of opposite responses. Children offered opposite responses in *both* positive and negative environments, suggesting that they access a space of inferences like that in Table 1A. By contrast, adults basically only offered opposite responses in *negative* contexts, suggesting that they access a space like that in Table 1B.

5 CORPUS STUDIES

In order to better understand children's experience with and usage of *even*, we conducted two corpus studies, in which we compare the features of our stimuli to tokens of *even* found in child and adult corpora. Our investigation resulted in three relevant findings:

1. Children produce *even* essentially adult-like (error-free) as early as 3 years of age, hence much earlier than they comprehend *even* in an adult-like manner.

[22] Attempts at modelling this contrast with a multinomial mixed-effects model unfortunately failed to converge.

[23] Our scoring of justifications was charitable towards adult-like interpretations or true guesses so as to minimize the chance of inflating the rate of opposite responses.

[24] Recall that the rate of opposite responses in Experiment 1 was inflated due to the ambiguity of the filling stories. Reanalysis of the results from Experiment 1 that accounts for this inflation in fact demonstrates a similar polarity asymmetry: opposite responses are primarily concentrated in negative environments (3 in POS vs. 17 in NEG).

2. Adults show a use-bias for *even* in negative environments (here, 'use-bias' refers to a bias towards using a word in a particular environment). Children, by contrast, initially favor *even* in positive environments and acquire the adult use-bias for negative *even* in the ages of 4-5.
3. The form of stimuli in our comprehension study instantiate a *low*-frequency use pattern for *even* in child and child-directed speech.

To study the distribution of *even* in both child-produced speech and child-directed speech[25], we examined token instances of *even* in the American English CHILDES corpus (MacWhinney, 2000).[26] Data were analyzed using childes-coder (Gowda, 2020; Sanchez et al., 2018), which presents instances of target words along with their broader contexts, and provides an interface for users to save metadata about these instances in a database (Supplementary Material; Figure S6). Instances of *even* produced by speakers (marked Target_Child for child-produced speech, and marked Mother, Father, Adult, Uncle, Grandmother, Aunt, Grandfather, Family_Friend, or Teacher for child-directed speech) were coded by mutual agreement among the authors for several criteria, including:

1. Presence of negation: {Yes, No, Unclear}
2. Order of negation and *even*: {N/A, *even*-NEG, NEG-*even*, Unclear}
3. Whether negation is sentential: {Yes, No, Unclear}
4. Order of *even* and the subject: {*even*-Subj, Subj-*even*, Unclear}
5. The focus associate of *even*: {Subject, Object, Verb, Adjunct, Unclear}
6. The likelihood inference: {Most-Likely, Least-Likely, Unclear}

In addition to examining the surrounding context of each instance of *even* to determine the correct coding, we also made use of the metadata and audio recordings available in CHILDES.

5.1 Results

Our comprehension studies focused on the interaction between polarity and the likelihood inference in *even* sentences. Because we coded tokens of *even* in production by both polarity and the contextually salient likelihood inference, we can similarly investigate this interaction in production. Results for child-produced speech are summarized in the first four rows of Table 9. Results for child-directed adult speech are in the last row.

In Table 9, for both adults and children, the rates of "opposite" uses of *even* – most-likely inferences for *even* in in positive environments and least-likely inferences in negative environments – are so low that they are essentially absent from our findings. This is quite striking given that we see robust rates of opposite responses in comprehension and suggests that opposite responses are, in fact, a comprehension phenomenon.[27]

Another salient pattern we observe in both child and child-directed corpus data is the overall prevalence of negative-*even* sentences (Table 10). Children start off using *even* in positive environments more often than in negative environments at age 3. According to a chi-squared test for homogeneity, the distribution of

[25] Here, we use 'child-directed speech' to refer to all utterances produced by non-child speakers in the CHILDES corpora. Thus, this data includes all speech in CHILDES that a child was exposed to, not necessarily just speech that was specifically directed at a child.

[26] Due to metadata consistency issues, data from the MacWhinney corpus was excluded from the analysis of both child-produced and child-directed speech.

[27] In actuality, because our coding schema kept track of the polarity of the environment rather than upward/downward entailment, there were higher numbers of most-likely inferences in apparent positive environments, as indicated in Table 9 by the gray figures. However, these are not true opposite responses because most-likely inferences are predicted by the grammar in downward entailing environments more generally, whether or not there is sentential negation. Indeed, a study of most-likely responses in positive sentences for adults shows that they all involve downward entailing environments.

positive and negative uses of *even* in 3yos is significantly different from adults ($p < .001$), while there is no significant difference ($p > .1$) between the distribution of positive and negative *even* sentences in 4, 5, and 6yos and adults. That is, by age 4, children appear to have an adult-like use-bias for *even* in negative environments.

Lastly, we investigated several features of our comprehension study stimuli in the corpus and found that our stimuli instantiate a low-frequency usage pattern for *even*. Our stimuli can be described according to the following feature specification:

1. Focus associate = subject
2. *Even* precedes the subject
3. *Even* precedes negation

We parametrically compared our stimuli to other *even* constructions by coding for whether a given utterance with *even* focused the subject, whether it contained sentential negation, and the linear order of *even*, the focus associate, and negation (if applicable). Table S1 in the Supplementary Material shows that subject focus was relatively infrequent compared to VP or object focus. When the subject was focused, however, pre-subject *even* was preferred to post-subject *even*.

In addition to the placement of the subject with respect to *even*, we can compare the placement of *even* with respect to negation (Supplementary Material; Table S2). *Even* follows negation in the majority of cases. When this requirement interacts with the previous tendency for subject-associating *even* to appear pre-subject, we see a preference for maintaining both constraints, resulting in *not even*. Examples with both subject focus and sentential negation were therefore most often of the form in (16b). Our stimuli, by contrast, took the form in (16a)[28].

(16) a. Even Linda$_F$ didn't write to me. (dispreferred)
b. Not even Linda$_F$ wrote to me. (preferred)

5.2 Analysis and Discussion

In our corpus studies, children as young as 3yo produced *even* as if they were adults.[29] Taking their behavior at face value, and assuming Snyder (2007)'s principle of Grammatical Conservatism, this should not be possible unless they have identified a grammatical basis for *even*'s scalar inferences.

(17) **Grammatical Conservatism**: Children do not begin making productive use of a new grammatical construction in their spontaneous speech until they have both determined that the construction is permitted in the adult grammar, and identified the adult's grammatical basis for it. (Snyder, 2007)

The fact that they exhibit non-adult-like behavior in comprehension should therefore not indicate that they lack a grammar entirely, as is argued by (Ito, 2012; Kim, 2011, e.g.). Indeed we argue that children must have a grammar of sorts for *even*, which happens to invite at times non-adult-like behavior in comprehension. Moreover, children must have some appreciation for which cells in the space of inferences allowed by their grammar are not available to adults, or else they would not be so adult-like in production. We will explore this in Section 6.

[28] Table S1 in the Supplementary Material contains all tokens of *even* in which any type of negation was present; sentential or constituent. *Even-neg* order was mostly available for clauses with constituent rather than sentential negation.

[29] Although in this paper we only present CHILDES data from ages 3-6, adult-like production of *even* is apparent as early as 2 years old, with 32 instances of POS *even* sentences and 16 instances of NEG *even* sentences.

This conclusion is supported by another finding from our corpus studies, namely the fact that our comprehension task stimuli instantiate a low-frequency use pattern for *even* in both child and adult speech. Children's lack of experience with our *even* sentences, however, apparently did not deter them. Children ages 4-6 still offered adult-like responses as the dominant response pattern in our comprehension studies. We must therefore conclude that their command of *even* is quite sophisticated. They are able to abstract away from the particular *even* sentences that they hear and use most frequently, and generalize to other less-frequent uses.

Lastly, our corpus studies offered a perspective on the polarity asymmetry observed in our comprehension studies. Adults apparently produce *even* approximately 1.5 times more often in negative environments than positive environments, which translates into a comparative abundance of negative *even* in children's input. We propose that children show (slightly) better rates of comprehension in negative environments because they have more experience trying to interpret negative *even*. This effect is, however, less pronounced than our other findings because of the natural tension between inherent knowledge and experience. The proposed child grammar of *even* affords children the same abstract knowledge of *even* in negative as well as positive environments, which accounts for their overall adult-like competence. Their performance with *even* in real time, however, can be marginally affected by their confidence with each polarity environment.[30]

6 LEARNING SCALAR INFERENCES

The previous sections presented novel findings from a series of comprehension and corpus studies which significantly enrich our knowledge of the empirical landscape of how *even* is acquired. The picture that emerges reveals a surprisingly intricate developmental path.

Our evidence from the comprehension studies suggests that children as young as 4 years of age systematically draw scalar inferences that are comparable in nature to those generated by the adult grammar of *even*. Interestingly, their knowledge of the environmental conditions controlling when to draw which type of inference (least-likely-to in positive environments and most-likely-to/least-likely-to-not in negative environments) is not completely adult-like at this age. This gives rise to a comprehension behavior that runs at times directly opposite to that of adult speakers.

With regard to the basic effect of polarity on the nature of the scalar inference, we saw a rather striking lack of opposite scalar inferences in the corpus data. Even occurrences of *even* produced by 3yo conform to the adult pattern. This is quite surprising since we know from the comprehension experiments that opposite inferences are allowed by their developing grammar as late as age 6. Furthermore, children ages 4-6 were adult-like in production to the extent that they showed an adult-like use-bias for negative *even*, indicating that they are also sensitive to the conversational settings in which *even* sentences are predominantly used.

To explain why the acquisition of *even* unfolds along such an intricate path is a non-trivial task. It involves specifying the initial hypothesis space that the learners start out with as well as the final state that characterizes the adult grammar of *even*. Furthermore, learners must identify the relevant evidence as well as develop or adopt strategies that together enable them to transform the former cognitive structure into the latter. Though we are not yet in a position to offer a full account of the acquisition of *even* that lives up to all of these demands, we do think that our findings allow us to make significant progress towards that goal.

[30] Recall that this effect is fragile in our experiments (i.e. statistically present in Experiment 1 but only present as a trend in Experiment 2). This fragility might be due to a rather fine-grained effect of experience. It is conceivable that the absence of the ability modal and the presence of the present perfect make our stimuli in Experiment 2 even less well-represented in the input than the stimuli in Experiment 1. If this is true, it would thereby delay the advantage of negative environments to a later age when enough of such cases have been encountered.

6.1 Initial hypothesis space for *even*

The findings from our comprehension studies provide persuasive information about the nature of the initial hypothesis space: it needs to allow for all four combinations of likelihood inferences and polarity of the environment to be expressible by *even*, Table 1A. A simple way of implementing such a grammar would be to postulate a polysemous *even* which can freely occur in positive and negative environments.

Our argument in support of this conclusion is straightforward. We found both adult-like and non-adult-like scalar inferences in the earliest stages of comprehending *even* in both positive and negative environments. In other words, we saw that all four cells in Table 1A are utilized as soon as learners start to appreciate the scalar nature of *even*. Importantly, we saw that all four inference patterns occurred stably throughout an extended period of learning, and we found them to be regularly accompanied by reasoned justifications that referenced the relevant scale properties. This shows that the non-adult-like responses (just like the adult-like responses) were sanctioned by the developing grammar. Thus, they should be analyzed as exemplars that are predicted by the initial hypothesis space rather than as errors whose source is unrelated to the grammar of the learner.

6.2 Adult grammar of *even*

From the perspective of the two competing views on the adult grammar described in Section 2, the initial hypothesis space in Table 1A can be described as an as yet unconstrained form of the grammar predicted by the ambiguity theory. Recall that the main tenet of the ambiguity theory is that *even* can in principle carry a least-likely as well as a most-likely inference. The distribution of these variants needs to be constrained via the addition of a grammatical feature (in this case an NPI-feature on most-likely *even*). Without that addition, the distribution of *even* would remain unconstrained allowing for all logically possible combinations to be realized, Table 1A.

An as yet unconstrained version of the grammar predicted by the scope theory, by contrast, is not a viable option for the initial hypothesis space since it makes only three of the four required cells available, Table 1B. The reason is, again, transparent. The main tenet of the scope theory is that *even* can only carry a least-likely inference. In combination with negation we can generate a most-likely inference if *even* out-scopes negation since the resulting least-likely-to-not inference is equivalent to a most-likely-to inference. However, without the presence of negation, the scope theory can only generate a least-likely inference, which leaves the most-likely inferences in positive environments unaccounted for.

Though we have identified (on empirical grounds) a greater similarity between the initial hypothesis space for *even* and the ambiguity theory, it would be unjustified to conclude at this point that the ambiguity theory has to be correct for the adult grammar of *even*. Both theories of adult *even* are, in fact, compatible with the initial state postulated for children. They simply require different transformations on the initial state. Deciding which one offers the better account for the adult grammar, therefore, depends on what the actual steps are that allow learners to transform the initial hypothesis space (Table 1A) into either the constrained version of Table 1A or Table 1B.

To arrive at an ambiguity grammar of adult *even*, the learner must replace the polysemous *even* with two separate *even*s, each specified for a particular likelihood inference and level of polarity sensitivity (NPI or unmarked). Under the scope theory, learning amounts to eliminating the possibility of *even* triggering a most-likely inference altogether. This is arguably a simpler transformation on the initial hypothesis space. However, it yields the target grammar only at the expense of adopting an unprecedented constraint on the syntactic scope of *even*.

Importantly, both theories also have to explain why other logically possible combinations of likelihood inferences and polarity-sensitivity/scope constraint are never selected by learners of English *even*. Below we argue that a plausible source to rule out unattested combinations is the limited conversational utility of those combinations. Interestingly, these considerations will also provide us with a possible account of why acquiring the adult grammar takes relatively long and what might be responsible for the puzzling production-comprehension asymmetry we have observed.

6.3 Pragmatics of likelihood-inferences

Throughout the paper we have described the scalar inferences triggered by *even* in terms of likelihoods – "least-likely-to" in positive environments and "most-likely-to/least-likely-to-not" in negative environments. In doing so, we adopted the terminology of Karttunen and Peters (1979) which is intuitive and sufficiently transparent to characterize the differences between the various scalar inferences we have encountered. Whether likelihood is (always) the correct way to characterize the dimension of the scalar inferences of *even* is, however, debated in the literature. Alternatives include various formulations of expectedness, noteworthiness, informativity as well as scales introduced by gradable predicates.[31] We cannot provide a full assessment of this debate here. Instead, our aim is to clarify the connection between the inferred relative likelihood of a proposition and its noteworthiness in a given conversational situation. This will be sufficient to diagnose the conversational status of the adult-like as well as the opposite scalar inferences we have observed.

Taking a closer look at the opposite inferences preschoolers draw in our comprehension studies, we observe that they are not simply absent from the adult grammar of *even*, but are in fact conversationally odd if we try to render their content anyway. Compare, for example, (18), which features the content of the adult-like scalar inferences via an appositive relative clause, to its rather odd sounding counterpart in (19), which features the "opposite" content of *even*'s inferences.

(18) Adult-like inferences

a. Everybody has reached the book, including Jessiepillar, who was the least likely to have done so.

b. Nobody has reached the book, including Jessiepillar, who was the most likely to have done so.

(19) Opposite inferences

a. # Everybody has reached the book, including Jessiepillar, who was the *most likely* to have done so.

b. # Nobody has reached the book, including Jessiepillar, who was the *least likely* to have done so.

We propose that a pragmatic explanation of this contrast can provide insight into the factors that help learners constrain their initial hypothesis space for *even*. To see how, let us examine the conversational context in which our *even* sentences were uttered.

Recall that the three characters in our stories either all succeeded at the relevant task or they all failed. This fact was highlighted explicitly with a universal statement immediately preceding the *even* sentence. The truth-conditional content of the *even* sentence was therefore redundant, which puts the burden to

[31] See Fillmore (1965); Fauconnier (1975); Kay (1990); Rullmann (2007); Giannakidou (2007); Greenberg (2016), among others. Refinements of, or alternatives to likelihood-based characterization can be envisioned that are consistent with our findings. Choosing among them is, however, not topical for us here.

provide conversational utility for the utterance squarely on its not-at-issue content. A question we might ask is whether a pragmatic requirement on conversational utility constrains the space of possible likelihood inferences at all.

Given the oddity of (19), we argue that it does. Moreover we propose that this oddity is derived if a connection between a character's likelihood of success with some notion of propositional noteworthiness is important to adult-like competence with *even*. To see why, it is important to consider the context in which the *even* sentence occurs very carefully.

After the universal statement but before the *even* sentence is uttered, the context contains a proposition of the form, *Every x in C is such that x has reached the book*. If the modal horizon against which the likelihood inference were evaluated contained all and only the verifying situations characterized by the universal statement, the likelihood inference would be moot. Comparing the relative likelihood of *Jessiepillar has reached the book* to *Some other x in C has reached the book* is almost nonsensical because both sentences are already true in the context.

What allows the likelihood inference to be meaningful and useful is to consider a context in which it was *not* given that any character would reach a book. In other words, for the *even* sentence to have a sensible inference, it must take as its context variable a set of propositions that does *not* contain the universal statement that preceded the *even* sentence in our experiment. But what drives this move? Why does a listener bother to accommodate a different common ground in which to make sense of the likelihood inference, rather than just ignore it? We propose that this move is connected to a notion of propositional noteworthiness.

We propose that when an adult listener hears an *even* sentence in our context, they detect its redundancy and ask, what makes this proposition worth repeating? I.e. what makes this result surprising or deserving of comment? It is this question that allows the listener to specify a useful common ground in which to consider the likelihood inference. The logic goes as follows: in order for "Jessiepillar has reached the book" to be noteworthy, it must be unexpected. Jessiepillar must therefore be the character whose success was least likely.

Notice, however, that there is no way to connect to a most-likely inference on this logic, hence capturing the oddness of (19). Considering characters who were *likely* to succeed in no way explains why emphasizing those characters' success is interesting.

If this is the type of pragmatic reasoning that adults employ generally in a conversation, opposite responses are predicted to be infelicitous, thus providing insight into why the grammar of *even* is constrained to just two adult-like inferences. Therefore, the content that our child comprehenders end up with when they select the opposite character is odd from the perspective of the adult grammar.

To clarify, for Jessiepillar to be singled out even though her height doesn't justify it, as happens when children provide opposite responses, is of course logically possible.[32] However, for adult speakers this requires a different "backstory", e.g. Jessiepillar might be the most/least motivated of the three to do what it takes to reach the book making her the most/least likely to succeed/fail despite her height. Our stories did not provide any useful information about the characters other than their height, however. Thus, adult comprehenders are stuck with Jessiepillar's height as the only available basis for anchoring the likelihood

[32] Our cases are not characterized by entailment relations between alternatives. In such cases, opposite responses would only be possible at pains of accepting a contradiction. See e.g Lahiri (1998).

inference triggered by *even*. They therefore pick the shortest character to be Jessiepillar when all candidates succeed and the tallest when all of them fail.

For our child comprehenders the situation is different. While they may be practiced enough conversationalists to consider a common ground in which a likelihood inference is meaningful, their grammar overgenerates. Because their grammar allows for a least-likely as well as a most-likely specification of the scalar inference regardless of the polarity of the sentence, they consider both the tallest and the shortest character as grammatically viable candidates for Jessiepillar. This is only possible, however, if we assume that children are less attuned to the relevant conversational pragmatics than adults are, thus allowing both options to remain viable in our conversational setting. Thus, they can in principle choose the character at either end of the scale.

With regard to why children are less adept at this kind of pragmatic reasoning than adults, a number of options seem plausible. For instance, it may be that "explicit" Theory of Mind level reasoning about the motivations of speakers, which is required to detect this sort of pragmatic oddness, is not yet fully developed or sufficiently practiced.[33] Alternatively, or additionally, it may be that children's processing resources for this kind of reasoning are still not up to full capacity. Whatever the true underlying causes, it seems reasonable to characterize their pragmatic reasoning as more tolerant than those of adults towards not knowing exactly what the speaker had in mind when they issued their *even* sentence.[34] A willingness to proceed in a state of partial ignorance leaves both inferences in play making both extrema live possibilities. Of course, this follows only if the developing grammar generates both types of inferences in both environments to begin with. On our proposal this is so because their grammar allows for all four cells of the initial hypothesis space to be expressible by *even*.[35]

The fact that opposite choices occur at all is therefore predicted on the basis of children's more limited conversational experience compared to adults. However, the fact that opposite choices occur less frequently than adult-like choices may be seen as a reflection of those choices being less optimal even from the perspective of the learner. After all, these opposite choices require a willingness to proceed without having figured out exactly what the speaker meant with their *even* sentence.[36]

Turning to the question of how children actually acquire the adult grammar of *even*, the following picture emerges: learning, under the present view, is a function of becoming more adept at recognizing conversational goals and more intolerant when those goals are not identified during comprehension. In other words, the more pragmatically skilled a learner is, the better they will be at recognizing and recording the specific conversational setting in which *even* sentences are used. This growing conversational confidence favors the adult grammar, which does not support opposite inferences, and correspondingly discriminates against a grammar that does support opposite inferences. Eventually, the absence of evidence in favor of

[33] Though it is now widely accepted that some aspects of Theory of Mind reasoning, often called "implicit" Theory of Mind inferences, are in place earlier than our age range (cf. Onishi and Baillargeon (2005)) it is plausible that the relevant skills in our task include assessing a speaker's conversational assumptions and goals, which has been argued to come online much later in development and hence are likely to scale with the mount of practice young reasoners have, see e.g. Perner and Roessler (2012).

[34] See Katsos and Bishop (2011) for a similar notion that children are pragmatically more tolerant than adults.

[35] We have evidence from the adult control studies that the adult grammar, by contrast, does not provide access to all four cells. Rather, the adult error pattern we have observed is compatible with only three of the four cells in the space (Table 1B). This potentially indicates that the adult grammar is most like the scope theory, since only one type of error - least-likely-to inferences in negative environments - occurred. Most-likely-to inferences in positive environments never occurred, which is straightforwardly predicted by the scope theory. The ambiguity theory, by contrast, rules them out by a mechanism – licensing of NPI – that is known to be error prone during processing (Drenhaus et al., 2005).

[36] If we assume that whenever they proceed in a state of ignorance they are guessing which character they should pick our observed rate of opposite responses of 25% translates into a rate 50% of not being able to figure out what exactly the speaker had in mind. To assess whether this is a reasonable estimate would, however, require a more fleshed out theory of what makes pragmatic reasoning of this sort difficult for our learner.

opposite inferences is taken by the learner to suggest an adjustment to the grammar to ensure that sentences that would generate opposite inferences are no longer generated to begin with.[37]

Turning to the question from Section 6.2, i.e. why learners do not consider other logically possible adult grammars with different combinations of likelihood inferences and polarity, two factors emerge: 1) no data from the input supports such grammars, and 2) the learner's own pragmatic knowledge discourages them.

Last but not least, the present perspective also allows us to sketch a plausible account of the production-comprehension asymmetry. Recall that we never see children use *even* in a non-adult way, even at the earliest stages. Specifically, they underutilize the opposite inferences that their grammar apparently licenses.

A key difference between production and comprehension is that the speaker knows what that intended message is, while the comprehender has to figure out what the message is that the speaker intended.[38] Our account allows us to exploit this difference directly: we proposed that opposite inferences surface during comprehension when the listener is unable to figure out what the speaker's conversational goals might be (e.g. why they singled out Jessiepillar), but is nevertheless willing to go along with the task at hand (in our case selecting one of the three characters).[39][40] Importantly, we did not require that children not appreciate the connection between likelihood and expectedness. Indeed, the fact they they mostly interpret our target sentences in an adult way suggests that they do, just not as reliably as adults. This means that we can reasonably assume that when they issue an *even* sentence they do so with the intent to convey information that rides on the intuitive connection between likelihood and expectedness.

7 CONCLUSION

This paper has advanced a view of the scalar inferences associated with *even* in child grammar that stands in contrast with much prior literature. Previous acquisition studies of *even* and scalar inferences treated children's non-adult-like behavior as evidence of simple confusion about *even* and likelihood inferences in discourse. Our studies present evidence to the contrary. Children are not simply confused. They are, in fact, rather keenly sensitive to the scalar nature of *even* and generate robustly both least-likely and most-likely inferences very early on. They are non-adult-like only in that they exhibit more tolerance to uses of *even* during comprehension where the speaker's conversational goals carried by *even* are left unresolved.

On our view, there is nothing difficult per se about detecting scalar inferences. As soon as children learn to associate them with the particle *even*, they immediately access a relevant hypothesis space of scalar inferences associated with *even* along the lines of Rooth (1985)'s ambiguity theory (Table 1A). This is evident from their "error" patterns in our comprehension studies, as well as the absence of such errors in production.

We presented two comprehension studies that shared Kim (2011)'s "Guess who?" format. In these experiments, children as young as 4yo performed well above chance. Most notably, children in our age

[37] Our proposal is part of the growing literature on the role of pragmatics in language acquisition which has uncovered a great deal of pragmatic sophistication that young learners bring to the task of language acquisition in variety of different situations, ranging from referential word learning (e.g. Horowitz and Frank, 2015; Sullivan and Barner, 2015; Sullivan et al., 2019) to speech act pragmatics and its role in the acquisition of propositional attitudes (Hacquard and Lidz, 2018). What the precise relation of our proposal is to the type of pragmatic reasoning in these cases is not immediately obvious since expunging a grammatical option that is underutilized because speakers tend to not highlight likely outcomes bears little resemblance to determining when and how a novel word is used by a speaker to refer to a novel object or the cases Hacquard and Lidz (2018)'s "pragmatic syntactic bootstrapping hypothesis" is meant to account for.

[38] See especially Hendriks (2014) for discussion of production-comprehension asymmetries in language acquisition.

[39] See Aravind (2018) for evidence that children in our age range who are otherwise quite astute at picking up the status of presupposed information as contextually entailed nevertheless prefer a fully redundant reading of an utterance over one that is informative only at pains of accommodation.

[40] The error pattern we observed in our adult comprehenders – opposite responses in negative environments – can be understood in a similar vein if we assume, as seems plausible, that our web-based task environment invites shallow processing as a form of satisficing, (Ferreira, 2003). Shallow processing will generate scalar inferences for *even* but may not require a full reconstruction of the (imagined) speaker's conversational purpose in singling out Jessiepillar.

range predominantly chose either the most- or the least-likely character to succeed in any given story, but rarely chose the middle character. They additionally justified their choices with comments that demonstrated a sensitivity to scalar properties about the characters. This behavior is predicted by the space of inferences in Table 1A, but is not expected if children are merely guessing.

Further motivation for this treatment of child non-adult-like behavior comes from our adult control studies. Adults likewise were occasionally susceptible to "opposite" likelihood inferences, and justified these choices with normal reference to the scalar properties of the characters. We argued that this wouldn't be possible unless these inferences were made in principle available by the grammar.

In addition to the comprehension experiments, we conducted two corpus studies that examined tokens of *even* in child and child-directed speech. Strikingly, neither adults nor children exhibited "opposite of adult-like" uses of *even*. We argue on the basis of this production-comprehension asymmetry that children not only hypothesize a space of inferences like that in Table 1A, but they also command some of the knowledge necessary to constrain this space (or else they wouldn't be so adult-like in production).

Learning to transform the space in Table 1A to the adult grammar amounts to making use of that knowledge in comprehension as well as production. This is a gradual process of becoming increasingly intolerant to certain inferences, as they become increasingly confident in their ability to reason about, and identify, speakers' conversational goals.

REFERENCES

Aravind, A. (2018). *Presuppositions in context*. Ph.D. thesis, MIT

Crnič, L. (2014). Non-monotonicity in NPI licensing. *Natural Language Semantics* 22, 169–217. doi:10.1007/s11050-014-9104-6

Crnič, L. (2009). *Getting* even. Ph.D. thesis, MIT

Drenhaus, H., Frisch, S., and Saddy, D. (2005). Processing negative polarity items: When negation comes through the backdoor. In *Linguistic Evidence* (Mouton de Gruyter). 145–164. doi:10.1515/9783110197549.145

Drummond, A. (2012). Ibex: A web interface for psycholinguistic experiments. Available at: https://github.com/addrummond/ibex [Accessed April 9, 2018].

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry* , 353–375

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology* 47, 164 – 203. doi:10.1016/S0010-0285(03)00005-7

Fillmore, C. J. (1965). Entailment rules in semantic theory. *POLA report* 10

Francis, N. (2018). Presupposition-denying uses of *even*. In *Proceedings of SALT 28*, eds. S. Maspong, B. Stefánsdóttir, K. Blake, and F. Davis (Ithaca, NY: CLC Publications), 161–176. doi:10.3765/salt.v28i0.4409

Gast, V. and van der Auwera, J. (2011). Scalar additive operators in the languages of Europe. *Language* 87, 2–54. doi:10.1353/lan.2011.0008

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science* 7, 457–472. doi:10.1214/ss/1177011136

Giannakidou, A. (2007). The landscape of *even*. *Natural Language and Linguistic Theory* 25, 39–81. doi:10.1007/s11049-006-9006-5

Gowda, Y. (2020). childes-coder: Utility for coding data from CHILDES corpora. Available at: https://github.com/tlonic/childes-coder [Accessed March 13, 2020].

881 Greenberg, Y. (2015). *Even*, comparative likelihood and gradability. In *Amsterdam Colloquium*, eds.
882 T. Brochhagen, F. Roelofsen, and N. Theiler. vol. 20, 147–156
883 Greenberg, Y. (2016). A novel problem for the likelihood-based semantics of *even*. *Semantics and*
884 *Pragmatics* 9, 1–28. doi:10.3765/sp.9.2
885 Greenberg, Y. (2018). A revised gradability semantics for *even*. *Natural Language Semantics* 26, 51–83.
886 doi:10.1007/s11050-017-9140-0
887 Guerzoni, E. (2004). *Even*-NPIs in yes/no questions. *Natural Language Semantics* 12, 319–343. doi:10.
888 1007/s11050-004-8739-0
889 Hacquard, V. and Lidz, J. (2018). Children's attitude problems: Bootstrapping verb meaning from syntax
890 and pragmatics. *Mind & Language* 34, 73–96. doi:10.1111/mila.12192
891 Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The
892 MCMCglmm R package. *Journal of Statistical Software* 33, 1–22. doi:10.18637/jss.v033.i02
893 Halle, M. and Marantz, A. (1993). Distributed Morphology and the pieces of inflection. In *The View from*
894 *building 20: essays in linguistics in honor of Sylvain Bromberger*, eds. K. L. Hale, S. J. Keyser, and
895 S. Bromberger (Cambridge, MA: MIT Press), no. 24 in Current studies in linguistics. 111–176
896 Hendriks, P. (2014). *Asymmetries between language production and comprehension*, vol. 42 of *Studies in*
897 *Theoretical Psycholinguistics* (Cambridge, Massachusetts: Springer). doi:10.1007/978-94-007-6901-4
898 Herburger, E. (2000). *What counts: focus and quantification* (Cambridge, Massachusetts: MIT Press).
899 doi:10.7551/mitpress/7201.001.0001
900 Horn, L. (1969). A presuppositional analysis of *only* and *even*. In *Chicago: Chicago Linguistic Society*.
901 vol. 5, 98–107
902 Horn, L. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of
903 California: Los Angeles
904 Horn, L. (1989). *A natural history of negation* (University of Chicago Press)
905 Horowitz, A. C. and Frank, M. C. (2015). Young children's developing sensitivity to discourse continuity
906 as a cue for inferring reference. *Journal of Experimental Child Psychology* 129, 84–97. doi:10.1016/j.
907 jecp.2014.08.003
908 Ito, M. (2012). Japanese-speaking children's interpretation of sentences containing the focus particle
909 *datte* 'even': Conventional implicatures, QUD, and processing limitations. *Linguistics* 50. doi:10.1515/
910 ling-2012-0004
911 Karttunen, L. and Peters, S. (1979). Conventional implicature. In *Syntax and semantics*, eds. C.-K. Oh and
912 D. A. Dineen (Academic Press), vol. 11: Presupposition. 1–56
913 Katsos, N. and Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of
914 informativeness and implicature. *Cognition* 120, 67–81. doi:10.1016/j.cognition.2011.02.015
915 Kay, P. (1990). *Even*. *Linguistics and Philosophy* 13, 59–111. doi:10.1007/BF00630517
916 Kim, S. (2011). *Focus particles at syntactic, semantic and pragmatic interfaces: The acquisition of* only
917 *and* even *in English*. Ph.D. thesis, University of Hawaii
918 Krifka, M. (1991). A compositional semantics for multiple focus constructions. In *Proceedings of SALT I*,
919 eds. S. K. Moore and A. Z. Wyner (Ithaca, NY: CLC Publications), 127–158. doi:10.3765/salt.v1i0.2492
920 Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come: Bayesian methods for data
921 analysis in the organizational sciences. *Organizational Research Methods* 15, 722–752. doi:10.1177/
922 1094428112457829
923 Lahiri, U. (1998). Focus and negative polarity in Hindi. *Natural language semantics* 6, 57–123. doi:10.
924 1023/A:1008211808250

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, vol. 1 (Psychology Press)

Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part II. *Language and Linguistics Compass* 10, 591–613. doi:10.1111/lnc3.12207

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science* 308, 255–258. doi:10.1126/science.1107621

Perner, J. and Roessler, J. (2012). From infants to children's appreciation of belief. *Trends in Cognitive Sciences* 16, 519 – 525. doi:10.1016/j.tics.2012.08.004

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* 6, 7–11

Rooth, M. (1985). *Association with focus*. Ph.D. thesis, University of Massachusetts Amherst

Rooth, M. (1996). Focus. In *The Handbook of Contemporary Semantic Theory*, ed. S. Lappin (Oxford: Blackwell Publishers). 271–297

Rullmann, H. (1997). *Even*, polarity, and scope. In *Papers in Experimental and Theoretical Linguistics*, eds. M. Gibson, G. Wiebe, and G. Libben (University of Alberta), vol. 4. 40–64

Rullmann, H. (2007). What does *even* even mean? *Ms. University of British Columbia*

[Dataset] Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2018). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System. doi:10.31234/osf.io/93mwx

Schwarz, B. (2005). Scalar additive particles in negative contexts. *Natural Language Semantics* 13, 125–168. doi:10.1007/s11050-004-2441-0

Snyder, W. (2007). *Child language: the parametric approach*. Oxford linguistics (Oxford University Press)

Sullivan, J. and Barner, D. (2015). Discourse bootstrapping: preschoolers use linguistic discourse to learn new words. *Developmental Science* 19, 63–75. doi:10.1111/desc.12289

Sullivan, J., Boucher, J., Kiefer, R. J., Williams, K., and Barner, D. (2019). Discourse coherence as a cue to reference in word learning: Evidence for discourse bootstrapping. *Cognitive Science* 43, e12702. doi:10.1111/cogs.12702

von Stechow, A. (1991). Current issues in the theory of focus. In *Semantics: An International Handbook of Contemporary Rsearch*, eds. A. von Stechow and D. Wunderlich (Walter deGruyter). 804–825

Wilkinson, K. (1996). The scope of *even*. *Natural Language Semantics* 4, 193–215. doi:10.1007/BF00372819

FIGURE CAPTIONS

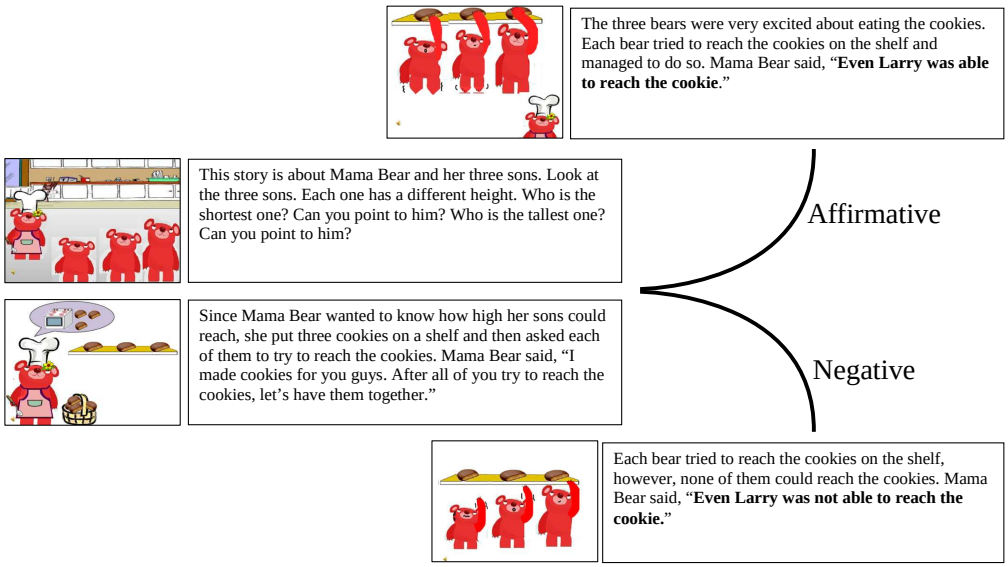
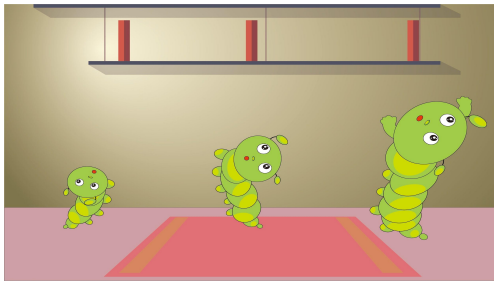


Figure 1. Sample story setup. (Kim, 2011)



"Even Jessiepillar$_F$ wasn't able to reach a book!"

"Even Rufus$_F$ was able to lift a pumpkin!"

"Even Mary$_F$ wasn't able to fit into the hole!"

"Even Sammy$_F$ was able to fill her basket!"

Figure 2. Four different scale types for the target stimuli: reach, lift, fit, fill.

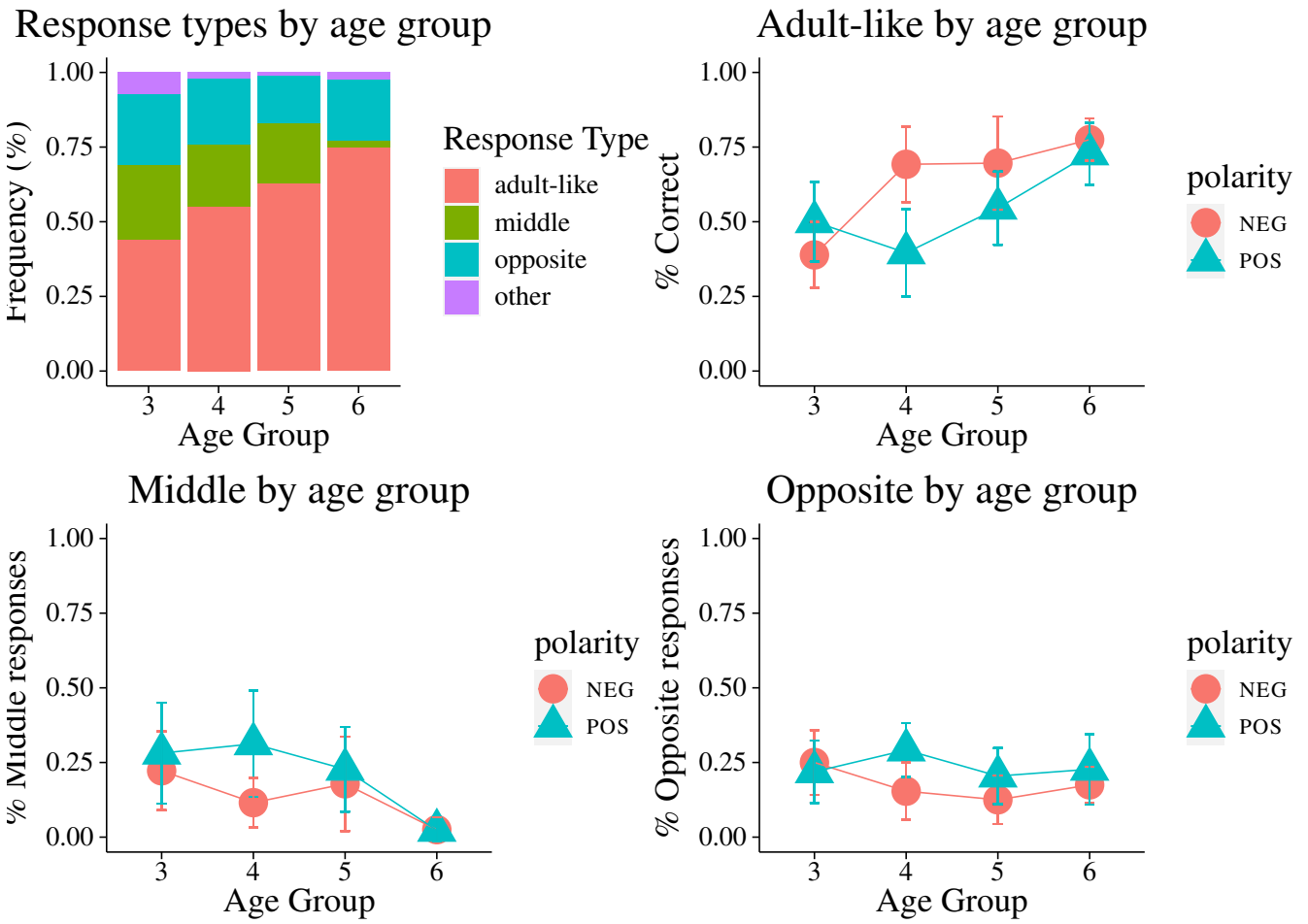
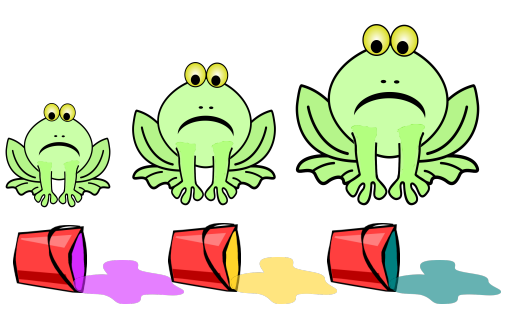
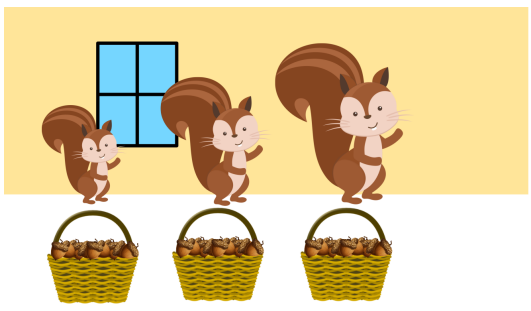


Figure 3. Results for Experiment 1. **(A)** Rate of response types by age group. **(B)** Rate of adult-like responses by age group and polarity. **(C)** Rate of middle responses by age group and polarity. **(D)** Rate of opposite responses by age group and polarity.



"Even Frida$_F$ has spilled her paint!"



"Even Sammy$_F$ hasn't spilled her basket!"

Figure 4. Spilling stimuli

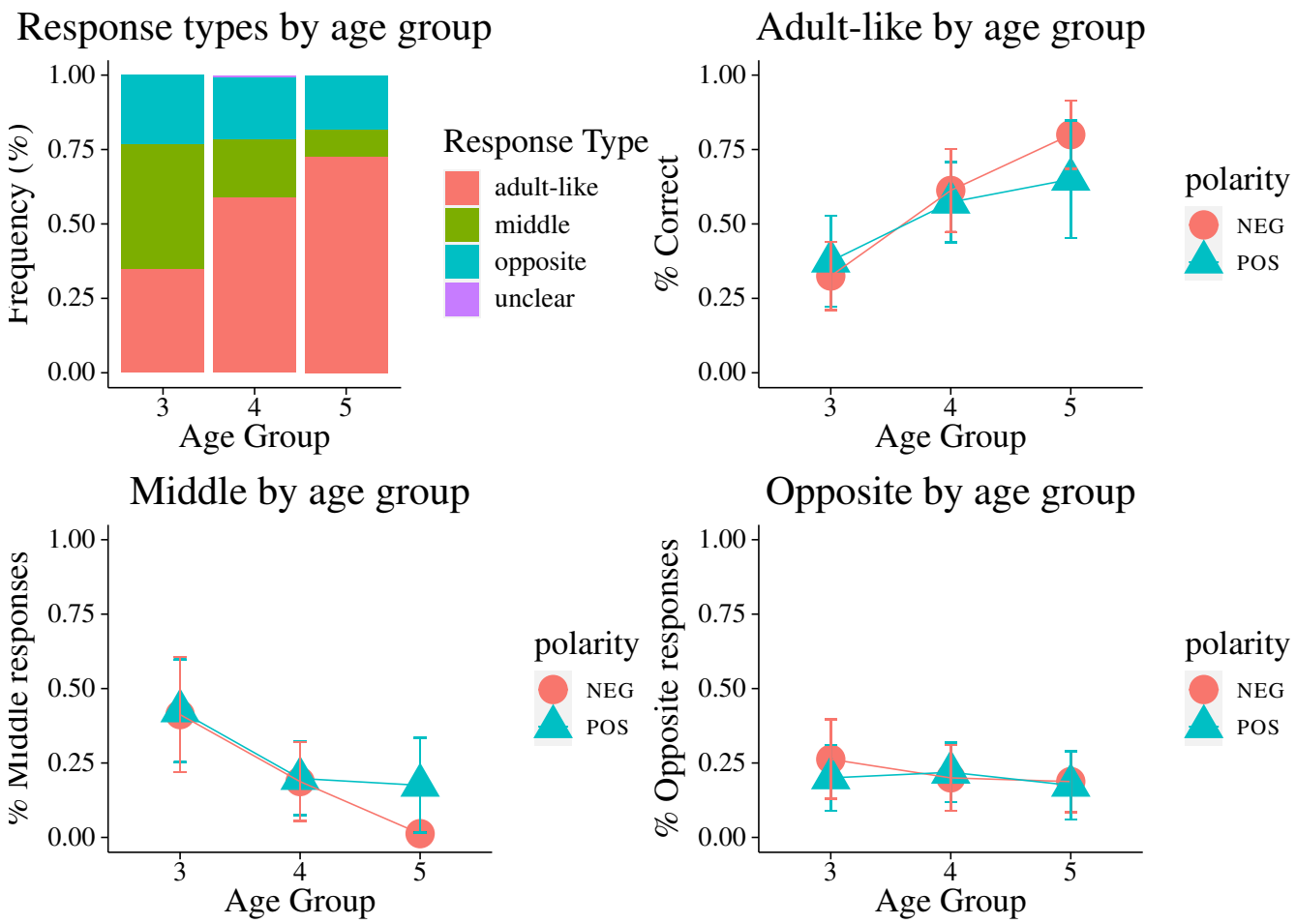


Figure 5. Results for Experiment 2. **(A)** Rate of response types by age group. **(B)** Rate of adult-like responses by age group and polarity. **(C)** Rate of middle responses by age group and polarity. **(D)** Rate of opposite responses by age group and polarity.

TABLES

		Likelihood inference	
		Least-likely	Most-likely
Sentence polarity	POS	✓	✓
	NEG	✓	✓

(A) Ambiguity theory of *even*

		Likelihood inference	
		Least-likely	Most-likely
Sentence polarity	POS	✓	
	NEG	✓	✓

(B) Scope theory of *even*

Table 1. Space of in principle available inferences associated with *even* on each theory.

		Selection pattern			
		Target characters for both sentence types	Opposite characters for both sentence types	Always rightmost or leftmost character	Any characters
Sentence type	Test sentences	33.3% (36/90)	38.9% (35/90)	27.8% (25/90) (22.2% for rightmost, 5.6% for leftmost)	
	Control sentences	20% (18/90)	40% (36/90)	36.7% (33/90) (26.7% for rightmost, 10% for leftmost)	3.3% (3/90)

Table 2. Rate of responses out of different types of pragmatics for test and control sentences in children's group. From Kim (2011).

Response Type	Justification Type	Age Group: 3	4	5	6	Total
adult-like	none	17	16	9	12	54
	random	0	4	7	6	17
	scalar	13	35	47	45	140
middle	none	15	9	5	1	30
	random	2	9	11	1	23
	scalar	0	3	4	0	7
opposite	none	10	10	3	1	24
	random	3	3	2	1	9
	scalar	3	9	11	15	38

Table 3. Experiment 1: Justifications for each response type by age.

Response Type	Polarity	% of responses	Std.dev
adult-like	POS	88%	32%
	NEG	82%	38%
middle	POS	0%	0%
	NEG	2%	14%
opposite	POS	12%	32%
	NEG	16%	37%

Table 4. Experiment 1, Adult control: Rates of response types by polarity.

Age Group vs. Age Group		**4yo** Mean	CI		**5yo** Mean	CI		**6yo** Mean	CI
3yo	Adult-like	13	-11;39	Adult-like	33	-2;67	Adult-like	37	10;66
	Middle	-6	-34;16	Middle	-12	-47;21	Middle	-17	-43;0
	Opposite	-6	-26;13	Opposite	-21	-44;-1	Opposite	-20	-44;2
4yo				Adult-like	20	-10;48	Adult-like	24	2;47
				Middle	-5	-30;22	Middle	-11	-28;0
				Opposite	-15	-33;0	Opposite	-13	-33;4
5yo							Adult-like	4	-13;29
							Middle	-6	-28;3
							Opposite	1	-9;13

(A) All data.

Age Group vs. Age Group		**4yo** Mean	CI		**5yo** Mean	CI		**6yo** Mean	CI
3yo	Adult-like	29	0;59	Adult-like	29	0;60	Adult-like	34	5;65
	Middle	-13	-48;11	Middle	-10	-47;16	Middle	-17	-50;1
	Opposite	-17	-44;7	Opposite	-19	-46;4	Opposite	-17	-45;8
4yo				Adult-like	0	-22;22	Adult-like	5	-15;27
				Middle	2	-14;22	Middle	-5	-19;3
				Opposite	-3	-20;13	Opposite	0	-18;18
5yo							Adult-like	5	-17;28
							Middle	-7	-25;2
							Opposite	3	-15;21

(B) Only data from negative stimuli.

Response Type vs. Response Type	**Adult-like**	Mean	CI	**Opposite**	Mean	CI
Middle	None	-57	-89;-19	None	-57	-90;-18
	Random	-8	-29;2	Random	-6	-28;6
	Scalar	63	26;94	Scalar	62	23;93
Opposite	None	1	-24;25			
	Random	-2	-8;2			
	Scalar	1	-24;27			

(C) Mean difference in rate of justification types by response type.

Table 5. Summary of statistical analysis for experiment 1: **(A)** Mean differences between age groups in rate of adult-like, middle, and opposite responses, for all data **(B)** Mean differences between age groups in rate of adult-like, middle, and opposite responses, for just negative stimuli. **(C)** Mean differences in rate of justification types by response type. Cells highlighted in green indicate a significant difference (i.e., not centered around 0).

Response Type	Justification Type	Age Group: 3	4	5	Total
adult-like	none	27	27	8	62
	random	6	13	4	23
	scalar	23	64	104	191
middle	none	47	15	1	63
	random	16	13	8	37
	scalar	4	6	6	16
opposite	none	18	10	6	34
	random	7	6	2	15
	scalar	12	21	21	54

Table 6. Experiment 2: Justifications offered for each response type by age.

Response Type	Polarity	% of responses	Std.dev
adult-like	POS	95%	21%
	NEG	91%	29%
middle	POS	3%	17%
	NEG	1%	11%
opposite	POS	2%	13%
	NEG	8%	27%

Table 7. Experiment 2, Adult control: Rates of response types by polarity.

Age Group vs. Age Group		4yo Mean	4yo CI		5yo Mean	5yo CI
3yo	Adult-like	39	6;71	Adult-like	56	27;86
	Middle	-37	-77;-1	Middle	-45	-80;10
	Opposite	-2	-23;20	Opposite	-12	-33;7
4yo				Adult-like	18	-2;41
				Middle	-7	-23;2
				Opposite	-10	-27;6

(A) Rates of different response types.

Response Type vs. Response Type		Adult-like Mean	Adult-like CI		Opposite Mean	Opposite CI
Middle	None	-18	-53;61	None	-25	-63;1
	Random	-13	-36;0	Random	-12	-35;2
	Scalar	31	2;61	Scalar	37	5;68
Opposite	None	7	-10;33			
	Random	-1	-6;2			
	Scalar	-6	-31;12			

(B) Rates of different justification types by response type.

Table 8. Experiment 2: Mean differences in (A): the rate of adult-like, middle, and opposite responses, and (B): the rate of justification types (none, random, and scalar) by response type. Cells highlighted in green indicate a significant difference (i.e., not centered around 0).

Age group	Sentence Polarity	Inference: Least-likely	Most-likely	Unclear	Total
3	POS	47	1	3	51
	NEG	1	17		18
4	POS	99	1	2	104
	NEG	1	161	5	167
	Unclear	1	1	1	3
5	POS	25	(1)	1	(29) 28
	NEG		42		42
	Unclear	1	1		2
6	POS	6		2	8
	NEG		23	1	24
Adults	POS	754	(57) 0	31	(842) 785
	NEG	16	1143	27	1186
	Unclear	1	2	7	10

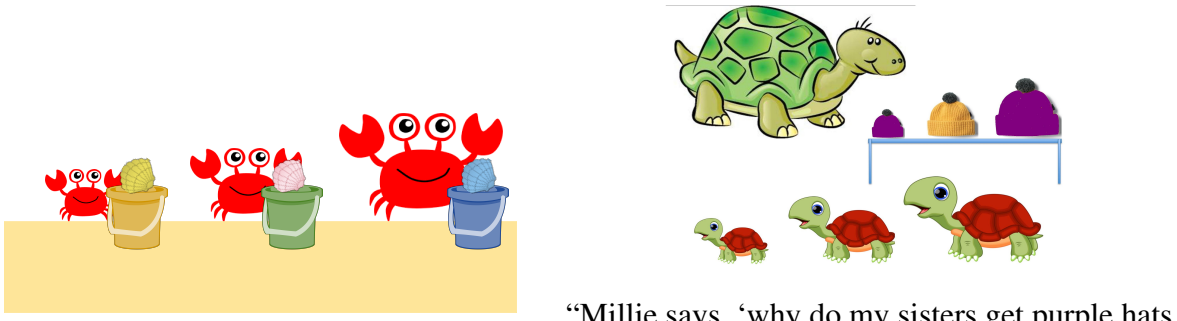
Table 9. Child and child-directed (adult) production of *even* in negative and positive environments between 3-6 years old. Figures in gray show numbers before post-hoc reanalysis (cf. footnote 27).

Age	NEG:POS *even* ratio	NEG count	POS count
3 yrs	0.35	18	51
4 yrs	1.6	167	104
5 yrs	1.5	42	28
6 yrs	3	24	8
Adult	(1.5) 1.6	(1186) 1243	785

Table 10. The ratio of negative to positive *even* sentences by age. In adults, we include downward-entailing sentences which lack sentential negation in the NEG figures, with ratios/counts excluding these sentences in gray.

Supplementary Material

0.1 Figures



"Carl's bucket doesn't match his shell!"

"Millie says, 'why do my sisters get purple hats but I get a yellow one?"

Figure S1. Two different filler items: matching by color/size.

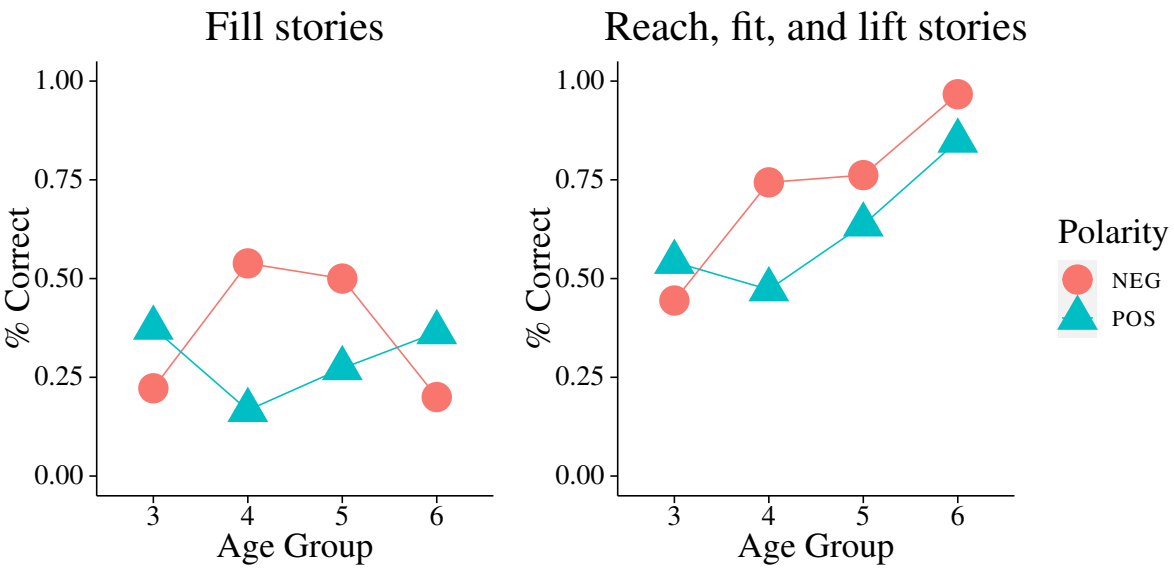


Figure S2. Adult-like behavior in Experiment 1, by story type.

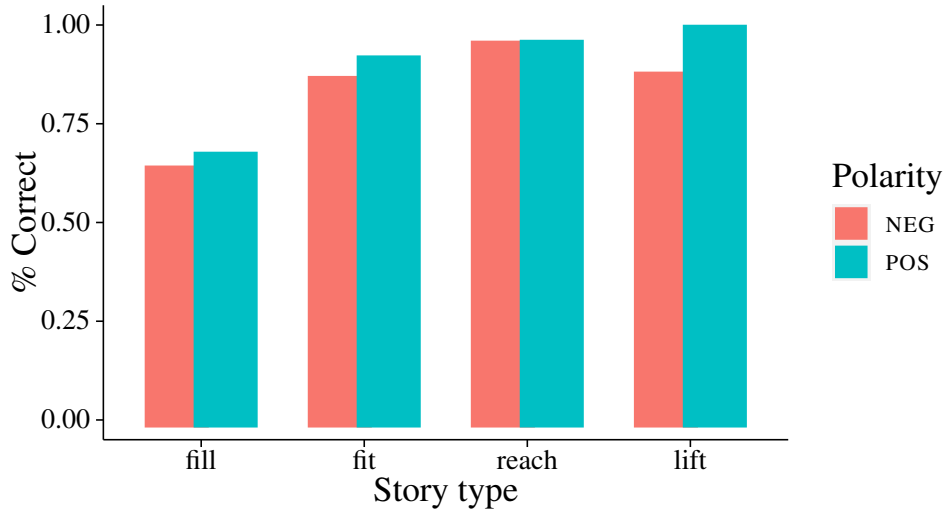


Figure S3. Adult study 1: adult performance by story type.

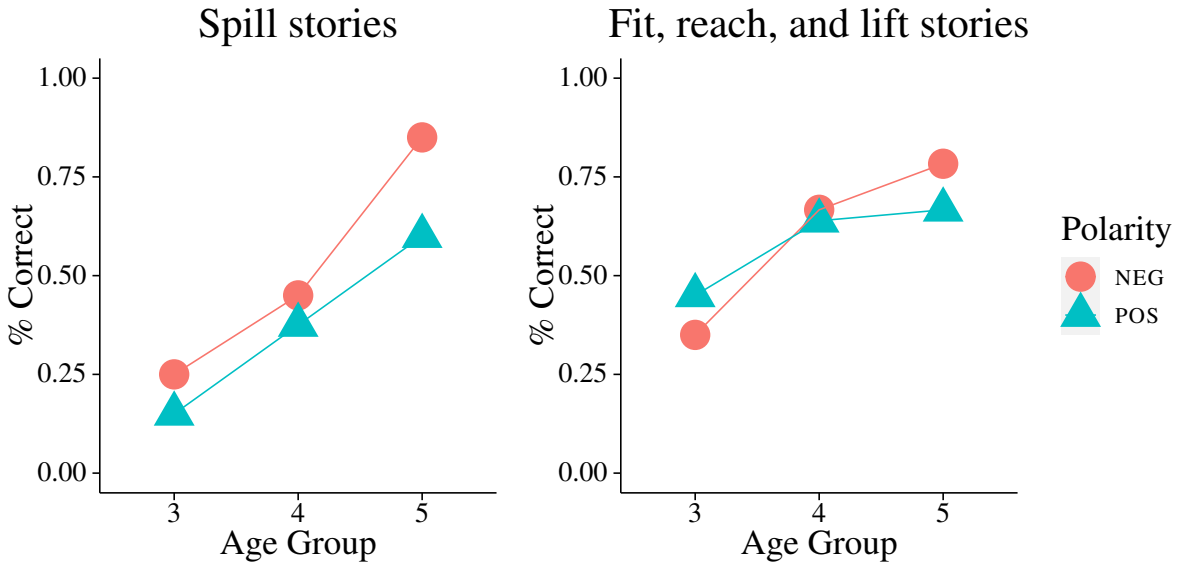


Figure S4. Adult-like behavior in Experiment 2, by story type.

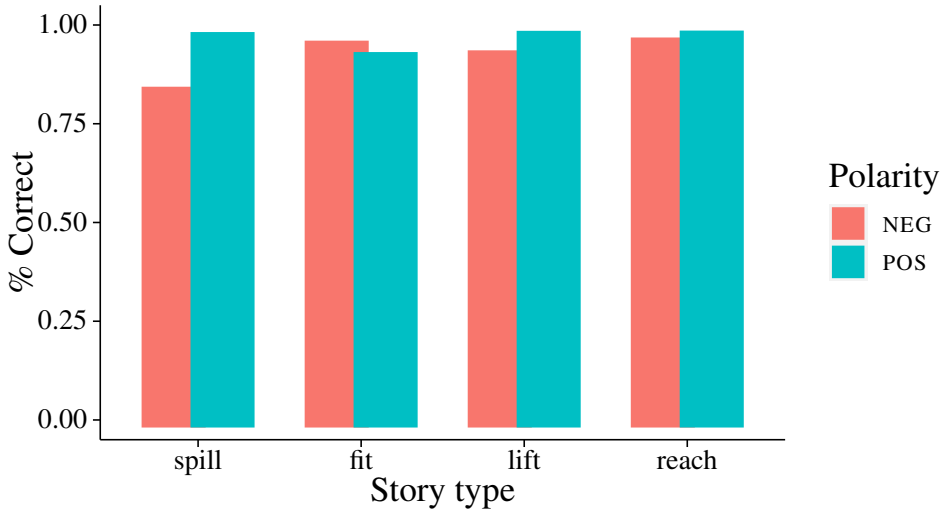


Figure S5. Adult study 2: adult performance by story type.

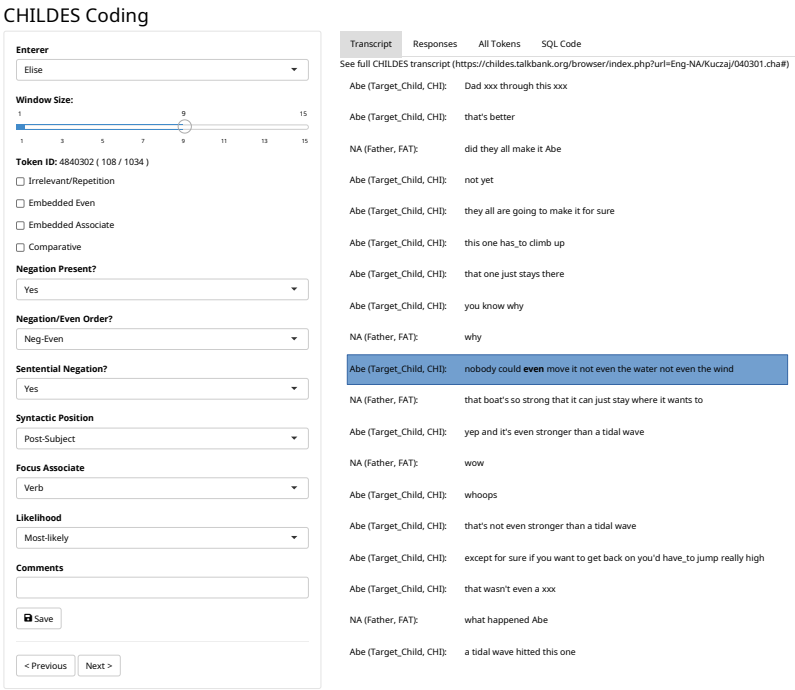


Figure S6. An example instance of *even* from the CHILDES corpus, along with its coding.

	Even's position	Sentence Polarity	Focus associate: Subject	Other	Total
Children	Pre-subject	POS	41	27	68
		NEG	15	3	18
	Post-subject	POS	13	207	220
		NEG	10	371	381
		Unclear		5	5
Adults	Pre-subject	POS	89	89	178
		NEG	24	23	47
		Unclear		2	2
	Post-subject	POS	11	637	648
		NEG	8	1123	1131
		Unclear		5	5

Table S1. Both children and adults most often associate pre-subject *even* with subject focus and post-subject *even* with VP-internal focus. Very few instances of pre-subject *even*. Highlighted cells are those which instantiate the form of our comprehension study stimuli.

Even-NEG Order	Sentential Negation	Child Total	Adult Total
NEG-*Even*	Yes	457	1156
	No	6	6
	Unclear	5	4
Even-NEG	Yes	16	23
	No	26	58
	Unclear	4	1

Table S2. In negative polarity sentences (Sentential Negation=Yes), both children and adults have a strong tendency to place *even* after the negative element. Highlighted cells are those which instantiate the form of our comprehension stimuli.