On the Nature of Alternatives in the Acquisition and Processing of *Only*

M. Hackl[a], E. Flor[a], E. Olson[b], L. Rosenstein[c], A. Sugawara[d] and Ken Wexler[a,e]

[a]Department of Linguistics and Philosophy, MIT, [b]Independent Researcher, [c]Harvard University; [d]Waseda University, [e]Departments of Brain and Cognitive Sciences, MIT

ARTICLE HISTORY
Compiled November 24, 2026

ABSTRACT
Focus particles such as *only* rely on alternatives to make their contribution to sentence meanings. Central questions remain: Are these alternatives syntactic or semantic objects? What are the mechanisms or principles responsible for determining the set of alternatives, and how sensitive are they to syntactic, semantic and pragmatic factors? Following the pioneering work of Crain et al. (1994) and much subsequent work, this paper brings data from language acquisition and language processing to bear on these questions. We argue, based on a series of acquisition and processing experiments, that the alternatives relevant for *only* are A) **syntactic objects** (following Fox and Katzir 2011) rather than purely semantic objects (Rooth 1985, 1992), and B) by default subject to a semantic orderability requirement which surfaces as **scalar presupposition**. We account for our data by developing a general comprehension model for sentences with *only* which predicts when they are relatively easy or difficult to parse. Importantly, we show that the predictions of this model are borne out both in L1-learners and adults, albeit with different surface manifestations.

KEYWORDS
First Language Acquisition, Language Processing, *Only*, Alternatives, Focus

1. Introduction

1.1. Crain's Puzzle

In a series of influential studies, Crain and colleagues (Crain et al. 1992, 1994) discovered a curious asymmetry in children's ability to interpret sentences with *only* such as those in (1). Specifically, children between ages three and six were shown to have a pronounced tendency to interpret sentences such as (1-a), where *only* seems to be attached to the subject of the sentence (henceforth "S-Only"), in a non-adult-like way, while sentences like (1-b), where *only* seems to be attached to the VP (henceforth "VP-Only"), always received an adult-like interpretation.

Different parts of the material in this paper have been presented at various venues. We thank the audiences at these venues, which include Beijing Foreign Studies University, BUCLD 39, Bunkyo Gakuin University, ESSL, KEIO University, KSLI 15, LAWNE, LFRG, LLCC and HUJI, McMaster University, Mie Univeristy, Nanzan University, NYU, Sinn und Bedeutung 19, UMass Amherst, SNU, and ZAS for their questions and comments. Data availability: The data that support the findings of this study together with the analyses scripts are available at https://github.mit.edu/ESSL/only-hackl-et-al-2026.

CONTACT M. Hackl. Email: hackl@mit.edu

(1) a. Only the cat is holding a flag. S-Only
b. The cat is only holding a flag. VP-Only

Crain et al. used a canonical truth-value judgment task to determine children's understanding of sentences like (1). Participants were told a story, supported by a picture, in which a cat is holding a flag, a goose is holding a flag and a balloon, and a frog is holding a balloon. Their task was to judge whether Kermit the Frog was correct in describing the scene using either (1-a) or (1-b), having been prompted by the proctor with *Kermit, can you tell me what happened in the story?*.

Crain et al. found that 35 out of 38 children ages three to six accepted (1-a) as a correct description of the scene even though two characters were holding a flag. Moreover, when asked to justify their response, 21 of them pointed out that the cat was holding a flag and nothing else.

These justifications reveal two critical pieces of information about this phenomenon. First, they indicate that the participants had accessed the core meaning of *only* as an exclusive operator even when giving non-adult-like responses. Second, they seemed to assign (1-a) a meaning comparable to what adults would assign to (1-b). These observations, together with the fact that children do not seem to exhibit non-adult-like interpretations for (1-b), suggest the characterization in (2).

(2) **Crain's Puzzle**

a. Children, ages three to six, exhibit a pronounced difficulty understanding S-Only sentences but not VP-Only sentences.
b. They resolve their comprehension difficulty by mapping an S-Only sentence to the meaning they assign to its VP-Only counterpart.

Crain's puzzle exhibits three broad robustness signatures. First, it is observed across both truth-values. (1) demonstrates a case where the adult interpretation of the S-Only sentence was false while the VP-Only interpretation was true. Follow-up research has revealed that the same non-adult interpretation is given to S-Only sentences when they are adult-true and their VP-Only counterpart is false (Crain et al. 1994; Notley et al. 2009; Hackl et al. 2015). Second, Crain's puzzle has been replicated across a wide developmental range, from children as young as 2;6 to children older than 6 (Notley et al. 2009). Third, the phenomenon has been documented across a substantial variety of typologically diverse languages, including English (Philip and Lynch 2000; Notley et al. 2009), Mandarin Chinese (Yang 2002; Notley et al. 2009; Zhou and Crain 2009; Su 2024), German (Müller et al. 2011; Höhle et al. 2016), and Japanese (Endo 2004; Sano 2011). The robustness of Crain's puzzle raises the stakes for understanding its true nature. Its cross-linguistic robustness, in particular, strongly suggests that Crain's puzzle is rooted in a central component of the linguistic system which might be shared across the world's languages.

To account for (2) Crain et al. (1994) suggest that the VP-Only interpretation is somehow more basic and therefore acquired earlier.[1] This means that children initially have only one interpretation for sentences with *only*. Given this, Crain's puzzle can be

[1]For the case at hand, Crain et al. (1994) intend this interpretation to be loosely paraphrased by *The cat is holding a flag and nothing else*. Adopting the canonical terminology, it's the interpretation that results from associating VP-Only with the object *a flag*. Crain et al. do not claim that this is the only way VP-Only can be interpreted by children. VP-Only could also associate with the entire VP, for instance, giving rise to an interpretation paraphrased by *The cat is holding a flag and satisfies no other relevant property*. In other words, their hypothesis is that children have access to a limited set of meanings for sentences with *only* and that set corresponds to the range of interpretations adults have for VP-Only.

explained by assuming that children can map any *only*-sentence to that interpretation, ignoring the surface position of *only*. Concretely, S-Only sentences are predicted to receive a VP-Only interpretation whereas the reverse error – misinterpreting VP-Only sentences with the S-Only meaning – is predicted never to occur.

While this proposal accounts for the asymmetry in (2), it comes with several non-trivial implications and questions. One important implication is that the underlying source is a genuinely immature, non-adult-like grammar, which lacks a specific form-meaning mapping. A consequence of this assumption is that the asymmetry should be evident in language production as well as language comprehension. The present paper argues instead that Crain's puzzle is fundamentally a comprehension phenomenon and that explicating the child behavior requires a comprehension model for sentences with *only* which takes into account how various factors affecting comprehension difficulty interact with their grammar.

1.2. Reassessment of Crain's Puzzle

Crain's puzzle has garnered a lot of interest over the years leading to many replications across languages, more nuanced assessments of child behavior, and restatements and refinements of Crain et al.'s original proposal. Within the space allotted here, we cannot do full justice to the extant literature. Instead, we summarize the results of Hackl et al. (2015), which lays the foundation for the present paper, referencing the work Hackl et al. (2015) builds on when appropriate.

Hackl et al. (2015) expanded on a finding by Notley et al. (2009) indicating that under certain contextual manipulations children at an age where they would otherwise exhibit Crain's puzzle can interpret S-Only in an adult-like manner. Most prominent among these manipulations was a change of the leading question from *What happened?* to a subject question like *Who is holding a flag?*[2] Hackl et al. (2015) followed up on this finding and conducted a more systematic investigation of the role that the preceding question (the question-under-discussion, or QUD) plays in children's understanding of *only*.

The starting point of their study is the observation that Crain et al.'s original setup involves an incongruent pairing of question and answer, (Roberts 2012; Rooth 1996). This, they argue, imposes an additional processing load on comprehenders because they are forced to accommodate a congruent sub-question of *What happened?* suitable for the *only* sentence at hand: For S-Only it is a subject question and for VP-Only it is an object question, as is indicated in (3) and (4).

(3) a. Who is holding a flag?
b. Only the cat is holding a flag.
c. *The cat is only holding a flag.

(4) a. What is the cat holding?
b. *Only the cat is holding a flag.
c. The cat is only holding a flag.

[2]To account for this, Notley et al. (2009) offer the following proposal: *Only* is initially an adverbial quantifier which takes scope over the entire sentence when it immediately precedes the subject and so can, in principle, associate with any expression in the sentence. By default it associates with the object, giving rise to Crain's puzzle. However, this default can be overridden e.g. when the sentence with *only* is preceded by a subject question. Later in development, having encountered positive evidence e.g. in the form of term answers, children reanalyze S-Only as involving a quantificational noun phrase ("DP-Only").

The congruence requirement for answers with *only* and their corresponding questions exemplified in (3) and (4) can be subsumed under the following descriptive generalizations, (Rooth 1996; Roberts 2012, a.o.).

(5) **Question-Answer Congruence**
A declarative sentence S is a **congruent** answer to a single *wh*-question Q_{wh} if the answer term in S corresponding to the *wh*-phrase in Q_{wh} is F(ocus)-marked.

(6) **Focus Co-occurrence Constraint** for *Only* (**"FCC-*Only*"**)[3]
A sentence S containing an occurrence of *only* has to have an **accessible F-feature** identifying a constituent, X_F, as the associate of *only* .

a. When *only* occurs to the immediate left of the subject (S-Only) F is accessible if it is on or inside the subject.
b. When *only* occurs to the immediate left of the VP (VP-Only) F is accessible if it is on or inside the VP.

Principles (5) and (6) jointly predict the acceptability pattern between subject and object questions, on the one hand, and S-Only vs. VP-only answers, on the other: a congruent question places F on a constituent accessible to *only*, (3-a)⊕(3-b) and (4-a)⊕(4-c), while an incongruent question demands F in an inaccessible position, (3-a)⊕(3-c) and (4-a)⊕(4-b).

Interestingly, the discourse in Crain et al.'s original setup where *What happened?* precedes the *only*-answer appears to be neither fully acceptable nor fully unacceptable.

(7) a. What happened in the story?
b. #Only the cat is holding a flag.
c. #The cat is only holding a flag.

Here neither answer is strictly congruent and, indeed, notably less well matched to the preceding question than the congruent pairs from above. Nevertheless, unlike the mismatches in (3)–(4), they seem to form a grammatical discourse with *What happened in the story?*. Following Roberts (2012) we can attribute this to the fact that comprehenders can accommodate a congruent sub-question of the original QUD (e.g., *Who is holding a flag?* for (7-b), *What is the cat holding?* for (7-c)). This is possible, on Roberts (2012)'s account, because the narrower questions seek subsets of the information requested by *What happened?* and question accommodation to repair a mismatched dialogue is, by assumption, constrained to sub-questions. No comparable repair is available for the mismatched pairs in (3)–(4), where the explicit question is, informally speaking, seeking information at the same level of granularity as its congruent counterpart and a move to that question would be lateral.

This perspective allows us to reframe Crain's puzzle: the asymmetry between S-Only and VP-Only seen in Crain et al. (1992) can be attributed to the relative difficulty of accommodating a subject question versus an object question, rather than to a non-adult-like grammar. A prediction of this shift in perspective is that Crain's puzzle should disappear with congruent questions, and the reverse mis-analysis should emerge when VP-Only is preceded by a subject question. Hackl et al. (2015) confirmed both

[3]This statement is evidently tailored towards the cases discussed here and not meant to cover all occurrences of *only*. Moreover, it is purposefully descriptive and avoids any reference to syntactic structure even though many theories of *only* attempt to explicate (6) as a consequence of *only* requiring its associate to be c-commanded by *only* (at LF).

predictions, as summarized in Table 1.[4]

Table 1. Rates of adult-like interpretations in Hackl et al. (2015) by question type.

	Question Type		
Position of *only*	What happened?	Who	What
S-Only	29.6%	72%	15.5%
VP-Only	79.6%	50%	93%

As is evident from Table 1, both congruent conditions yield high rates of adult-like interpretations, both incongruent conditions yield low rates, and the accommodation-requiring baseline (*What happened?*) produces the original Crain asymmetry. Crucially, the data show that Crain et al.'s characterization both under- and overestimates children's abilities. [5]

i. Crain et al. underestimate children's ability because children can interpret S-Only sentences in an adult-like way, and do so reliably in supportive discourse environments that do not require question accommodation.
ii. Crain et al. overestimate children's ability because children also sometimes misinterpret VP-Only sentences to mean what their S-Only counterparts would. In fact, they do so at substantial rates when the discourse environment suggests an S-Only meaning via a preceding subject question.

These observations suggest that Crain's puzzle is fundamentally a comprehension phenomenon which may be exhibited by children even if they have an adult-like grammar for S-Only and VP-Only sentences. Understanding why and when it arises would, then, require a general comprehension model for sentences with *only* that explains when they are easy or difficult to understand. We will propose such a model in Section 3 based on the idea that the search for an appropriate associate of *only* during comprehension is guided by potentially conflicting factors, including but not limited to the preceding question. Before we get to this, however, we discuss another expectation coming out of this change in perspective, namely that Crain's puzzle may not be evident in child language production.

2. Corpus study: *Only* in child and child-directed adult speech

Viewing Crain's puzzle as a comprehension phenomenon rooted in factors governing the search for the associate of only rather than as a direct reflection of a non-adult-like grammar leads to the expectation that utterances of S-Only sentences conveying a VP-Only meaning (or, for that matter, VP-Only sentences conveying an S-Only meaning) do not occur in child production. Simply put, this is because speakers are assumed to know the message they want to convey and so know what the associate of *only* is before they utter the sentence. This means that child productions of *only* sentences should be essentially adult-like.

[4]Table 1 presents the aggregate results of two versions of the same experiment. See Hackl et al. for details. The *What happened?* data are reanalyzed as Child Experiment 1 below, where applying the same exclusion criteria to all three child experiments leaves 46 rather than 38 children in the analysis and yields slightly different rates (29.9% and 79.9%).

[5]Su (2024) finds similar effects for S-Only with implicit subject or object questions induced by context manipulations.

The predictions of proposals like Crain et al. (1994) or Notley et al. (2009), who analyze Crain's puzzle as a reflection of a non-adult-like grammar, are quite different. They expect that S-Only occurs in child production at roughly the same rate as VP-Only, but with a non-adult-like, VP-Only interpretation unless there is sufficient contextual support for the S-Only interpretation. They also predict that adult-like uses of DP-only, irrespective of whether it is the subject or any other part of the sentence, do not occur until late in development, after DP-Only has been recognized as a construction of the language.

To evaluate these predictions, we examined child and child-directed adult speech in the Brown and Providence corpora from CHILDES (Brown 1973; Rose and MacWhinney 2014). We extracted all utterances containing *only* along with their conversational context using the childes-db database version 2021.1 (Sanchez et al. 2019): all nine children in the two corpora produced sentences with *only* (for a total of 177 observations). We coded each occurrence of *only* by hand for its syntactic position, the identity of its associate in the utterance, the semantic properties of its associate (whether it was in some sense scalar, a conjunction, or neither), and the licensing properties of the surrounding discourse. This is summarized in the scoring scheme in Table 2.[6]

Table 2. Scoring scheme for utterances with *only* in CHILDES

POSITION	S	VP		Other	$*$
ASSOCIATE	S	VP	V	Other	$*$
SEMANTIC	Scalar	Conjunction	Neither		$*$
DISCOURSE	Question	Focus	Neither		$*$

We established 3 possible values for the category POSITION, distinguishing between *only* found immediately preceding the subject ("S"), immediately preceding the VP ("VP"), or attached to some other category ("other") such as object DPs or adjuncts. We use the category ASSOCIATE to record which expression in the utterance serves as the associate of *only* given the intended meaning. The possible values included all the values for POSITION as well as a value of "V" to mark narrow focus on the main verb.

Each associate was classified in terms of SEMANTIC properties potentially relevant for *only*. To keep track of so- called "scalar *only*", we classified associates as "scalar" if they had meanings that were either intrinsically scalar such as numerals or gradable predicates (*one, little*, etc.) or scalar given the context or world knowledge (e.g. *only a baby, not a grown up*). Furthermore, we flagged associates if they were conjunctions taking them to be proxies for so-called "quantity *only*".[7] Finally, we scored the surrounding discourse in terms of its licensing potential for the *only* utterance. We distinguish two types of supportive contexts, a preceding question or a contrastive focus.

For each of these four categories we used the value $*$ to mark that unequivocal scoring was impossible and we marked repetitions ("r") and excluded both from any analyses. We added two additional flags to this general scoring schema to identify cases of so called "adversative only" and cases of DP-internal only. Such utterances were marked with "a" and "d" for both POSITION and ASSOCIATE respectively: these are treated as special cases for the purposes of our investigation and are therefore excluded from the main analysis.

Before turning to the results we describe the sample that the counts below are drawn

[6]See https://github.mit.edu/ESSL/only-hackl-et-al-2026 for raw scoring data.

[7]See Section 3 for theoretical justifications.

from. Brown contributes three children, aged 18.0 to 62.4 months across the recordings (mean 42.2), and Providence six, aged 11.1 to 48.1 months (mean 26.9). Table 3 gives the size of each sub-sample together with the density of *only* in it. Two properties of the sample bear on what it can and cannot show. *Only* occurs 2.37 times per 10,000 word tokens in child speech and 4.15 times per 10,000 word tokens in child-directed speech. Children produce *only* at about 0.57 times the rate at which they hear it when the comparison is made per word token, and at about 0.35 times that rate when it is made per utterance. The two figures differ because child utterances are shorter than the utterances addressed to children (2.82 against 4.61 MLU), and we give both so that the comparison does not turn on the choice of denominator. The 896 tokens of *only* in Table 3 are distributed over the 881 utterances discussed below, fifteen utterances containing two.

Table 3. Size of each sub-sample and density of *only* in it

Corpus	Role	Children	Utterances	Word tokens	*only*	per 10k utt.	per 10k tokens
Brown	Child	3	96,952	305,540	122	12.58	3.99
Brown	Adult	3	87,687	366,889	218	24.86	5.94
Providence	Child	6	173,286	457,085	59	3.40	1.29
Providence	Adult	6	286,775	1,357,528	497	17.33	3.66
Both	Child	9	270,238	762,625	181	6.70	2.37
Both	Adult	9	374,462	1,724,417	715	19.09	4.15

Out of a total of 881 utterances containing *only* from the two corpora, 177 are child productions and 704 are child directed adult speech (henceforth, for brevity, "Child" and "Adult"). We excluded 36 from Child and 41 from Adult because it was impossible to determine either the syntactic position of *only* or its associate or both. We also excluded 19 occurrences from Child and 35 from Adult because they were verbatim or near-verbatim repetitions of something that had just been uttered (either by the same speaker, or someone else). These two exclusion criteria resulted in a 31% and 11% loss from Child and Adult, respectively. After excluding DP-internal and adversative *only* because they are not directly relevant to the core predictions we are left with 113 occurrences of *only* in Child and 501 in Adult.

Table 4 presents a summary of *only* occurrences by position, aggregating the data from the 9 children from the Brown and Providence corpora and performing the exclusions described above. Figure 1 displays the development in the distribution of the three different types of *only* in Child (bottom panel) and Adult (upper panel) from 15 months of age to 62 months of age, binned into 6-month intervals.[8]

Table 4. Number of *only* occurrences by Position and Associate

	S-Only	VP-Only	Other-Only	Total post excl.	S-Only/(VP+Other)
Adult	44	252	205	501	9.6%
Child	15	50	48	113	15.3%
total	59	302	253	614	10.6%

Even though the overall number of occurrences of *only* is relatively low, we can identify distributional properties which speak to our questions. First, we observe that the number of S-Only cases is relatively small in both Adult and Child, amounting

[8]Missing from Figure 1 are 3 occurrences of VP-Only and 3 of Other found in Adult. They were removed from the chart because they occurred before the first age bin (target child's age less than 15 months), when no relevant utterance of *only* occurred in Child.

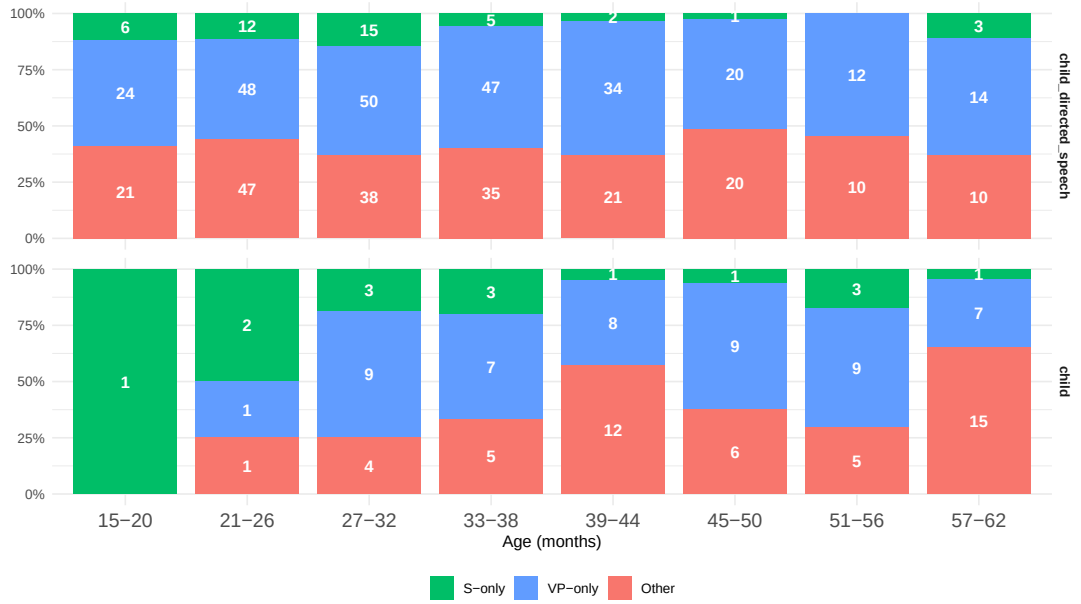

Figure 1. Distribution of S-Only, VP-Only and Other over time

to barely a tenth of the occurrences overall.[9] The majority of cases are VP-Only or Other-Only cases where *only* is attached to the object or an adjunct. Examples of the latter class from CHILD are in (8).

(8) a. see mines only has one (Adam 4;4; Brown, 040401.cha:2820)
b. dey need only one (Adam 3;5; Brown, 030515.cha:3918)
c. but only if you get dat one out (Sarah 5;0; Brown, 050016.cha:175)

Interestingly, the ratio of S-Only is actually higher in CHILD than in ADULT (15.3% of cases vs. 9.6%). Moreover, as can be seen in the stacked bar plot, the rate of S-Only is even higher early in development (15-38 months) before it levels out at the adult rate (39-62 months). This is quite striking, especially from the vantage point of Crain's puzzle. A relevant example is shown in (9).

(9) only Mommy use it. (Eve 1;10; Brown, 011000b.cha:1187)

Even more striking is the fact that we found no occurrences that might be instances of Crain's puzzle in either CHILD or ADULT and only one case in ADULT which might be the reverse error.[10] In fact, more generally, we find that *only* is deployed by both types of speakers with a remarkable degree of adherence to the grammatical and discourse principles regulating its use. Mis-speaking, as it were, is next to non-existent in our corpora even though both types of speakers exhibit robust uses of *only* in a variety of different syntactic environments including not only VP-Only and S-Only but also

[9]We cannot say at this point whether the low frequency of S-Only is a specific property of child and child directed speech.

[10]The only case that at first glance appears to be an instance of illegitimate association comes from Adam's mother who says *the firemen only can do that* (Brown, 030418.cha:560). The associate of *only* here is clearly the subject. However, *only* is plausibly right adjoined to the subject here and so possibly not a genuine case of illegitimate association. Nevertheless, to be conservative, it is counted as instance of VP-Only in our analysis.

DP-Only, (8-b), "adversative" *only*, (10-a), and DP-internal *only*, (10-b).

(10) a. only I don't want to make them (William 3;1; Providence, 030125.cha:1552)
b. yyy this is the only new one I've got here (Naima 2;10; Providence, 021001.cha:2558)

Table 5 summarizes the results of the corpus study as it pertains to the SEMANTIC and DISCOURSE properties of *only*. One notable fact is that more than half of the cases have a scalar associate (59.7% in ADULT and 53% in CHILD). This includes inherently scalar associates like bare numerals or gradable adjectives but also associates that are scalar in virtue of contextual information or world knowledge. Examples from CHILD are in (11).

(11) a. you only need one quart (Adam 3;5; Brown, 030515.cha:2514)
b. I'm only little (Ethan 2;8; Providence, 020810.cha:570)
c. he's only a baby (Sarah 3;6; Brown, 030630.cha:733)

With regard to discourse properties we see that only about 30% of cases had an identifiable licensor in the surrounding discourse (a question or a focused expression) and that almost all of those cases involved a contrastively focused antecedent (27.5% in ADULT and 26.5% in CHILD). Examples are in (12).

(12) a. You cannot use it — only $[\text{Mommy}]_F$ use it (Eve 1;10; Brown, 011000b.cha:1187)
b. Oh and what else? — only a wig (Sarah 4;4; Brown, 040425.cha:246)

We also see an interesting interaction between the scalarity of the associate and discourse licensing via focus: the rate of focus licensing is noticeably higher for non-scalar associates. In the case of ADULT, 25.4% (76/299) of all scalar associate cases are focus licensed while 32.4% (60/185) of non-scalar ones are. In the case of CHILD the difference is even greater (18.3% (11/60) versus 40.9% (18/44)). In other words, associates that are scalar appear to be less in need of additional discourse licensing via focus.[11]

Table 5. Results for SEMANTIC and DISCOURSE

ADULT	Focus	Qu.	Neither	∗	Tot.
Scalar	76	1	217	5	299
Conj.	1	0	6	0	7
Neither	60	1	116	8	185
∗	1	0	6	3	10
Tot.	138	2	345	16	501

CHILD	Focus	Qu.	Neither	∗	Tot.
Scalar	11	5	43	1	60
Conj.	0	0	1	0	1
Neither	18	4	21	1	44
∗	1	0	2	5	8
Tot.	30	9	67	7	113

[a] "∗" here indicates that unequivocal scoring according to the given category was impossible; "Neither" indicates that that the occurrence was scored has bearing neither (semantic or discourse) property. The data reported here are post exclusions.

In sum, our corpus study provides strong indications that children's productions are adult-like when it comes to *only*, and they are so from the earliest points of use onward. Children seem to know the meaning of *only* in exquisite detail—they use it to express propositions we have labeled S-Only, VP-Only, Object-Only, "adversative only", "scalar only", they use it discourse-appropriately, and they do not misspeak in the way we might expect if Crain's puzzle were a direct product of a non-adult-like grammar lacking certain form-meaning pairings. In particular and most relevantly to our discussion, neither Crain's error nor the "reverse Crain" error which Hackl et al.

[11] As we will see, below, our comprehension model offers a way of understanding these observations in terms of the scalar presupposition triggered by *only* and the role that information provided by the discourse context plays.

(2015) observed with incongruent VP-Only sentences is evident in their productions. Importantly, as we have seen, this is not for lack of opportunities as children use *only* in various syntactic positions, including immediately preceding the subject, the VP, the object, and various adjuncts, and even sentence-finally.

Given what appears to be a relatively small number of S-Only cases in the corpus, we might wonder whether the absence of Crain-style misuse is statistically meaningful.[12] Recall that there are 15 instances of S-Only, all of which are adult-like in having the subject as the associate and 50 instances of VP-Only, all of which are adult-like in having the object or the VP as associate. I.e., none of our cases are an instance of the puzzling ungrammatical errors found in comprehension studies.

To statistically assess the meaningfulness of these observations we can ask what the probability is that these 15 adult-like S-Only examples and 50 adult-like VP-Only cases could have arisen by chance. On the assumption that there is no underlying difference in the relevant child grammars with regard to comprehension and production—an immature grammar will affect comprehension and production in the same way as will an adult-like grammar—the null hypothesis can be derived by estimating the base rate for making the respective speech error in both cases from our comprehension data.[13]

Using the mean for correct comprehension from Experiment 1 in section 5.1 of Hackl et al. (2015), which is a direct replication of Crain et al. (1992), the probability for S-Only to be adult-like is 29.9% and for VP-Only 79.9%. For a set of 15 occurrences of S-Only, which can reasonably be assumed to be independent (given that utterances of *only* deemed to be "repetitions" were excluded), the probability of all exhibiting an adult-like associate is $.299^{15}$ (approximately $1.3648e-8$), obviously an extremely unlikely outcome far below customary confidence thresholds for concluding that an outcome is consistent with the null hypothesis. The same is true for VP-Only ($.799^{50} = 1.3407e-5$). This means that there is almost no chance that the associates in the production data were chosen by chance if there was no difference in the grammars for comprehension and production.[14]

It may be worth noting that the age range for the children in this replication study (3;3–6;9, M=4;10) differed notably from that in the corpus data, which included productions from children as young as 1;7. Moreover, 9 of the 15 S-Only cases in the corpus occurred by 36 months of age. Yet the youngest child in the comprehension experiment was 39 months. So 3/5 of the corpus S-Only productions were from even younger children than the youngest child in the comprehension experiment. However, this difference in age range does not undermine our argument. If anything, it adds further support for our conclusion because younger children are expected to make more errors of association than older children if grammatical knowledge of *only* is still developing in the age range where we observe Crain's puzzle. In other words, comparing the S/VP-Only productions of younger children to the errors made by the older children in the comprehension experiment would be intrinsically skewed in favor of finding higher rates in the corpus study if there is grammatical development.

[12] We thank Reviewer 2 for urging us to ask and respond to this question.

[13] The underlying assumption here is that speech errors, which in general are part of adult as well as child speech, may occur at different rates even when they are quite similar in form and they may do so for grammar independent reasons.

[14] See e.g. Poeppel and Wexler (1993) for a parallel use of this type of exact probability test to assess significance of occurrence rates in a corpus.

3. A comprehension model for *only S*

The corpus data together with the findings of Hackl et al. (2015) and related work (Notley et al. 2009) suggest that Crain's puzzle is fundamentally a comprehension phenomenon. To gain traction on how it arises we propose in this section a comprehension model for sentences with *only* that allows us to explicate how various factors affecting comprehension interact to produce the observed data.

The core idea of our model is that the realtime comprehension ease or difficulty of a sentence *Only S* is a function of how easy or difficult it is to locate the focus-marked constituent, X_F, within S serving as its associate. Identifying X_F is critical because it provides the key to generating the set of alternatives, $ALT(S)$, that *only* negates. We propose that multiple, potentially conflicting factors govern the search for X_F during online comprehension. These factors include (i) the anaphoric link between $ALT(S)$ and the current Question Under Discussion (QUD), (ii) the scalar presupposition triggered by *only* which requires the elements of $ALT(S)$ to be orderable, (iii) the algorithm for generating $ALT(S)$ from X_F which constrains which input constituents can give rise to an orderable $ALT(S)$, and (iv) the syntactic position of *only* in S . When these factors converge – pointing to the same constituent as the associate of *only* – comprehension proceeds smoothly. When they conflict – for example, when the constituent that most readily supports the scalar presupposition of *only* is in a position where F is inaccessible according to (6) – comprehension is difficult.

3.1. Terms of Discussion

To make these ideas precise, we follow Fox and Katzir (2011) in analyzing focus association with *only*, assuming a Rooth-style LF where *only* is a sentence-level operator no matter where it appears on the surface (Rooth 1996). *Only* combines with its clausal sister S, the so-called prejacent of *only*, and uses it in two ways. First, following Horn (1969), it conveys that the proposition expressed by S must be entailed by the common ground, i.e. it is presupposed.[15] We indicate the presuppositional status of that inference by underlining it in (13). Second, it asserts that all relevant alternatives to S collected in $ALT(S)$ express propositions that are false unless entailed by S.

(13) $$[\![\textbf{only S}]\!]^w \Leftrightarrow \underline{[\![\textbf{S}]\!]^w = 1} \;\&\; \forall S' \in ALT(S)[[\![\textbf{S}']\!]^w = 1 \rightarrow \lambda w'.[\![\textbf{S}]\!]^{w'} \subseteq \lambda w'.[\![\textbf{S}']\!]^{w'}]$$

If we want to know what a given instance of *only S* asserts we need to look at the particular make-up of $ALT(S)$ whose members are being negated. On Fox and Katzir's proposal ALT(S) is a set of sentences (i.e. a set of syntactic rather than semantic objects) that are derived from the prejacent, S, by replacing an F-marked constituent, X_F, in S with contextually relevant alternative expressions derivable from X_F by the rules in (14), while leaving everything else in S unchanged.

(14) **Alternatives to X_F given the discourse context C**
An expression Y is an alternative for X_F in the discourse context C if Y is

[15]The precise content and informational status of the non-exclusive inference of Only-S is the topic of a longstanding debate going back to at least Horn (1969). Since our contribution is orthogonal to these issues we adopt Horn's classical treatment but envision that appropriate changes can be made when the debate is settled.

either a lexical item, an expression salient in C, or is a sub-constituent of X.

Applying this theory in conjunction with the FCC-Only in (6) to our cases leads to the following basic analysis. In (15-a) *only* immediately precedes the subject. Its corresponding LF is sketched in (15-b). Given the surface position of *only* and (6), the focus feature of $[\textit{the cat}]_F$ is accessible for *only*. $ALT(S)$ is the set of alternative LFs we get by replacing *the cat* with contextually relevant alternatives to *[the cat]*. Since the context provides *the goose* and *the frog* (by explicit mention in the story) we have (15-c) as the set of LFs that are being negated by *only*. Parallel comments apply to (16-a).

(15) a. Only $[\text{the cat}]_F$ is holding a flag.

b. LF: Only $[[\text{the cat}]_F$ is holding a flag].

c. ALT($[[\text{The cat}]_F$ is holding a flag]) = {[The cat is holding a flag], [The goose is holding a flag], [The frog is holding a flag]}

d. $[\![\textbf{(15-a)}]\!]^w \Leftrightarrow \underline{\text{The cat is holding a flag}}$ & The goose is not holding a flag & The frog is not holding a flag.

(16) a. The cat is only holding $[\text{a flag}]_F$.

b. LF: Only [The cat is holding $[\text{a flag}]_F$].

c. ALT([The cat is holding $[\text{a flag}]_F$]) = {[The cat is holding a flag], [The cat is holding a balloon], [The cat is holding a flag and a balloon].}

d. $[\![\textbf{(16-a)}]\!]^w \Leftrightarrow \underline{\text{The cat is holding a flag}}$ & The cat is not holding a balloon & The cat is not holding a flag and a balloon.

This analysis does not yet offer a foothold to develop an explanation of Crain's puzzle. To do that we need an amendment that makes (16-a) easier to comprehend than (15-a) when the preceding discourse does not identify the X_F via a congruent question. We argue the source of this asymmetry is a second presupposition triggered by *only* which licenses a scalar inference about the prejacent (e.g. Beaver and Clark 2008). In particular, we suggest that *only* requires of its prejacent that it expresses a relatively low ranked proposition among the members of $ALT(S)$, (17).

(17) $$[\![\textbf{only S}]\!]^w \Leftrightarrow \underline{[\![\mathbf{S}]\!]^w = 1} \;\&\; \underline{\exists S' \in ALT(S)[\lambda w'.[\![\mathbf{S}]\!]^{w'} \prec \lambda w'.[\![\mathbf{S'}]\!]^{w'}]} \;\&\; \forall S' \in ALT(S)[[\![\mathbf{S'}]\!]^w = 1 \rightarrow \lambda w'.[\![\mathbf{S}]\!]^{w'} \subseteq \lambda w'.[\![\mathbf{S'}]\!]^{w'}]$$

The scalar presupposition in (17) is stated in rather weak terms – there has to be at least one member of $ALT(S)$ that expresses a proposition that is ranked higher than the prejacent proposition. A consequence of this assumption is that $ALT(S)$ needs to be orderable for *only S* to be felicitous. This can be the case either because the members of $ALT(S)$ can be (non-trivially) ranked by entailment – following the literature we call this "Q(uantity)-only" and symbolize it as $only_{\prec_Q}$ – or because there is a contextually supplied ordering relation $\prec_C$ giving rise to so called "scalar only", symbolized as $only_{\prec_C}$.

To illustrate, imagine a card game where the highest card is the Ace of Hearts. Intuitively (18) can be understood in two ways – either that Jake has no card other than the Queen of Spades ($only_{\prec_Q}$) or that he has no higher ranked card ($only_{\prec_C}$).

(18) a. Jake only has the Queen of Spades.

b. *Jake has the Queen of Spades and no other card.* $only_{\prec_Q}$
c. *Jake has the Queen of Spades and no higher card.* $only_{\prec_C}$

Replacing the Queen of Spades with the Ace of Hearts, the highest ranked card in the game, eliminates the $only_{\prec_C}$ reading because there is no element in $ALT(S)$ that is ranked above the prejacent if $ALT(S)$ were ranked based on the rules of the game. The $only_{\prec_Q}$ reading remains, however, because $ALT(S)$ can be ranked by entailment and so might include, for instance, conjunctive alternatives like *Jake has the Queen of Spades and the Ace of Hearts* which are stronger than the prejacent.

(19) a. Jake only has the Ace of Hearts.
b. *Jake has the Ace of Hearts and no other card.* $only_{\prec_Q}$
c. #*Jake has the Ace of Hearts and no higher card.* $^{*}only_{\prec_C}$

The reverse effect can be observed if we eliminate the possibility of ordering $ALT(S)$ by entailment. For instance, if the preceding question presupposes uniqueness as in (20). Now the $only_{\prec_Q}$ interpretation is unavailable but the sentence is still felicitous under the $only_{\prec_C}$ reading. If we eliminate that option as well by picking the strongest card, *only* cannot be used felicitously at all,(20-d).

(20) Which card does Jake have?
a. Jake only has the Queen of Spades.
b. *Jake has the Queen of Spades and no higher card.* $only_{\prec_C}$
c. #*Jake has the Queen of Spades and no other card.* $^{*}only_{\prec_Q}$
d. #Jake only has the Ace of Hearts.

Postulating that *only* always (or by default) triggers a scalar presupposition imposes an additional constraint on its use – its prejacent must give rise to an orderable set of alternatives. Given our assumptions about how $ALT(S)$ is generated, this can interact with the internal structure of the prejacent. Specifically, if X_F supports alternatives that can be ranked based on entailment, then further contextual support, e.g. in the form of a salient scale, is not needed. If it doesn't, the scalar presupposition can only be met by ranking $ALT(S)$ in terms of a contextually provided scale and if such a scale is not salient in C comprehenders must accommodate a scale on the fly, making comprehension more involved. Crain's puzzle, we suggest, can be understood in these terms, assuming that accommodation is a conversational skill that children are less adept at than adults.[16]

3.2. Revisiting Crain's puzzle

Recall that Crain's puzzle in its original form arises when the preceding question does not determine where X_F is in *only S*. This opens the door for other factors to play a more prominent role in locating X_F during realtime comprehension, including whether the resulting set of alternatives is orderable for a given choice of X_F.

Crain et al.'s story, repeated in (21), purposefully unfolds without providing a contextually salient scale of any sort. That is, no part of the story suggests, for instance, that the cat holding a flag is less expected, hence more noteworthy, than any other animal holding a flag. Nor is there an indication that for the cat to hold a flag is more or less expected than for it to hold a balloon. Satisfying the scalar presupposition

[16]See e.g. Aravind et al. (2022) for an independent argument to that effect.

therefore rides on ranking $ALT(S)$ by entailment unless comprehenders accommodate an ad hoc scale of some sort.[17]

(21) Here is a story about a cat, a goose, and a frog.
The cat is holding a flag. The goose is holding a flag and a balloon. And the frog is holding a balloon.

When it comes to ranking the members of $ALT(S)$ by entailment Crain et al.'s story does favor the object to be the associate over the subject because, as we saw in (16-c), $ALT(S)$ contains a conjunctive alternative *The cat is holding a flag and a balloon* which entails the prejacent. On the assumptions in (14) this alternative can be generated without further ado because the discourse contains a conjunctive alternative to *a flag*, namely *a flag and a balloon*.

For the subject DP, matters are different. If F is located on *the cat* the resulting set of alternatives is (15-c). Its members cannot be ordered based on entailment because *the cat is holding a flag* is not entailed by e.g. *the goose is holding a flag* nor the other way around. In fact, the members of $ALT(S)$ are simply logically independent. Thus, adopting a parse for *only S* where X_F is the subject entails accommodation of a scale of some sort to support the scalar presupposition of *only* and so leads to a more complex comprehension task.[18]

This asymmetry between F-marking the subject or the object as the associate of *only* can explain why S-Only sentences are more difficult to comprehend in Crain et al.'s discourse situation. Of course, to complete an account of children's behavior we also need an assumption about the status of the FCC-Only in (6) in child language but before we get to that we explain how our assumptions provide the foundation for a general comprehension model for *only S*.

3.3. Searching for X_F in Only S

Given that *only* requires access to an F-marked expression, X_F, which serves as its associate, comprehension of *Only S* can be modeled as involving a search for X_F. The search is successful if X_F is found in an accessible position according to the FCC-Only in (6) and the resulting set of alternatives supports the scalar presupposition of *only S*. Various factors can influence this search in realtime, including the surface position of *only*, the alternatives derivable from X_F as well as discourse properties such as the presence and form of a question under discussion or the salience of a potential scale to order $ALT(S)$.

We call a constituent X in S an "easily scalable constituent" (ESC) if $ALT(S)$ derived via substitutions of X according to the rules in (14) alone, without accommodation of a suitable scale, supports the scalar presupposition of *only S*. Examples of $ESCs$ are lexically scalar expressions such as numerals or gradable adjectives as well as expressions for which the context provides stronger alternatives (e.g. in the form of

[17] Against this background it is instructive to revisit Notley et al.'s Experiment 3 which notably departed from Crain et al.'s story format by explicitly introducing a scale in the story along which the subject characters were ranked.

[18] Note that the scale may very well be one based on quantity. I.e. it seems plausible that for adults, comprehension involves accommodation of a "who"-question which is number neutral and so may include conjunctive answer terms. We do not know, at this point, whether children represent accommodated sub-questions as such during comprehension since our experiment does not offer a direct test of this claim. Note, however, that the proposal does not rely on this being the case. What is important is that children can detect that *What happened?* is not a congruent QUD for the target sentence and that in the given context they find it easier to compute its meaning if they generate alternatives based on the object being F-marked.

a conjunctive alternative).

The fact that *ESCs* offer a direct path to satisfy the scalar presupposition makes them natural targets for locating F during realtime sentence comprehension. They are, as it were, attractors for F because they make it easy to satisfy the scalar presupposition. However, whether they can actually serve as the associate of *only* depends on their location in S. If they are in an eligible position given (6) all is well because all cues for locating F point to the same location. If they are in an ineligible position, however, comprehenders face conflicting cues making the comprehension task more difficult. This is summarized in Table 6 and illustrated by the examples in (22) and (23), which use the numeral *one* to illustrate the effect of an ESC in various positions inside S.

Table 6. Comprehension for *Only S* given the presence of an *ESC* in *S*

	Subject *ESC*	Object *ESC*
S-Only	EASY	DIFFICULT
VP-Only	DIFFICULT	EASY

(22) a. Only one$_F$ animal is holding a flag. S-Only + Subject ESC
b. Only the cat is holding one$_F$ flag. S-Only + Object ESC

(23) a. One$_F$ animal is only holding a flag. VP-Only + Subject ESC
b. The cat is only holding one$_F$ flag. VP-Only + Object ESC

Bare numerals such as *one* are ESCs according to our definition.[19] They have lexical alternatives which are intrinsically ranked. Because of that, locating F on a bare numeral yields a set of alternatives that can satisfy the scalar presupposition directly, without input from the discourse context. For instance, ALT(22-a) contains the prejacent *One animal is holding a flag* as well as stronger alternatives like *Two animals are holding a flag* which entail it. Because of that, we expect bare numerals to attract F during realtime sentence comprehension.

This property interacts with the surface position of *only*, however. When the numeral is placed in an eligible position given the location of *only* – inside the subject for S-Only, (22-a), or inside the VP for VP-Only, (23-b) – both factors are aligned and comprehension should be relatively easy. When the numeral is placed in an ineligible position, by contrast, – inside the VP for S-Only,(22-b), or inside the subject for VP-Only, (23-a) – comprehension should be more difficult because the factors are misaligned and point to different locations for F in S.

The results in Hackl et al. (2015) can be described in similar terms. The preceding question provides a cue for the location of F via QAC. Given that both *Who is holding a flag?* and *What is the cat holding?* are number neutral with regard to their answer terms, the question provides salient alternatives via its denotation (roughly the set of its possible answers as sketched in (24)) which support the scalar presupposition triggered by *only*.

(24) a. $[\![\textbf{Who is holding a flag?}]\!]^w \approx$ {The cat is holding a flag, the goose is holding a flag, ..., the cat and the goose are holding a flag, ...}

[19] By "bare numerals" we intend to refer to unmodified numeral expressions which, unlike modified numerals like *more than n, at least n, ...*, can serve as associates of *only*.

b. $[\![\textbf{What is the cat holding?}]\!]^w \approx$ {The cat is holding a flag, the cat is holding a balloon, the cat is holding a flag and a balloon}

Crain's puzzle arises when the preceding question leaves it open where the F-marked associate of *only* should be and comprehenders need to determine its location based on the properties of the prejacent and other aspects of the discourse context. On the assumption that focus alternatives are syntactic rather than semantic objects, derivable from a Fox and Katzir-style algorithm, locating F on the object (which corresponds to accommodating an object-question) offers a more direct path to satisfying the scalar presupposition because in the utterance context *a flag* is an ESC. Locating it on the subject, by contrast, requires accommodating a scale of some sort since *the cat* is not an ESC in C resulting in the set of alternatives coming out of the prejacent as such not being orderable.

Our comprehension model is based on very general assumptions that are not specific to any population of comprehenders, whether children or adults. Thus, if correct, we should see both adults and children exhibit behavioral patterns reflecting comprehension ease or difficulty depending on the location of the ESC and *only*. The ways in which a given speaker exhibits the predicted sensitivity may, of course, differ between children and adults. One plausible expectation is that children may deal with the comprehension difficulty by assigning non-adult meanings – assuming that they can somehow override FCC-Only – while adults might only be slowed down in their processing. The next two sections present experimental evidence from adults and children in support of these predictions.

4. Adult comprehension of *only*

Our approach to Crain's puzzle leads us to expect a parallel pattern for adults. This section presents three experimental studies supporting that expectation.

4.1. Method and materials

All of our adult experiments used the same task and environment, differing only minimally in the composition of the target items, with all other aspects constant. We used a timed inference task as shown in Figure 2.

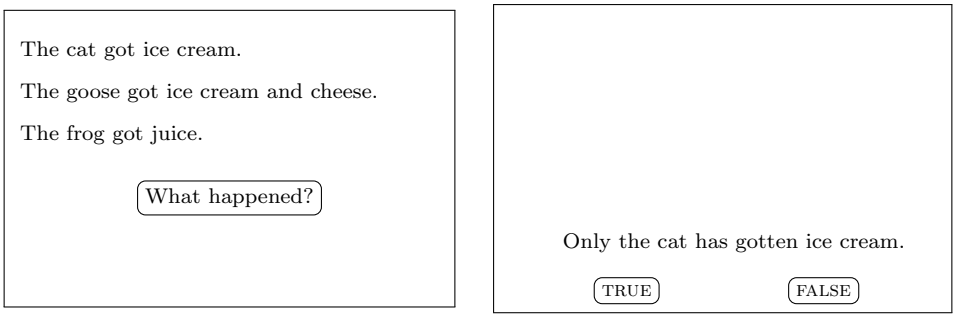

Figure 2. Screen 1 of timed inference task. Screen 2 of timed inference task.

Participants read a set of statements on a computer screen, which jointly describe a situation similar to the one depicted in Crain et al. (1992, 1994)'s stories. Below these

"premises" was a button labeled "What happened?". Upon clicking that button, the premises disappeared and a new statement was presented on the screen that did or did not truthfully describe the situation characterized by the story. Participants were asked to indicate whether the set of statements supported the target sentence by clicking TRUE or FALSE as quickly and reliably as they could.

Participants saw 12 S-Only and 12 VP-Only sentences pseudo-randomly interspersed between 34 filler items. For each group of target sentences six were true under the adult interpretation and six were false. Moreover, the stories were constructed so that a target sentence that was true under the S-Only interpretation was false under the VP-only counterpart interpretation and vice versa. For instance, the target sentence in Figure 2 is false under the adult interpretation but true if it were given a VP-Only parse.

The experiment was presented over the internet using the IbexFarm platform hosted on institutional servers. Participants were recruited through Amazon's Mechanical Turk and compensated for their participation at current market rates.[20] Each participant gave written consent, completed a short questionnaire and then 2 practice items to get accustomed to the task, before starting the experiment.

Participants' responses and response times were recorded using a customized Ibex script. A participant's data were excluded from analysis if they indicated that they were a non-native English speaker during recruitment or if they responded accurately less than 75% of the time across filler items.[21] Response times were excluded if they accompanied an incorrect response or if they exceeded 15 seconds. Accuracy was analyzed with a generalized linear mixed-effects model with a logit link and log-transformed response times with a linear mixed-effects model, in both cases of the form `dependent attachment + (1 | participant)`. We report a by-participant random intercept only: a by-item intercept accounted for effectively zero variance and produced singular fits, and a by-participant random slope for attachment cannot be estimated in the between-subjects child experiments, so we adopted this structure uniformly across all experiments rather than varying it case by case (see footnote 33). Degrees of freedom for the LMEMs are Satterthwaite approximations. For the attachment contrast in each experiment we report an effect size in the text, an odds ratio for accuracy, and for logRT the multiplicative response-time ratio $\exp(\beta)$, each with a 95% confidence interval. These are profile-likelihood intervals wherever the profile converges; for some of the logistic models it does not, and we fall back on Wald intervals built from the fixed-effect standard errors. The analysis script records which of the two was used for every model.

4.2. Adult Experiment 1 – Baseline

Our baseline experiment tests a basic prediction of our comprehension model: S-Only sentences require a more complex comprehension process than VP-Only sentences in an environment like Crain et al.'s where the preceding question leaves the location of the X_F underspecified and the discourse context does not offer a salient ad hoc scale along which the alternatives to the prejacent could be ranked. To this end, Experiment 1 used target sentences in the mold of Crain et al. (1992, 1994)'s original paradigm

[20] The current market rate refers to the rate of participating in short linguistic experiments via Amazon's Mechanical Turk in the years 2013 and 2014, when the studies were conducted.

[21] Since participants were recruited online and were incentivized to complete the study as quickly as possible, as many as 60% of participants could be dropped using these two criteria. We aimed to have at least 25 remaining participants before proceeding with any further analysis.

with minor modifications suggesting more strongly than the original that there was a sequence of events (e.g. distribution of food items) leading to an outcome described by the target sentence. An example item in all four conditions is shown in (25) and (26).

(25) The goose got ice cream and cheese.
The cat got ice cream.
The frog got juice.

What happened?

Only the cat has gotten ice cream. **False**

The cat has only gotten ice cream. **True**

(26) The goose got cheese.
The cat got ice cream and cheese.
The frog got juice.

What happened?

Only the cat has gotten ice cream. **True**

The cat has only gotten ice cream. **False**

Results: We recruited 64 participants and analyzed data from 40 participants after performing participant exclusions. Participants were equally accurate on both S-Only and VP-Only target sentences (85.5% versus 84.7%, respectively, $Pr(> |z|) = 0.695$). However, their response times were significantly faster in the VP-Only condition than in the S-Only condition (7.92 versus 8.07 *log*-transformed RT, $t = -5.08$, $df = 769$, $p = 4.64e{-}7$). The corresponding effect sizes are an odds ratio of 0.93, 95% CI $[0.65, 1.34]$, for accuracy, and a response-time ratio of $\exp(\beta) = 0.87$, 95% CI $[0.83, 0.92]$, for RT, that is a reduction of 12.7%; back-transformed, the geometric mean response time is 3212 ms for S-Only against 2744 ms for VP-Only. This is shown in Figure 3.[22,23]

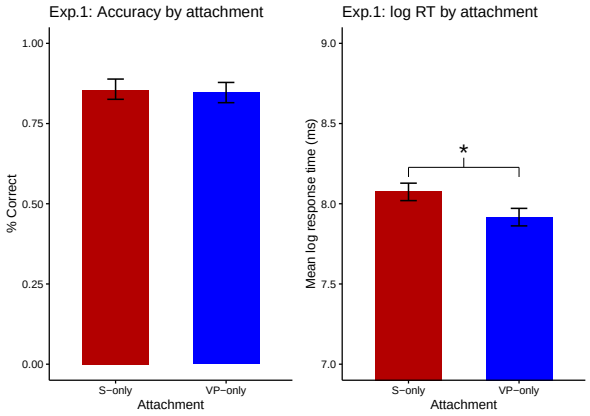

Figure 3. Mean accuracy and Log-RT for Adult Experiment 1.

Discussion: Even though our timed inference task put non-trivial demands on participants' short-term memory, accuracy was quite high across both S-Only and VP-Only sentences. This suggests that participants had little difficulty with determining the correct associate of *only*, regardless of attachment site. However, there is a notable contrast between their response times between S-Only and VP-Only, with the former taking significantly longer. This is predicted by our model, which claims S-Only sen-

[22]The plots here and below collapse data across truth-values, unless stated otherwise, and display 95% confidence interval based error bars.

[23]Restricting the RT model to correct trials only, as reported in the main text, follows our practice throughout the paper. As a robustness check, we also ran the RT model on all target trials, including incorrect ones: the VP-Only advantage remains significant ($\beta = -0.110$, $SE = 0.0254$, $t = -4.35$, $p = 1.54e{-}5$), qualitatively unchanged from the correct-trials-only estimate reported above.

Table 7. Results of statistical tests for Adult Experiment 1

	Accuracy Estimate	Std. Error	z value	$Pr(>\|z\|)$	**log-RT** Estimate	Std. Error	t value	df	$Pr(>\|t\|)$
(Intercept)	1.94	0.182	10.7	$1.25e-26$	8.07	0.0646	125	42.3	$6.28e-56$
VP-Only	−0.0725	0.185	−0.392	0.695	−0.135	0.0266	−5.08	769	$4.64e-7$

tences to be more difficult to comprehend in this environment than VP-Only sentences. Thus, Crain's puzzle has its counterpart in adult comprehension: an increase in RTs which we take to reflect an increase in comprehension complexity.

4.3. Adult Experiment 2a,b – Numeral in subject position

Our second adult experiment tests our predictions for bare numerals illustrated in (22) and (23) above. Recall that by our model, bare numerals are ESCs: they are intrinsically scalar, with lexical substitutions giving rise to sets of alternatives directly rankable by quantity, without requiring accommodation of an ad hoc scale. This makes them attractors for F as comprehenders search for the associate of *only* in realtime and so should, in conjunction with the FCC-Only, produce a specific pattern of comprehension difficulty or ease given the surface of position of *only* and the bare numeral in S.[24] To test this prediction we adapted our materials from Experiment 1 in two ways.

In **Experiment 2a** we replaced the subject DP of our original target sentences with *one of the animals* while leaving everything else unchanged. This is shown in (27) and (28), where the target sentences have the same truth values as the ones above.

(27)
The goose got ice cream and cheese.
The cat got ice cream.
The frog got juice.

What happened?

Only one of the animals has gotten ice cream.

One of the animals has only gotten ice cream.

(28)
The goose got cheese.
The cat got ice cream and cheese.
The frog got juice.

What happened?

Only one of the animals has gotten ice cream.

One of the animals has only gotten ice cream.

Results: We recruited 64 participants and included data from 26 participants in our analysis after performing participant exclusions as described in the methods section. As shown in Figure 4, adult participants performed significantly less accurately on the VP-Only sentences than they did on the S-Only sentences (62.6% versus 90.3%, respectively, $Pr(>|z|) = 2.04e-14$). They also took significantly longer to respond to the VP-Only sentences when compared to the S-Only sentences (8.47 versus 8.30 *log*-transformed RT, $t = 5.72$, $df = 441$, $p = 1.93e{-}8$). The effect sizes are an odds ratio of 0.17, 95% CI $[0.11, 0.27]$, for accuracy, and a response-time ratio of $\exp(\beta) = 1.21$, 95% CI $[1.13, 1.29]$, an increase of 20.6%; back-transformed, 4023 ms for S-Only against 4770 ms for VP-Only. This is shown in Figure 4.

[24]Reviewer 3 points out that there may be non-trivial differences between S-Only and VP-Only sentences whose subjects contain a numeral as in (27) with regard to what questions they might be naturally understood to answer. Specifically, *One of the animals has only gotten ice cream* may be naturally seen as an answer to a question that itself contains *only* as in *How many animals have only gotten ice cream?* while its S-Only counterpart wouldn't. How this difference interacts with our comprehension model is not entirely clear to us and left for future research.

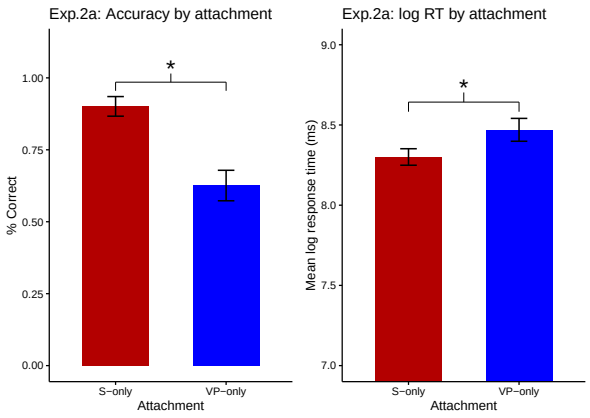

Figure 4. Mean accuracy and Log-RT for Adult Experiment 2a.

Table 8. Results of statistical tests for Adult Experiment 2a

	Accuracy Estimate	Std. Error	z value	$Pr(>\|z\|)$	**log-RT** Estimate	Std. Error	t value	df	$Pr(>\|t\|)$
(Intercept)	2.29	0.214	10.7	$9.33e-27$	8.31	0.0727	114	26.4	$3.83e-37$
VP-Only	−1.76	0.230	−7.65	$2.04e-14$	0.187	0.0327	5.72	441	$1.93e-8$

Discussion: Given the results of our adult baseline experiment, the contrast in comprehension accuracy and RT between S-Only and VP-Only conditions is quite striking. We saw earlier that in our Crain-style environment (*What happened?*, no contextual support for an ad hoc scale) S-Only is harder to comprehend than VP-Only. A seemingly minor change in the make-up of the subject so that it includes an ESC reverses this asymmetry. Now VP-Only is more difficult to comprehend as shown by both dependent measures. Comprehension accuracy remained high for S-Only but suffered a marked drop in the case of VP-Only. Likewise, taking the difference in target sentence length into account, RTs for S-Only remained basically unchanged when compared with the results in Experiment 1 but increased substantially for VP-Only.

Experiment 2b is a slight variant of the previous study to examine whether the effect of a numeral inside the subject is specific to *one* or generalizes to other numerals. To this end we adapted the materials to support a switch from *one of the animals* to *two of the animals*. Specifically, in order to accommodate the use of *two* in the target sentence, additional premises were added as shown in (29) - (30).

(29)
The goose got cheese.
The cat got ice cream.
The frog got cheese.
The penguin got cheese and juice.

What happened?

Only two of the animals have gotten cheese.

Two of the animals have only gotten cheese.

(30)
The goose got cheese.
The cat got ice cream.
The frog got juice.
The penguin got cheese and juice.

What happened?

Only two of the animals have gotten cheese.

Two of the animals have only gotten cheese.

Results: Data from 39 participants were included in the analysis out of 68 recruited. As in Experiment 2a, participants were significantly less accurate in the VP-Only

condition than in the S-Only condition (80.4% versus 87.6%, respectively, $Pr(>|z|) = 2.49e-3$). Likewise, they also responded to the VP-Only sentences slower than the S-Only sentences, and here too the contrast reaches significance (8.26 versus 8.19 *log*-transformed RT, $t = 3.04$, $df = 744$, $p = 2.44e{-}3$). The effect sizes are an odds ratio of 0.57, 95% CI $[0.40, 0.82]$, for accuracy, and a response-time ratio of $\exp(\beta) = 1.08$, 95% CI $[1.03, 1.13]$, an increase of 7.8%; back-transformed, 3596 ms for S-Only against 3869 ms for VP-Only. While these contrasts are somewhat weaker than in the case of Experiment 2a, they do suggest that the effects of *one* on the comprehension of *only* extend to other bare numerals, see Figure 5.

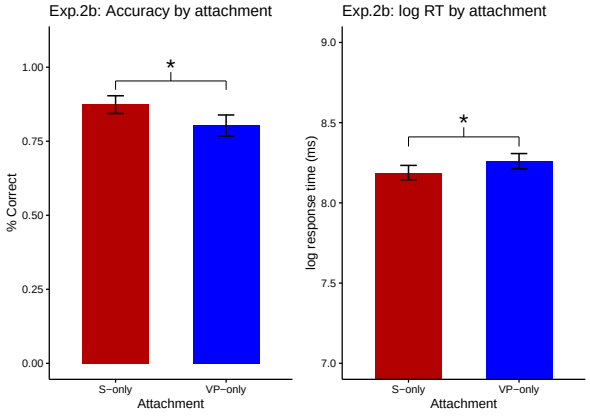

Figure 5. Mean accuracy and Log-RT for Adult Experiment 2b.

Table 9. Results of statistical tests for Adult Experiment 2b

	Accuracy Estimate	Std. Error	z value	$Pr(>\|z\|)$	**log-RT** Estimate	Std. Error	t value	df	$Pr(>\|t\|)$
(Intercept)	2.06	0.173	11.9	$1.05e-32$	8.20	0.0549	149	41.8	$1.13e-58$
VP-Only	−0.558	0.185	−3.02	$2.49e-3$	0.0750	0.0247	3.04	744	$2.44e-3$

Discussion: The findings of Experiment 2a and 2b are in line with the predictions of our model. Putting an ESC such as a bare numeral inside the subject of an S-Only sentence creates a situation where both factors governing the search for X_F point to the same location. This makes comprehension easier compared to the baseline in Experiment 1. For a VP-Only sentence, putting an ESC in subject position increases comprehension difficulty because the bare numeral distracts the search for X_F from its grammatical target which, by the FCC-Only, has to be on or inside the VP.

4.4. General discussion of adult experiments

In Section 3 we proposed a comprehension model to account for Crain's puzzle which makes parallel predictions for children and adults. The model relied on the scalar presupposition of *only* and a syntactic method of generating the set of alternative, which jointly govern the search for the F-marked associate of *only* when the preceding question leaves its location under-determined. This section presented evidence supporting

our approach from three experiments on adults.

Experiment 1 translated the original setup of Crain's puzzle into a timed inference task and revealed that adults do find S-Only sentences more difficult to comprehend in realtime than VP-Only sentences. Experiments 2a,b investigated whether this basic contrast in comprehension difficulty can be modulated by strategically placed ESCs (either *one of the animals* or *two of the animals* rather than a simple definite subject like *the cat*). We found that these intrinsically scalar subjects indeed made S-Only sentences easier to comprehend but also that they made VP-Only sentences more difficult. These findings provide striking and clear support for our approach.

5. Child studies

All of our child experiments used a Truth-Value Judgment task (Crain and Thornton 1998) emulating the set-up of Crain et al.'s original experiments with minor modifications specific to our predictions. Experiment 1 (our baseline) replicates Crain et al. (1992, 1994)'s original observation with our own methods and materials. Experiments 2a,b test the hypothesis that ESCs inside the subject increase the rate of adult-like interpretations for S-Only sentences, mirroring the results from our adult Experiments 2a,b. Participants were English-speaking children from Boston-area daycares and the Boston Children's Museum, representing a range of socioeconomic and ethnic backgrounds. Sessions took about ten minutes per participant and were conducted with caretakers present throughout.

5.1. Child Experiment 1 – Baseline

Method and materials: Children were told simple stories supported by PowerPoint slides which showed simple animations of animal characters receiving food items. The experimenter, a native speaker of English, narrated the story and advanced the slides. At the end of the story, the experimenter asked Kermit, *Can you tell me what happened?* Kermit responded with a target sentence (either with S-Only or VP-Only) or a filler sentence that did not contain *only*. A sample item is shown in Figure 6.[25]

P: This is a story about a cat, a goose and a frog.

The cat got ice cream.

The goose got ice cream and cheese.

The frog got ice cream and juice.

Kermit, can you tell me what happened?

K: Only the cat got ice cream.

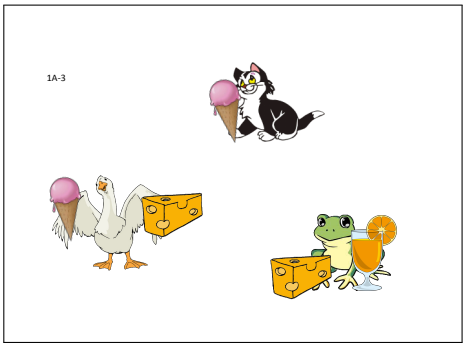

Figure 6. Trial structure for Experiment 1. Final Power point slide in Experiment 1.

After Kermit gave his answer, the child participant was asked to judge whether the

[25]Experiment 1 serves as the baseline for Experiments 2a and b. It was previously reported in Hackl et al. (2015) where it also served as baseline replication of Crain et al. (1994).

puppet was right or wrong/silly, and if silly, they were encouraged to provide a justification. While making judgments and providing justifications, children could see the events as a still image on the computer screen, comparable to the picture verification task used by Crain et al. and others. Children were encouraged and praised equally for *Right* and *Silly* answers.

The experiment had 8 target trials (4 with S-*Only* and 4 with VP-*Only*) and 4 filler trials within one session. Half the items in each condition were designed to be true under an adult interpretation relative to the story, and the other half to be false. Importantly, stories were constructed such that if a target sentence in a given condition (e.g., S-*Only)* was true under the adult S-*Only* interpretation, it would be false under a non-adult, VP-*Only* interpretation, and vice versa. The presentation order of target and filler items was pseudo-randomized into two lists which were then randomly assigned to participants.

Results: Fifty-two participants were recruited for this experiment. Children who incorrectly answered two or more filler items or who gave the same answer on nearly all trials (the yes-/no-sayer criterion described for Experiment 2a below) were excluded (N=6). Data from 46 participants (3;3 – 6;9, M=4;10) were included in the analysis. We observed a significant effect of the attachment site on accuracy rates, with the rate of adult-like responses being significantly lower for S-*only* (29.9%) than for VP-*only* (79.9%) as revealed by a linear mixed effects logistic regression model, Table 10. Figure 7 summarizes the results of Experiment 1, plotting adult-like responses for S-Only and VP-Only.[26]

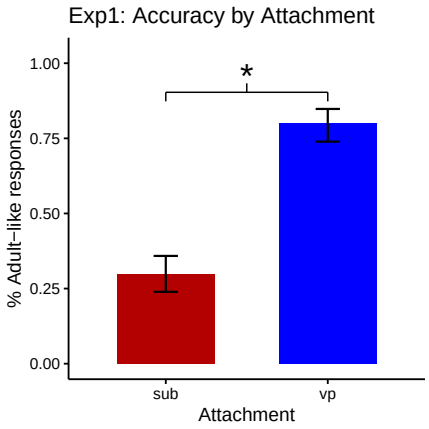

Figure 7. Mean accuracy Child Experiment 1.

Discussion: The difference in the rate of adult-like interpretations between S-Only and VP-Only sentences is striking. The former is well below chance while the latter is close to 80%, well above chance. This replicates Crain et al.'s original finding in a slightly stronger form and so provides a suitable baseline for our follow-up experiments.

[26]The model reported here is the frequentist GLMEM used throughout the paper. Its fit is singular: the by-subject variance is estimated at the boundary. As a check we refit the same model with a weakly informative prior on the random effect (`bglmer`, `blme`); that fit is likewise singular but converges on the same estimates ($\beta = 2.27$, $SE = 0.249$, $z = 9.10$ for the attachment contrast), so the result does not depend on the estimation method.

Table 10. Results of statistical tests for Child Experiment 1

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	Odds ratio	
(Intercept)	-0.852	0.161	-5.29	$1.20e-7$	0.43 [0.31, 0.59]	***
VP-Only	2.23	0.244	9.13	$6.84e-20$	9.32 [5.77, 15.0]	***

5.2. Child Experiments 2a,b — Numerals in subject position

Method and materials: The goal of Experiments 2a,b is to assess whether children's comprehension of S-Only sentences is facilitated by ESCs inside the subject similar to what we have seen in adults. The first experiment, Experiment 2a uses exactly the same materials as child Experiment 1 except that the subject of the target sentences changed from *the cat* to *one of the animals*, (31).

(31) Target sentences for Child Experiment 2a (stories and images identical to Child Experiment 1)

a. Only one of the animals got ice cream.
b. One of the animals only got ice cream.

Experiment 2b was designed analogously to Adult Experiment 2b and so used modified stories and images to support target sentences using *only two of the animals*. This is shown in Figure 8.[27] Both Experiment 2a and 2b had Attachment of *only* as a between factor. One group of participants saw only 4 S-Only target sentences combined with 4 filler items and the other group saw 4 VP-Only targets. In all other respects, Experiments 2a,b were structured like Experiment 1.

P: This is a story about a bird, a bunny, a penguin and a dolphin.

The dolphin got ice cream.

The bunny got jelly, cake and bread.

The penguin got ice cream.

The bird got ice cream, bread and juice.

Kermit, can you tell me what happened?

K: Only two of the animals got ice cream.

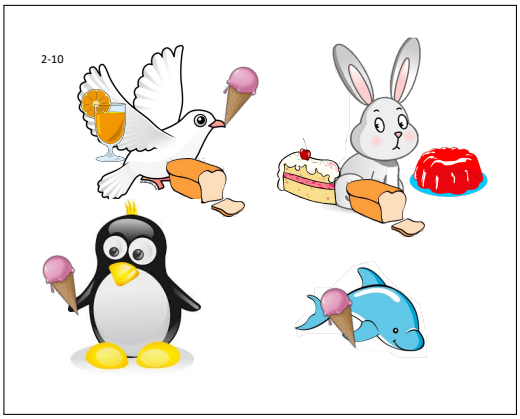

Figure 8. Trial structure for Experiment 2b. Final Power point slide in Experiment 2b.

Results for Experiment 2a: We recruited 155 children. Children who incorrectly answered two or more filler items, did not answer all questions, or favored one answer[28] were excluded, as well as those outside ages 3-6. Data from 128 participants (3;0 – 6;11, M=5;0) were included in the analysis. Among them, the S-Only condition included 56

[27]The stories and images for Experiment 2b had more food items than their counterparts in the adult experiment for unrelated reasons. We simplified the adult stories to reduce their memory load since participants did not see any images and had to memorize the premises before responding.

[28]Children were considered a "yes-sayer" or a "no-sayer" if more than 87.5% of their answers were yes or more than 87.5% were no, respectively. This is to catch children who said the same thing almost every time but still managed to get 3 of the 4 fillers correct by luck.

children (3;0-6;10, M=5;0), and the VP-Only condition included 72 children (3;1-6;11, M=5;0). Data were analyzed as above.[29]

Rates of adult-like interpretations were high for both S-Only and VP-Only and, in fact slightly higher even for S-Only (72.3%) than for VP-Only (66.0%), with no significant difference between them. Both rates were, however, significantly above chance as assessed by a one sample t-test of accuracy rates (t = 4.97, df = 55, p = 6.86e-06 and t = 3.54, df = 71, p = 7.13e-04). This is in stark contrast with the results from Experiment 1 where we saw an adult-like rate for S-Only of about 30%. Note, also, that the adult-like rate for VP-Only has dropped markedly from close to 80% in Experiment 1 to 66.0% here. This is shown in the panel on the left in Figure 9. As a supplementary analysis fit uniformly across all our accuracy data, we also ran a mixed-effects logistic regression on Experiment 2a's accuracy data: this model likewise reveals no significant effect of attachment ($\beta = -0.443$, $SE = 0.545$, $z = -0.813$, $p = 0.416$), consistent with the t-test results reported above.

Results for Experiment 2b: 60 participants were recruited, with the same exclusion criteria as 2a. Data from 51 participants (3;5 – 6;10, M=5;0) were included in the analysis. For the S-Only condition we had data from 27 children (3;5-6;6, M=4;10), and for the VP-Only condition we have data from 24 children (3;6-6;10, M=5;1).

The rate of adult-like interpretation for S-Only was 82.4%, even higher than in Experiment 2a, while it dropped markedly for VP-Only to 45.8%. This is shown in the panel on the right in Figure 9. Data analysis was performed as before and revealed that the S-Only condition was significantly more accurate than the VP-Only condition, Table 11.

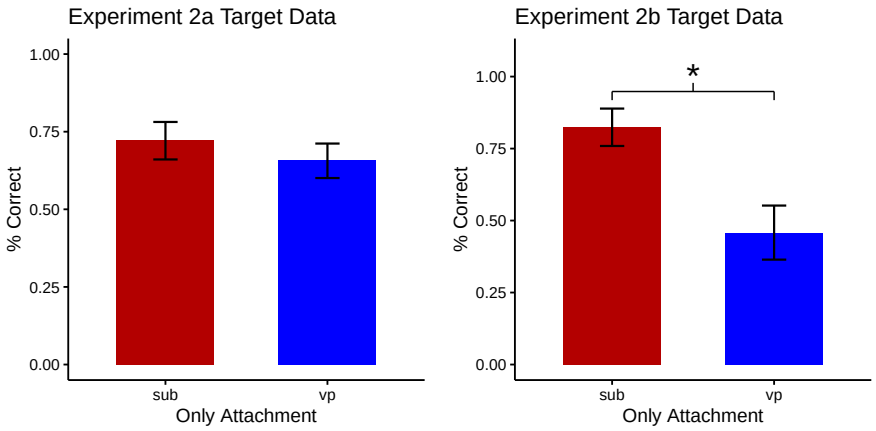

Figure 9. Mean Accuracy Child Experiment 2a,b

5.3. Discussion of Child Experiments

The main goal for our child experiments was to test whether comprehending S-Only is facilitated if the subject contains an ESC, in a way parallel to the adult results. To

[29]Except we did not include estimates for random slopes for participants due to the between-subjects nature of this experiment.

Table 11. Results of statistical tests for Child Experiment 2b

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	Odds ratio	
(Intercept)	2.25	0.524	4.29	$1.77e-05$	9.47 [3.80, 33.0]	***
VP-Only	−2.50	0.698	−3.59	$3.36e-04$	0.082 [0.016, 0.29]	***

this end, we replicated Crain's puzzle to establish a baseline and then tested children in the relevant age range on target sentences with bare numeral subjects like *one of the animals* and *two of the animals*.

Our predictions were confirmed by both child experiments: the rate of adult-like interpretations for S-Only rose from 29.9% in our baseline to 72.3% and 82.4% respectively. This is quite remarkable because DPs like *one/two of the animals* are in every way more complex than (contextually licensed) singular definites like *the cat*: they are longer, they involve a partitive structure based on a plural definite, the use an abstract counting expression, they evoke scalar alternatives, and their meaning is quantificational rather than referential. Yet, in the context of S-Only they make comprehension easier, even for our youngest participants as can be seen in Table 12, which displays the accuracy (the rate of adult-like interpretations) for S-Only and VP-Only by age.

Table 12. Accuracy in Experiments 2a and 2b by age

	3 ya		4 ya		5 ya		6 ya	
	S-Only	VP-Only	S-Only	VP-Only	S-Only	VP-Only	S-Only	VP-Only
ONLY ONE	71.9%	58.3%	61.3%	64.6%	73.6%	56.8%	92.5%	89.3%
ONLY TWO	85%	62.5%	75%	45%	84.1%	25%	100%	57.1%

The facilitation effect of bare numerals in subject position is specific to S-Only. Rates of adult-like interpretations for VP-Only drop compared to baseline, that is, bare numeral subjects make comprehension of VP-Only more difficult. This is most evident in the aggregate in Experiment 2b where the rate of adult-like interpretations for VP-Only was at chance level but it also appears to be a stable phenomenon for 3-6 year olds, Table 12.[30,31]

The main takeaway from our child experiments is that we see that children can reliably interpret S-Only in an adult-like way. Importantly, our cases did not offer any more discourse support than Crain et al.'s experiments did. The critical piece instead was the inherently scalar nature of the bare numeral subjects. This is expected under

[30]Our model does not immediately predict that *only two* has a stronger effect on comprehension than *only one*. We suspect that one reason for the finding lies in the fact that *two* is naturally more stressed, hence more prominent in our target sentences than *one*, even when *only* is on the VP. This means that it is a stronger attractor for F and a stronger facilitator for S-Only, and a more powerful distractor for VP-Only.

[31]Another noteworthy detail in the data is that 5-year-old children exhibit the strongest contrast between S-Only and VP-Only while younger participants show the significant improvement for S-Only but less of an effect for VP-Only. A cause for this age effect could be the development of number knowledge. Older children have a more solid grasp of numbers than younger ones and may be better at computing the meaning of scalar alternatives, making them more susceptible to the effect of numerals on the search for the associate of *only*. To test this descriptive age pattern formally, we fit a mixed-effects logistic regression on accuracy with attachment, (scaled) age in months, and their interaction as fixed effects, separately for Experiment 2a (age range 36–83 months) and Experiment 2b (age range 41–82 months). In Experiment 2a there is a significant main effect of age ($\beta = 0.951$, $SE = 0.414$, $z = 2.30$, $p = 0.0215$) but no significant attachment × age interaction ($\beta = -0.137$, $SE = 0.537$, $z = -0.256$, $p = 0.798$). In Experiment 2b neither the main effect of age ($\beta = 0.484$, $SE = 0.533$, $z = 0.908$, $p = 0.364$) nor the attachment × age interaction ($\beta = -0.412$, $SE = 0.670$, $z = -0.615$, $p = 0.538$) reaches significance. Thus, while accuracy in Experiment 2a improves with age overall, we find no statistical evidence that the size of the S-Only/VP-Only contrast itself changes with age in either experiment.

our comprehension model as is the fact that kids and adults exhibit the same selective sensitivity of S-Only and VP-Only to the presence of bare numerals in subject position.

6. General discussion

6.1. Cross-experiment comparisons

The discussion above compares Experiments 1, 2a and 2b qualitatively, and here we summarize the results of carrying out a statistically principled comparison.[32] We pooled the target trial accuracy data from Experiments 1, 2a and 2b separately for the adult and child population, coding participant and item identifiers as specific to their own experiment (i.e., no participant or item is treated as shared across experiments), and fit

```
correct ~ attachment * experiment + (1 | subject)
```

with a binomial link function, where `attachment` (S-Only vs. VP-Only) and `experiment` (Experiment 1, treated as the reference level, vs. 2a and vs. 2b) are both fixed effects. The coefficient of interest is the `attachment * experiment` interaction: it tests whether the size of the S-Only/VP-Only accuracy gap in Experiment 2a, and separately in Experiment 2b, differs from the gap observed in the Experiment 1 baseline.[33] For adults, the shift from Experiment 1 to Experiment 2a is highly significant ($\beta = -1.72, p < 10^{-8}$); the shift to Experiment 2b is smaller and only marginal by this measure ($\beta = -0.49, p = .061$). We ran the analogous model on log response times for the adult data, restricted to correct trials for consistency across the three experiments.[34] This shows a significant shift from baseline in both Experiment 2a ($\beta = 0.32, p < 10^{-13}$) and 2b ($\beta = 0.21, p < 10^{-8}$), consistent with our broader account that comprehension difficulty, which does not fully surface in accuracy, still surfaces in processing time. For children, where the manipulation produces a reversal rather than an attenuation of the baseline pattern, the same accuracy model shows a significant shift from Experiment 1 for both Experiment 2a ($\beta = -3.04, p < 10^{-10}$) and 2b ($\beta = -5.00, p < 10^{-13}$).

In summary, for comprehension, both children and adults showed a significant shift from the baseline to Experiment 2a. For Experiment 2b, children showed a significant shift while adults showed a marginally significant shift. Because correctness is often near ceiling for adults, the use of RT is taken as a behavioral measure that reflects the same underlying process as in children. For RT, the shifts from baseline to Ex-

[32] We thank Reviewer 2 for recommending that we do this.

[33] This is the most minimally-specified version of the random-effects structure that is still meaningful for every experiment. A by-item random intercept was dropped throughout because it consistently accounted for effectively zero variance and produced singular fits; a by-subject random slope for `attachment` was likewise dropped throughout, both because it has no within-subject contrast to be estimated from in Child Experiments 2a and 2b (attachment is between-subjects there) and because, even where a slope is structurally justified (all three adult experiments, Child Experiment 1, all within-subjects), it produced singular or degenerate fits. Across the individual per-experiment models, retaining the by-item intercept gave a non-singular fit in only 3 of 9 cases, with no coherent grouping among the three, which motivated dropping it uniformly rather than case by case. The fixed effect estimates reported here are consistent with those obtained from the individual models before this simplification.

[34] As a robustness check, we also ran this model on all target trials, including incorrect ones. The shifts from baseline remain significant for both Experiment 2a ($\beta = 0.23, p < 10^{-8}$) and Experiment 2b ($\beta = 0.17, p < 10^{-5}$), smaller in magnitude than the correct trials only estimates reported in the main text but qualitatively unchanged.

periments 2a and 2b are significant. These comparisons thus strongly corroborate the interpretation of our results. ESC associates inside the subject make S-Only sentences easier to comprehend for children and adults as reliably revealed by accuracy rates for the former and by response times for the latter population.

6.2. Conclusion

Crain et al. (1992, 1994) discovered a striking asymmetry between children's interpretation of S-Only and VP-Only and took it to show that S-Only is acquired later in development. Follow-up work has added considerable richness to the empirical landscape, but constant among all approaches to Crain's puzzle we are aware of is the claim that S-Only is special, either because its acquisition is delayed, (Crain et al. 1992, 1994; Notley et al. 2009; Zhou and Crain 2010, a.o.), or because its resultant sentence structure is in conflict with the default assignment of prosodic prominence in the language, leading to processing difficulty somehow resolved by assigning the S-Only sentence a VP-Only meaning, (Höhle et al. 2016; Su 2024, a.o.).[35,36]

This paper argues that Crain's puzzle is fundamentally a comprehension phenomenon, and that it is to be explained within a general comprehension model that applies to S-Only and VP-Only alike. Under our model it is predicted that Crain's puzzle can be reversed if the relevant factors favor the S-Only interpretation: a higher rate of adult-like interpretations for S-Only sentences than VP-Only sentences because they tend to interpret VP-Only sentences to mean what S-Only sentences say in such circumstances.

Hackl et al. (2015) documented such a reversal with incongruent question-answer pairs. Here we show that children exhibit this effect with bare numeral subjects even when all other factors are as in the original Crain et al. setup, which requires accommodation of a suitable question. This shows that scalar items can play the same role in helping or hindering comprehension of *only* as congruence. It also shows, more generally, that children do have access to the adult interpretation for S-Only and that an account of Crain's puzzle in terms of an immature grammar for S-Only is not tenable.

Our corpus study provides additional support for the claim that S-Only is part of children's grammar. We found adult-like uses of S-Only even by very young children and, most crucially, we found no occurrences of *only* that could be described as counterparts of Crain's puzzle or its reverse. In fact, their productions suggest that children know the syntactic, semantic and pragmatic properties of *only* in exquisite detail, even at a very young age. This is a remarkable fact in its own right, given the complexities involved in learning the grammar of *only*.

---

[35]The proposal in Notley et al. (2009) and Zhou and Crain (2010) combines a processing component similar to Höhle et al. (2016) with the assumption that S-Only in Child language is an adverbial quantifier while in adult English it forms a DP quantifier when preceding a DP. The latter assumption explains why processing difficulty can lead to the non-adult, VP-only interpretation. Höhle et al. (2016) and Su (2024) do not offer a specific hypothesis why the processing difficulty is resolved in this way rather than e.g. by guessing its meaning or by ignoring *only*, or by assigning some other interpretation.

[36]Reviewer 3 suggests yet another, quite intriguing account of Crain's puzzle on which it arises not because S-Only is (effectively) reanalyzed as VP-Only but because children parse the S-Only sentences with a (default) exhaust operator closing off the VP while leaving pre-subject *only* where it appears in the surface string. On this view, *Only the cat got ice cream* has a meaning akin to *The cat is the only animal who got nothing but ice cream*, which is supported by the relevant scene. However, while such a parse may indeed be responsible for some instances of non-adult-like comprehension of S-Only, we do not see how it could explain the effect of adding a bare numeral to the subject nor do we see how it could account for the reverse Crain effect observed in Hackl et al. (2015) where children seem to mis-analyze VP-Only sentences when preceded by a subject question as if they were S-Only sentences.

Crain's puzzle arises, on our account, during realtime comprehension because interpreting *only* involves a search for an F-marked expression that serves as its associate. This search, we argue, is subject to a variety of potentially conflicting factors key among which is the need to generate a set of formal alternatives that supports the scalar presupposition of *only*. A consequence of the model is that inherently scalar expressions such as bare numerals can play the same role in modulating the comprehension of sentences with *only* as congruent and incongruent questions do. Independent evidence for this comes from our adult experiments which revealed that comprehension difficulty in adults is sensitive to the same constellation of factors that produce Crain's puzzle or its reverse in child language.

A remaining question for the comprehension model we put forward concerns what distinguishes children from adults in their processing of sentences with *only*. Our account suggests that the difference is fundamentally one of comprehension rather than competence: children and adults share the same grammar for *only*, but differ in their ability to handle the demands that realtime comprehension places on that grammar, particularly when accommodation is required. We believe that Höhle et al. (2016)'s findings that measures of executive function correlate with children's preponderance to exhibit Crain's puzzle offer an interesting direction for further investigation.[37]

Finally, our account suggests several additional factors that may modulate comprehension difficulty beyond those directly tested here, including information packaging and syntactic correlates of focus. A fuller comprehension model will need to assess the relative contributions of these factors and how they interact with the syntactic position of *only* and the scalar properties of its possible associates. In this light, the rich cross-linguistic picture of focus sensitive operators is of particular interest, as it can provide a more comprehensive testing ground for the nuanced predictions of the comprehension model this paper has advanced. We leave these questions for future research.

References

Aravind A, Fox D, Hackl M. 2022. Principles of presupposition in development. Linguist Philos. 46(2):291–332. https://doi.org/10.1007/s10988-022-09364-z. `http://dx.doi.org/10.1007/s10988-022-09364-z`

Beaver D, Clark B. 2008. Sense and sensitivity: How focus determines meaning. Blackwell.

Brown R. 1973. A first language: The early stages. Harvard University Press.

Crain S, Ni W, Conway L. 1994. Learning, parsing, and modularity. In: Clifton C Jr, Frazier L, Rayner K, editors. Perspectives on sentence processing. Lawrence Erlbaum Associates. p 443–467.

Crain S, Philip W, Drozd K, Roeper T, Matsuoka K. 1992. Only in child language. In: Manuscript / presented at the University of Connecticut.

Crain S, Thornton R. 1998. Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics. MIT Press.

Endo M. 2004. Developmental issues on the interpretation of focus particles by Japanese children. In: BUCLD 28 Proceedings. Somerville, MA. Cascadilla Press. p 141–152.

Fox D, Katzir R. 2011. On the characterization of alternatives. Nat Lang Semant. 19(1):87–107. https://doi.org/10.1007/s11050-010-9065-3

[37] Höhle et al. (2016) present eye-tracking data that are adult-like with regard to the alternatives to the subject and conclude from this that children have the adult grammar of S-Only. Although this is in line with our claim we believe that their experimental set up allows for alternative explanations and that follow-up work is required to solidify their argument.

Hackl M, Sugawara A, Wexler K. 2015. Question–answer (in)Congruence in the acquisition of *Only*. In: BUCLD 39: Proceedings of the 39th Annual Boston University Conference on Language Development. Somerville, MA. Cascadilla Press. p 204–217.

Horn L. 1969. A presuppositional theory of *Only* and *Even*. CLS 5, Chic Linguist Soc.

Höhle B, Fritzsche T, Müller A. 2016. Children's comprehension of sentences with focus particles and the role of cognitive control: An eye tracking study with german-learning 4-year-olds. PLOS ONE. 11(3):e0149870. https://doi.org/10.1371/journal.pone.0149870. http://dx.doi.org/10.1371/journal.pone.0149870

Müller A, Schulz P, Höhle B. 2011. How the understanding of focus particles develops: Evidence from child German. In: Proceedings of the 4th Conference on Generative Approaches to Language Acquisition North America (GALANA 2010). Somerville, MA. Cascadilla Proceedings Project. p 163–171.

Notley A, Zhou P, Crain S, Thornton R. 2009. Children's interpretation of focus expressions in English and Mandarin. Lang Acquis. 16(4):240–282.

Philip W, Lynch E. 2000. Felicity, relevance, and acquisition of the grammar of *Every* and *Only*. In: BUCLD 24 Proceedings. Somerville, MA. Cascadilla Press. p 583–596.

Poeppel D, Wexler K. 1993. The full competence hypothesis of clause structure in early german. Language. 69(1):1–33; [accessed 2026-08-09]. http://www.jstor.org/stable/416414

Roberts C. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. Semant Pragmat. 5:1–69.

Rooth M. 1985. Association with focus [dissertation]. University of Massachusetts, Amherst.

Rooth M. 1992. A theory of focus interpretation. Nat Lang Semant. 1(1):75–116. https://doi.org/10.1007/bf02342617. http://dx.doi.org/10.1007/bf02342617

Rooth M. 1996. Focus. In: Lappin S, editor. The handbook of contemporary semantic theory. Blackwell Reference. p 271–297.

Rose Y, MacWhinney B. 2014. The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. Oxf Handb Corpus Phonol.

Sanchez A et al. 2019. childes-db: A flexible and reproducible interface to the Child Language Data Exchange System. Behav Res Methods. 51:1928–1941.

Sano T. 2011. Abstract CP/IP configuration in child Japanese. In: BUCLD 36 Proceedings. Somerville, MA. Cascadilla Press. p 525–535.

Su YC. 2024. Discourse effect on children's interpretation of pre-subject exclusive zhiyou 'only'. Lang Learn Dev. 20(4):324–342. https://doi.org/10.1080/15475441.2024.2313223. http://dx.doi.org/10.1080/15475441.2024.2313223

Yang XL. 2002. Focus and quantification in child language [dissertation]. University of Maryland.

Zhou P, Crain S. 2009. Focus in child language: Evidence from the acquisition of Chinese. In: Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition North America (GALANA 2008). Somerville, MA. Cascadilla Proceedings Project. p 336–346.

Zhou P, Crain S. 2010. Focus identification in child Mandarin. J Child Lang. 37(4):965–1005.